Add a permission의

AWS Batch

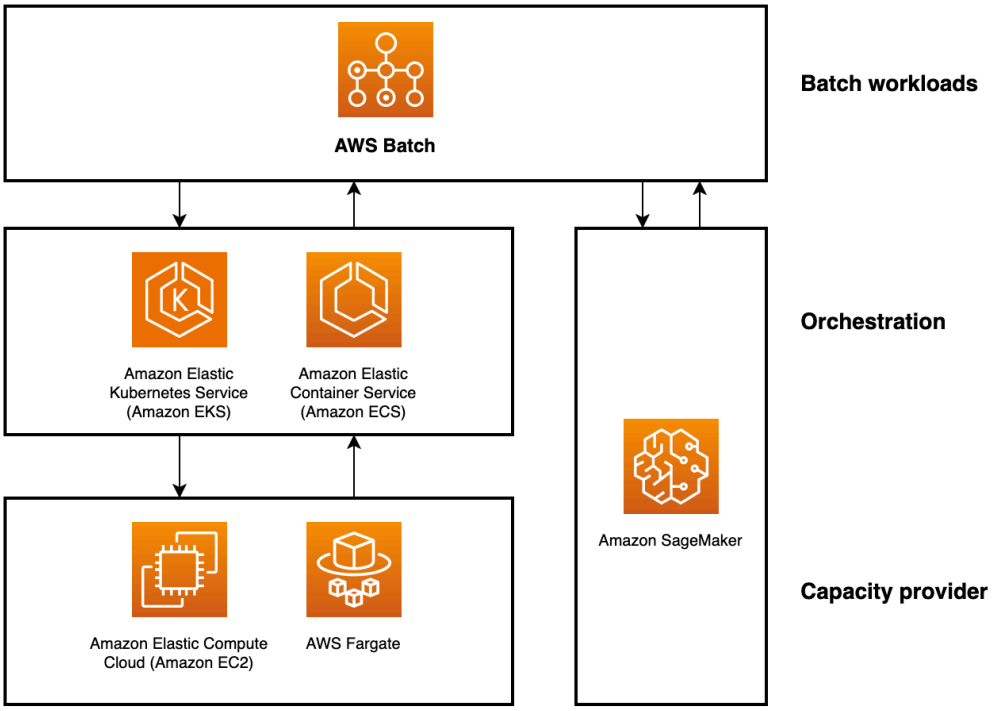
AWS Batch: Add a permission의

Table of Contents

AWS Batch란 무엇인가요?

AWS Batch 를 사용하면에서 배치 컴퓨팅 워크로드를 실행할 수 있습니다 AWS 클라우드. 배치 컴퓨팅은 개발자, 과학자, 엔지니어가 수많은 컴퓨터 리소스에 액세스할 때 일반적으로 사용하는 방법입니다. AWS Batch 는 기존의 배치 컴퓨팅 소프트웨어와 비슷하게 필요한 인프라를 구성하고 관리하는 획일적인 작업에 대한 부담을 덜 수 있습니다. 이 서비스는 제출된 작업에 응답하여 리소스를 효율적으로 프로비저닝함으로써 용량 제한을 해소하고, 컴퓨팅 비용을 줄이며, 결과를 신속하게 제공할 수 있습니다.

완전 관리형 서비스인는 모든 규모의 배치 컴퓨팅 워크로드를 실행할 수 있도록 AWS Batch 지원합니다.는 컴퓨팅 리소스를 AWS Batch 자동으로 프로비저닝하고 워크로드의 수량과 규모에 따라 워크로드 배포를 최적화합니다. 를 사용하면 배치 컴퓨팅 소프트웨어를 설치하거나 관리할 AWS Batch필요가 없으므로 결과를 분석하고 문제를 해결하는 데 시간을 집중할 수 있습니다.

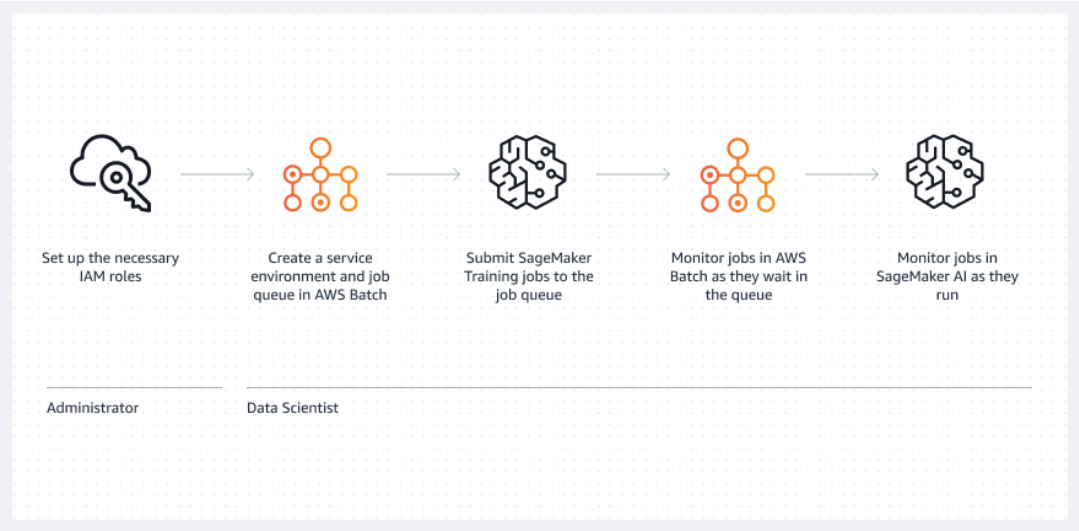


AWS Batch 는 AWS 관리형 컨테이너 오케스트레이션 서비스, Amazon ECS 및 Amazon EKS를 기반으로 대규모 컴퓨팅 집약적 워크로드를 실행하는 데 필요한 모든 기능을 제공합니다. AWS Batch 는 Amazon EC2 인스턴스 및 Fargate 리소스에서 컴퓨팅 용량을 확장할 수 있습니다.

AWS Batch 는 배치 워크로드를 위한 완전 관리형 서비스를 제공하며 처리량, 속도, 리소스 효율성 및 비용에 대해 이러한 유형의 워크로드를 최적화하는 운영 기능을 제공합니다.

AWS Batch 또한는 SageMaker 훈련 작업 대기열을 활성화하여 데이터 과학자와 ML 엔지니어가 우선 순위가 있는 훈련 작업을 구성 가능한 대기열에 제출할 수 있도록 합니다. 이 기능은 리소스를 사용할 수 있게 되는 즉시 ML 워크로드가 자동으로 실행되도록 하므로 수동 조정이 필요 없고 리소스 사용률을 향상합니다.

기계 학습 워크로드의 경우는 SageMaker 훈련 작업에 대한 대기열 기능을 AWS Batch 제공합니다. 특정 정책과 함께 대기열을 구성하여 ML 훈련 워크로드의 비용, 성능 및 리소스 할당을 최적화할 수 있습니다.



이는 관리자가 인프라와 권한을 설정하고 데이터 과학자가 ML 훈련 워크로드를 제출하고 모니터링하는 데 집중할 수 있도록 하는 공동 책임 모델을 제공합니다. 작업은 구성된 우선 순위 및 리소스 가용성에 따라 자동으로 대기열에 추가되고 실행됩니다.

처음 AWS Batch 사용하시나요?

를 처음 사용하는 경우 먼저 다음 섹션을 읽는 것이 AWS Batch좋습니다.

- 의 구성 요소 AWS Batch
- 에 가입 AWS 계정
- 설 AWS Batch정
- AWS Batch 자습서 시작하기
- SageMaker AI AWS Batch 에서 시작하기

관련 서비스

AWS Batch 는 Amazon ECS, Amazon EKS, 스팟 또는 온디맨드 인스턴스와 같은 전체 컴퓨팅 제품에서 AWS 컨테이너화된 배치 ML, 시뮬레이션 및 분석 워크로드를 계획 AWS Fargate, 예약 및 실행하는 완전관리형 배치 컴퓨팅 서비스입니다. 각 관리형 컴퓨팅 서비스에 대한 자세한 내용은 다음을 참조하세요.

- Amazon EC2 사용 설명서
- AWS Fargate 개발자 안내서
- Amazon EBS 사용 설명서
- Amazon SageMaker AI 개발자 안내서

액세스 AWS Batch

다음을 AWS Batch 사용하여에 액세스할 수 있습니다.

AWS Batch console

리소스를 생성하고 관리하는 웹 인터페이스.

AWS Command Line Interface

명령줄 셸에서 명령을 AWS 서비스 사용하여와 상호 작용합니다. AWS Command Line Interface 는 Windows, macOS 및 Linux에서 지원됩니다. 에 대한 자세한 내용은 AWS Command Line Interface 사용 설명서를 AWS CLI참조하세요. AWS Batch 명령 참조에서 AWS CLI 명령을 찾을 수 있습니다.

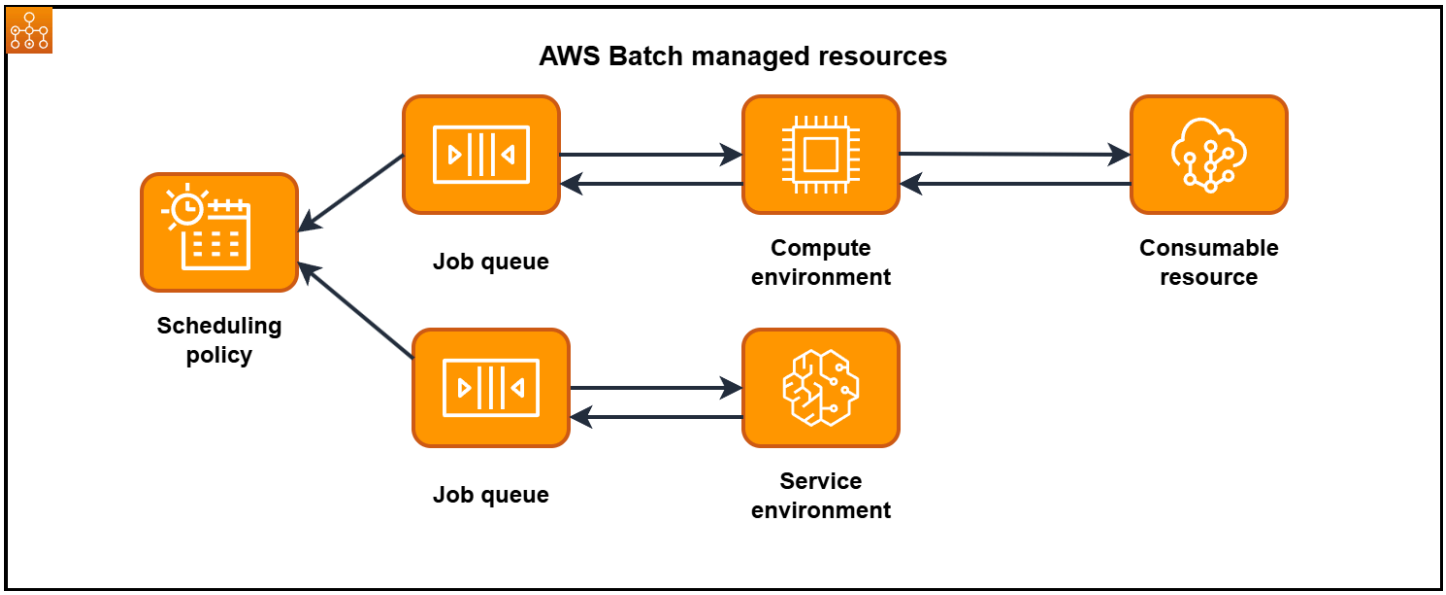
AWS SDK

HTTP 또는 HTTPS를 통해 요청을 제출하는 대신 언어별 APIs를 사용하여 애플리케이션을 빌드하려면에서 제공하는 라이브러리, 샘플 코드, 자습서 및 기타 리소스를 사용합니다 AWS. 이러한 라이브러리는 요청에 암호화 서명, 요청 재시도, 오류 응답 처리 등과 같은 작업을 자동으로 관리하는 기본 함수를 제공합니다. 이러한 함수를 사용하면 더 효율적으로 시작할 수 있습니다. 자세한 내용은 빌드 기반 도구를 참조하세요 AWS.

의 구성 요소 AWS Batch

AWS Batch 는 리전 내의 여러 가용 영역에서 배치 작업 실행을 간소화합니다. 새 VPC 또는 기존 VPC에서 AWS Batch 컴퓨팅 환경을 생성할 수 있습니다. 컴퓨팅 환경을 실행하고 작업 대기열과 연결하면

작업을 실행할 Docker 컨테이너 이미지를 지정하는 작업 정의를 정의할 수 있습니다. 컨테이너 이미지는 컨테이너 레지스트리에서 저장 및 pull됩니다. 레지스트리는 AWS 인프라 내부 또는 외부에 존재할 수 있습니다.



컴퓨팅 환경

컴퓨팅 환경은 작업을 실행하는 데 사용되는 관리형 또는 비관리형 컴퓨팅 리소스 세트입니다. 관리형 컴퓨팅 환경을 사용하면 원하는 컴퓨팅 유형(Fargate 또는 EC2)을 다양한 세부 수준에서 지정할 수 있습니다. 특정 유형의 EC2 인스턴스인 `c5.2xlarge` 또는 `m5.10xlarge`와 같은 특정 모델을 사용하는 컴퓨팅 환경을 설정할 수 있습니다. 또는 최신 인스턴스 유형만 사용하도록 지정할 수도 있습니다. 또한 스팟 인스턴스에 대해 지불할 금액과 함께 환경의 최소, 원하는 및 최대 vCPUs 수를 온디맨드 인스턴스 가격 및 대상 VPC 서브넷 세트의 백분율로 지정할 수 있습니다. 필요에 따라 컴퓨팅 유형을 AWS Batch 효율적으로 시작, 관리 및 종료합니다. 사용자 고유의 컴퓨팅 환경을 관리할 수도 있습니다. 따라서 AWS Batch 자동으로 생성하는 Amazon ECS 클러스터에서 인스턴스를 설정하고 확장하는 것은 사용자의 책임입니다. 자세한 내용은 컴퓨팅 환경 AWS Batch 단원을 참조하십시오.

작업 대기열

AWS Batch 작업을 제출할 때 작업이 컴퓨팅 환경에 예약될 때까지 작업이 상주하는 특정 작업 대기열에 제출합니다. 하나의 작업 대기열에 하나 이상의 컴퓨팅 환경을 연결할 수 있습니다. 또한 이러한 컴퓨팅 환경과 작업 대기열 자체에도 우선 순위 값을 할당할 수 있습니다. 예를 들어 시간에 민감한 작업을 제출할 때는 우선 순위가 높은 대기열을 가질 수 있습니다. 컴퓨팅 리소스 비용이 더 저렴할 때 언제든지 실행할 수 있는 작업은 우선 순위가 낮은 대기열을 가질 수 있습니다. 자세한 내용은 작업 대기열 단원을 참조하십시오.

작업 정의

작업 정의는 작업이 어떻게 실행될지를 지정합니다. 작업 정의는 작업에 들어가는 리소스에 대한 청사진이라고 할 수 있습니다. 작업에 IAM 역할을 제공하여 다른 AWS 리소스에 대한 액세스를 제공할 수 있습니다. 또한 메모리와 CPU 요구 사항을 모두 지정합니다. 작업 정의는 영구 스토리지의 컨테이너 속성, 환경 변수, 마운트 지점을 제어할 수도 있습니다. 작업 정의의 많은 사양은 개별 작업을 제출할 때 새 값을 지정하여 재정의될 수 있습니다. 자세한 내용은 작업 정의 섹션을 참조하세요.

작업

AWS Batch에 제출한 작업 단위(쉘 스크립트, Linux 실행 파일, Docker 컨테이너 이미지)입니다. 이름이 있으며, 작업 정의에서 지정한 파라미터를 사용하여 컴퓨팅 환경의 AWS Fargate 또는 Amazon EC2 리소스에서 컨테이너화된 애플리케이션으로 실행됩니다. 작업은 이름 또는 ID로 다른 작업을 참조할 수 있으며 다른 작업이 성공적으로 완료되었는지 여부 또는 지정한 리소스의 가용성에 따라 달라질 수 있습니다. 자세한 내용은 작업 단원을 참조하십시오.

예약 정책

사용자는 예약 정책을 사용하여 작업 대기열의 컴퓨팅 리소스를 사용자 또는 워크로드에 할당하는 방식을 구성할 수 있습니다. 공정 공유 예약 정책을 사용하면 워크로드 또는 사용자에게 서로 다른 공유 식별자를 할당할 수 있습니다. AWS Batch 작업 스케줄러는 기본적으로 선입선출(FIFO) 전략으로 설정됩니다. 자세한 내용은 공정 공유 예약 정책 단원을 참조하십시오.

소모성 리소스

소모성 리소스는 타사 라이선스 토큰, 데이터베이스 액세스 대역폭, 타사 API에 대한 호출 제한 필요성 등과 같이 작업을 실행하는 데 필요한 리소스입니다. 작업을 실행하는 데 필요한 소모성 리소스를 지정하면 Batch는 작업을 예약할 때 이러한 리소스 종속성을 고려합니다. 필요한 모든 리소스를 사용할 수 있는 작업만 할당하여 컴퓨팅 리소스의 사용 부족을 줄일 수 있습니다. 자세한 내용은 리소스 인식 일정 예약 단원을 참조하십시오.

서비스 환경

서비스 환경은가 작업 실행을 위해 SageMaker와 AWS Batch 통합되는 방법을 정의합니다. 서비스 환경을 사용하면 AWS Batch 가의 대기열, 예약 및 우선 순위 관리 기능을 제공하면서 SageMaker에서 작업을 제출하고 관리할 수 있습니다 AWS Batch. 서비스 환경은 SageMaker 훈련 작업과 같은 특정 서비스 유형에 대한 용량 제한을 정의합니다. 용량 제한은 환경의 서비스 작업에서 사용할 수 있는 최대 리소스를 제어합니다. 자세한 내용은 용 서비스 환경 AWS Batch 단원을 참조하십시오.

서비스 작업

서비스 작업은 서비스 환경에서 실행하기 AWS Batch 위해에 제출하는 작업 단위입니다. 서비스 작업은 AWS Batch의 대기열 및 예약 기능을 활용하는 동시에 실제 실행을 외부 서비스에 위임합니다. 예를 들어 서비스 작업으로 제출된 SageMaker 훈련 작업은에서 대기열에 추가되고 우선 순위가 지정 AWS Batch되지만 SageMaker 훈련 작업 실행은 SageMaker AI 인프라 내에서 이루어집니다. 이 통합을 통해 데이터 과학자와 ML 엔지니어는 SageMaker AI 훈련 워크로드에 대한 AWS Batch의 자동화된 워크로드 관리 및 우선 순위 대기열을 활용할 수 있습니다. 서비스 작업은 이름 또는 ID로 다른 작업을 참조하고 작업 종속성을 지원할 수 있습니다. 자세한 내용은 의 서비스 작업 AWS Batch 단원을 참조하십시오.

설 AWS Batch정

Amazon Web Services(AWS에 이미 가입했고 Amazon Elastic Compute Cloud(Amazon EC2) 또는 Amazon Elastic Container Service(Amazon ECS)를 사용하고 있는 경우, AWS Batch를 곧 사용할 수 있습니다. 이 서비스의 설정 프로세스는 유사합니다. 이는가 컴퓨팅 환경에서 Amazon ECS 컨테이너 인스턴스를 AWS Batch 사용하기 때문입니다. 를 AWS CLI 와 함께 사용하려면 최신 AWS Batch 기능을 AWS CLI 지원하는 버전의를 사용해야 AWS Batch합니다. 에서 AWS Batch 기능에 대한 지원이 표시되지 않으면 최신 버전으로 AWS CLI업그레이드합니다. 자세한 내용은 http://aws.amazon.com/cli/를 참조하세요.

> **Note**
>
> AWS Batch 는 Amazon EC2의 구성 요소를 사용하므로 이러한 많은 단계에서 Amazon EC2 콘솔을 사용합니다.

를 설정하려면 다음 작업을 완료합니다 AWS Batch.

주제

- 에 가입 AWS 계정
- 컴퓨팅 환경 및 컨테이너 인스턴스에 대한 IAM 역할 생성
- 인스턴스의 키 페어 생성
- VPC 생성
- 보안 그룹 생성
- AWS CLI 설치

에 가입 AWS 계정

를 시작하려면이 AWS필요합니다 AWS 계정. 생성에 대한 자세한 AWS 계정내용은 AWS Account Management 참조 안내서의 시작하기 AWS 계정를 참조하세요.

컴퓨팅 환경 및 컨테이너 인스턴스에 대한 IAM 역할 생성

AWS Batch 컴퓨팅 환경과 컨테이너 인스턴스를 이용하려면 사용자를 대신하여 다른 AWS API를 호출할 수 있도록 AWS 계정 자격 증명이 필요합니다. 컴퓨팅 환경과 컨테이너 인스턴스에 이러한 자격

증명을 제공하는 AWS Identity and Access Management 역할을 생성한 후, 이 역할을 해당 컴퓨팅 환경과 연결해야 합니다.

> **Note**
>
> AWS 계정에 필요한 권한이 있는지 확인하려면 계정에 초기 IAM 서비스 설정을 참조하세요. 콘솔 첫 실행 시, AWS Batch 컴퓨팅 환경과 컨테이너 인스턴스 역할이 자동으로 생성됩니다. 따라서 AWS Batch 콘솔을 사용하려는 경우 다음 섹션으로 진행할 수 있습니다. 대신 AWS CLI를 사용하려는 경우, 컴퓨팅 환경을 처음 생성하기 전에 에 대한 서비스 연결 역할 사용 AWS Batch, Amazon ECS 인스턴스 역할 및 자습서: IAM 실행 역할 생성의 절차를 완료합니다.

인스턴스의 키 페어 생성

AWS 는 퍼블릭 키 암호화를 사용하여 인스턴스의 로그인 정보를 보호합니다. AWS Batch 컴퓨팅 환경 컨테이너 인스턴스와 같은 Linux 인스턴스에는 SSH 액세스에 사용할 암호가 없습니다. 키 페어를 사용하면 인스턴스에 안전하게 로그인할 수 있습니다. 컴퓨팅 환경을 생성할 때 키 페어 이름을 지정한 후 SSH를 통해 로그인하면서 프라이빗 키를 입력합니다.

키 페어를 아직 생성하지 않은 경우 사용자는 Amazon EC2 콘솔을 사용하여 키 페어를 생성할 수 있습니다. 인스턴스를 여러 개 시작하려는 경우 각 리전에서 키 페어를 AWS 리전생성합니다. 리전에 대한 자세한 내용은 Amazon EC2 사용 설명서의 리전 및 가용 영역을 참조하세요.

키 페어 생성

1. https://console.aws.amazon.com/ec2/에서 Amazon EC2 콘솔을 엽니다.
2. 탐색 모음에서 키 페어에 AWS 리전 대한를 선택합니다. 위치에 상관없이 사용 가능한 어떤 리전도 선택할 수 있지만 키 페어는 리전별로 고유합니다. 예를 들어 미국 서부(오레곤) 리전에서 인스턴스를 시작하려면 동일 리전에서 인스턴스 키 페어를 생성해야 합니다.
3. 탐색 창에서 Key Pairs(키 페어), Create Key Pair(키 페어 생성)를 선택합니다.
4. Create Key Pair(키 페어 생성) 대화 상자의 Key pair name(키 페어 이름)에 새로운 키 페어 이름을 입력하고 Create(생성)를 선택합니다. 기억하기 쉬운 이름을 선택합니다(예: 즉, 사용자 이름, `-key-pair` 붙이기, 리전 이름 조합). 예를 들어, me-key-pair-uswest2로 지정할 수 있습니다.
5. 브라우저에서 프라이빗 키 파일이 자동으로 다운로드됩니다. 기본 파일 이름은 키 페어의 이름으로 지정된 이름이며, 파일 이름 확장명은 `.pem`입니다. 안전한 장소에 프라이빗 키 파일을 저장합니다.

> **Important**
>
> 이때가 사용자가 프라이빗 키 파일을 저장할 수 있는 유일한 기회입니다. 인스턴스를 시작할 때 키 페어의 이름을 제공하고, 인스턴스에 연결할 때마다 상응하는 프라이빗 키를 제공해야 합니다.

6. Mac 또는 Linux 컴퓨터에서 SSH 클라이언트를 사용하여 Linux 인스턴스에 연결하려면 사용자만 다음 명령으로 프라이빗 키 파일의 권한을 설정합니다. 이렇게 하면 사용자만 읽을 수 있습니다.

```
$ chmod 400 your_user_name-key-pair-region_name.pem
```

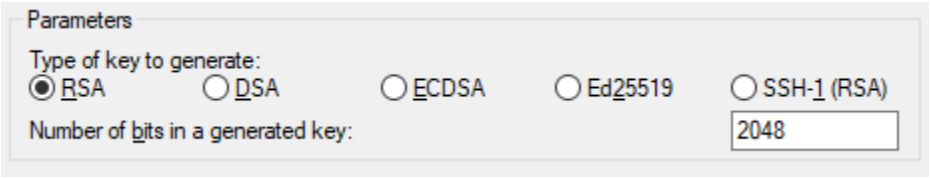
자세한 내용은 Amazon EC2 사용 설명서의 Amazon EC2 키 페어를 참조하세요.

키 페어를 사용하여 인스턴스에 연결하려면

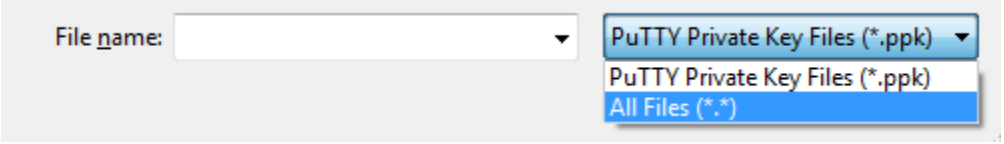
Mac 또는 Linux를 실행 중인 컴퓨터에서 Linux 인스턴스에 연결하려면 `.pem` 옵션과 프라이빗 키 경로를 사용하여 SSH 클라이언트에 `-i` 파일을 지정합니다. Windows 운영 체제의 컴퓨터에서 Linux 인스턴스에 연결하려면 MindTerm 또는 PuTTY를 사용합니다. PuTTY를 사용하려면 이를 설치하고 다음 절차에 따라 `.pem` 파일을 `.ppk` 파일로 변환해야 합니다.

(선택 사항) PuTTY를 사용하여 Windows에서 Linux 인스턴스에 연결하려면

1. http://www.chiark.greenend.org.uk/~sgtatham/putty/에서 PuTTY를 다운로드하고 설치합니다. 전체 제품군을 설치해야 합니다.
2. PuTTYgen을 시작합니다(예: 시작 메뉴에서 모든 프로그램, PuTTY, PuTTYgen을 선택합니다).
3. 생성할 키 유형(Type of key to generate)에서 RSA를 선택합니다. 이전 버전의 PuTTYgen을 사용한다면 SSH-2 RSA를 선택합니다.



4. 로드(Load)를 선택합니다. 기본적으로 PuTTYgen에는 확장명이 `.ppk`인 파일만 표시됩니다. `.pem` 파일을 찾으려면 모든 유형의 파일을 표시하는 옵션을 선택합니다.



5. 이전 절차에서 생성한 프라이빗 키 파일을 선택하고 열기(Open)를 선택합니다. 확인(OK)을 선택하여 확인 대화 상자를 닫습니다.
6. 프라이빗 키 저장(Save private key)을 선택합니다. PuTTYgen에서 암호 없이 키 저장에 대한 경고가 표시됩니다. 예(Yes)를 선택합니다.
7. 키 페어에 사용한 것과 동일한 키 이름을 지정합니다. PuTTY가 자동으로 `.ppk` 파일 확장자를 추가합니다.

VPC 생성

Amazon Virtual Private Cloud(VPC)를 사용하면 사용자가 정의한 가상 네트워크의 AWS 리소스를 시작할 수 있습니다. VPC에서 컨테이너 인스턴스를 시작할 것을 권장합니다.

기본 VPC가 있는 경우 이 섹션을 건너뛰고 다음 작업인 보안 그룹 생성으로 이동합니다. 기본 VPC가 있는지를 확인하려면 Amazon EC2 사용 설명서의 Amazon EC2 콘솔에서 지원되는 플랫폼을 참조하세요.

Amazon VPC 생성에 대한 자세한 내용은 Amazon VPC 사용 설명서의 VPC만 생성을 참조하세요. 다음 표를 참조하여 선택할 옵션을 결정하세요.

옵션	값
생성할 리소스	VPC 전용
명칭	선택적으로 VPC의 이름을 입력합니다.
IPv4 CIDR block	IPv4 CIDR 수동 입력 CIDR 블록 크기는 /16과 /28 사이여야 합니다.
IPv6 CIDR block	No IPv6 CIDR 블록
Tenancy	Default

Amazon VPC에 대한 자세한 내용은 Amazon VPC 사용 설명서의 Amazon VPC란 무엇인가? 섹션을 참조하세요.

보안 그룹 생성

보안 그룹은 연결된 컴퓨팅 환경 컨테이너 인스턴스에 대한 방화벽 역할을 하여 컨테이너 인스턴스 수준에서 인바운드 트래픽과 아웃바운드 트래픽을 모두 제어합니다. 보안 그룹은 보안 그룹이 생성된 VPC에서만 사용할 수 있습니다.

해당 IP 주소에서 SSH를 사용하여 컨테이너 인스턴스에 연결할 수 있게 하는 규칙을 보안 그룹에 추가할 수 있습니다. 어디서나 인바운드 및 아웃바운드 HTTP/HTTPS 액세스를 허용하는 규칙을 추가할 수도 있습니다. 작업에 필요한 포트를 여는 규칙을 추가합니다.

여러 리전에서 컨테이너 인스턴스를 시작하려면 각 리전에 보안 그룹을 생성해야 합니다. 자세한 내용은 Amazon EC2 사용 설명서에서 리전 및 가용 영역을 참조하세요.

> **Note**
>
> 로컬 컴퓨터의 퍼블릭 IP 주소가 필요하며, 서비스를 사용해 확인할 수 있습니다. 예를 들면, Amazon에서는 http://checkip.amazonaws.com/ 또는 https://checkip.amazonaws.com/ 서비스를 제공합니다. IP 주소를 제공하는 다른 서비스를 찾으려면 "what is my IP address"로 검색합니다. 인터넷 서비스 제공업체(ISP)를 통해서 혹은 고정 IP 주소가 없는 방화벽 뒤편에서 접속한다면 클라이언트 컴퓨터가 사용하는 IP 주소의 범위를 찾습니다.

콘솔을 사용하여 보안 그룹을 생성하려면

1. https://console.aws.amazon.com/vpc/에서 Amazon VPC 콘솔을 엽니다.
2. 탐색 창에서 [Security Groups]를 선택합니다.
3. 보안 그룹 생성을 선택합니다.
4. 보안 그룹의 이름과 설명을 입력합니다. 보안 그룹을 생성한 후에는 보안 그룹에 대한 이름과 설명을 변경할 수 없습니다.
5. VPC에서 VPC를 선택합니다.
6. (선택 사항) 기본적으로 처음에 새 보안 그룹에는 리소스에서 나가는 모든 트래픽을 허용하는 아웃바운드 규칙만 적용됩니다. 인바운드 트래픽을 사용하거나 아웃바운드 트래픽을 제한하려면 규칙을 추가해야 합니다.

   AWS Batch 컨테이너 인스턴스는 인바운드 포트를 열어야 할 필요가 없습니다. 하지만 SSH 규칙을 추가할 수는 있습니다. 이렇게 하여 사용자는 컨테이너 인스턴스에 로그인하고 도커 명령을 사용하여 작업 컨테이너를 검사할 수 있습니다. 또한 컨테이너 인스턴스에서 웹 서버를 실행하는 작

업을 호스팅하게 하려는 경우 HTTP 규칙을 추가할 수도 있습니다. 이러한 보안 그룹 규칙을 필요에 따라 추가하려면 다음 절차를 수행합니다.

Inbound(인바운드) 탭에서 다음 규칙을 생성하고 Create(생성)를 선택합니다.

- 규칙 추가(Add Rule)를 선택합니다. Type(유형)에서 HTTP를 선택합니다. Source(소스)에서 Anywhere(위치 무관)(`0.0.0.0/0`)를 선택합니다.
- 규칙 추가(Add Rule)를 선택합니다. Type(유형)에서 SSH를 선택합니다. 소스에서 사용자 지정 IP를 선택하고 사용자 컴퓨터 또는 네트워크의 퍼블릭 IP 주소를 Classless Inter-Domain Routing(CIDR) 표기법으로 지정합니다. 회사에서 주소를 범위로 할당하는 경우 전체 범위(예: `203.0.113.0/24`)를 지정합니다. CIDR 표기법으로 개별 IP 주소를 지정하려면 내 IP를 선택합니다. 퍼블릭 IP 주소에 라우팅 접두사 `/32`가 추가됩니다.

> **Note**
>
> 보안상, 테스트용으로 짧은 시간 동안만 허용하는 경우를 제외하고는 사용자 인스턴스에 대한 모든 IP 주소(`0.0.0.0/0`)의 SSH 액세스를 허용하지 않는 것이 좋습니다.

7. 태그를 지금 추가하거나 나중에 추가할 수 있습니다. 태그를 추가하려면 새 태그 추가(Add new tag)를 선택한 다음 태그 키와 값을 입력합니다.
8. 보안 그룹 생성을 선택합니다.

명령줄을 사용하여 보안 그룹을 생성하려면 >create-security-group(AWS CLI)을 참조하세요.

보안 그룹 작업에 대한 자세한 내용은 보안 그룹 작업하기를 참조하세요.

AWS CLI 설치

AWS CLI에 AWS Batch를 사용하려면 최신 AWS CLI 버전을 설치합니다. AWS CLI 설치 또는 최신 버전 업그레이드 방법에 대한 자세한 내용은 AWS Command Line Interface 사용 설명서의 AWS 명령줄 인터페이스 설치를 참조하세요.

AWS Batch 자습서 시작하기

AWS Batch 처음 실행 마법사를 사용하여 빠르게 시작할 수 있습니다 AWS Batch. 사전 요건을 완비한 후에는 최초 실행 마법사를 사용하여 컴퓨팅 환경, 작업 정의 및 작업 대기열을 생성할 수 있습니다.

또한 AWS Batch 최초 실행 마법사를 사용하여 샘플 "Hello World" 작업을 제출하여 구성을 테스트할 수 있습니다. 시작하려는 Docker 이미지가 이미 있는 경우 해당 이미지를 사용하여 작업 정의를 생성할 AWS Batch수 있습니다.

그런 다음 AWS Batch 처음 실행 마법사를 사용하여 컴퓨팅 환경, 작업 대기열을 생성하고 샘플 Hello World 작업을 제출할 수 있습니다.

마법사를 사용하여 Amazon EC2 오케스트레이션 시작하기

Amazon Elastic Compute Cloud(Amazon EC2)는 AWS 클라우드에서 확장 가능한 컴퓨팅 용량을 제공합니다. Amazon EC2를 사용하면 하드웨어에 사전 투자할 필요가 없어 더 빠르게 애플리케이션을 개발하고 배포할 수 있습니다.

Amazon EC2를 사용하여 원하는 수의 가상 서버를 구축하고 보안 및 네트워킹을 구성하며 스토리지를 관리할 수 있습니다. Amazon EC2에서는 스케일 업 또는 다운을 통해 요구 사항 변경 또는 사용량 급증을 처리할 수 있으므로 트래픽을 예측할 필요성이 줄어듭니다.

개요

이 자습서에서는 마법사 AWS Batch 를 사용하여를 설정하여 Amazon EC2를 구성하고를 실행하는 방법을 보여줍니다`Hello World`.

대상

이 자습서는 AWS Batch의 설정, 테스트 및 배포를 담당하는 시스템 관리자 및 개발자를 위해 설계되었습니다.

사용된 기능

이 자습서에서는 AWS Batch 콘솔 마법사를 사용하여 다음을 수행하는 방법을 보여줍니다.

- Amazon EC2 컴퓨팅 환경 생성 및 구성
- 작업 대기열을 생성합니다.
- 작업 정의 생성
- 실행할 작업 생성 및 제출

- CloudWatch에서 작업 출력 보기

필요한 시간

이 자습서를 완료하는 데 약 10~15분이 소요됩니다.

리전별 제한 사항

이 솔루션의 사용과 관련된 국가별 또는 리전별 제한은 없습니다.

리소스 사용 비용

AWS 계정 생성에는 요금이 부과되지 않습니다. 그러나 이 솔루션을 구현하면 아래 표에 나열된 비용이 일부 또는 전부 발생할 수 있습니다.

설명	비용(USD)
Amazon EC2 인스턴스	생성된 각 Amazon EC2 인스턴스에 대해 비용을 지불합니다. 요금에 대한 자세한 정보는 Amazon EC2 요금을 참조하십시오.

사전 조건

시작하기 전:

- 없는 AWS 계정 경우를 생성합니다.
- ecsInstanceRole 인스턴스 역할을 생성합니다.

1단계 컴퓨팅 환경 생성

> **Important**
>
> 이 자습서에는 가능한 한 간단하고 빠르게 시작하기 위해 기본 설정을 사용하는 단계가 포함되어 있습니다. 프로덕션 용도로 생성하기 전에 모든 설정을 숙지하고 요구 사항을 충족하는 설정으로 배포하는 것이 좋습니다.

Amazon EC2 오케스트레이션을 위한 컴퓨팅 환경을 만들려면

1. AWS Batch 콘솔 최초 실행 마법사를 엽니다.

2. 작업 및 오케스트레이션 유형 구성에서 Amazon Elastic Compute Cloud(Amazon EC2)를 선택합니다.
3. 다음을 선택합니다.
4. 이름 컴퓨팅 환경 구성에서 해당 컴퓨팅 환경의 고유한 이름을 지정합니다. 각 이름의 최대 길이는 128자입니다. 대문자 및 소문자, 숫자, 하이픈(-) 및 밑줄(_)을 포함할 수 있습니다.
5. 인스턴스 역할에서 필요한 IAM 권한이 연결되어 있는 기존 인스턴스 역할을 선택합니다. 이 인스턴스 역할을 사용하면 컴퓨팅 환경의 Amazon ECS 컨테이너 인스턴스가 필요한 AWS API 작업을 호출할 수 있습니다. 자세한 내용은 Amazon ECS 인스턴스 역할 단원을 참조하십시오.

   인스턴스 역할의 기본 이름은 `ecsInstanceRole`입니다.
6. 인스턴스 구성의 경우 기본 설정을 그대로 둘 수 있습니다.
7. 네트워크 구성에는 AWS 리전에 대한 기본 VPC를 사용합니다.
8. 다음을 선택합니다.

2단계: 작업 대기열 생성

작업 대기열은 AWS Batch 스케줄러가 컴퓨팅 환경의 리소스에서 작업을 실행할 때까지 제출된 작업을 저장합니다. 자세한 내용은 작업 대기열 섹션을 참조하세요.

Amazon EC2 오케스트레이션을 위한 작업 대기열을 만들려면

1. 이름에 대한 작업 대기열 구성에 작업 대기열의 고유한 이름을 지정합니다. 각 이름의 최대 길이는 128자입니다. 대문자 및 소문자, 숫자, 하이픈(-) 및 밑줄(_)을 포함할 수 있습니다.
2. 다른 모든 구성 옵션에는 기본값을 그대로 둘 수 있습니다.
3. 다음을 선택합니다.

3단계: 작업 정의 생성

AWS Batch 작업 정의는 작업 실행 방법을 지정합니다. 각 작업은 작업 정의를 참조해야 하지만, 작업 정의에 지정된 대부분의 파라미터는 런타임에 재정의될 수 있습니다.

작업 정의를 생성하려면

1. 작업 정의 생성의 경우

   a. 이름에 작업 대기열의 고유한 이름을 지정합니다. 각 이름의 최대 길이는 128자입니다. 대문자 및 소문자, 숫자, 하이픈(-) 및 밑줄(_)을 포함할 수 있습니다.

b. 명령 - 선택 사항의 경우, `hello world`를 사용자 지정 메시지로 변경하거나 그대로 둘 수 있습니다.

2. 다른 모든 구성 옵션에는 기본값을 그대로 둘 수 있습니다.
3. 다음을 선택합니다.

4단계: 작업 생성

작업을 생성하려면 다음을 수행합니다.

1. 이름의 작업 구성 섹션에서 작업의 고유한 이름을 지정합니다. 각 이름의 최대 길이는 128자입니다. 대문자 및 소문자, 숫자, 하이픈(-) 및 밑줄(_)을 포함할 수 있습니다.
2. 다른 모든 구성 옵션에는 기본값을 그대로 둘 수 있습니다.
3. 다음을 선택합니다.

5단계: 검토 및 생성

검토 및 생성 페이지에서 구성 과정을 검토합니다. 변경해야 하는 경우 편집을 선택합니다 입력이 끝나면 리소스 생성을 선택합니다.

1. 검토 및 생성에서 리소스 생성을 선택합니다.
2. 가 리소스를 할당하기 AWS Batch 시작하면 창이 열립니다. 완료되면 대시보드로 이동을 선택합니다. 대시보드에서는 할당된 모든 리소스를 볼 수있으며 작업이 `Runnable` 상태임을 확인할 수 있습니다. 작업 실행이 예약되고 2~3분 후에 완료됩니다.

6단계: 작업의 출력 보기

작업의 출력을 보려면 다음을 수행합니다.

1. 탐색 창에서 작업을 선택합니다.
2. 작업 대기열 드롭다운에서 자습서를 위해 생성한 작업 대기열을 선택합니다.
3. 작업 테이블에는 모든 작업과 현재 상태가 나열됩니다. 작업의 상태가 성공이 되면 작업 이름을 선택하여 작업의 세부 정보를 봅니다.
4. 세부 정보 창에서 로그 스트림 이름을 선택합니다. 작업에 대한 CloudWatch 콘솔이 열리고 `hello world` 메시지 또는 사용자 지정 메시지가 포함된 이벤트가 하나 있을 것입니다.

7단계: 자습서 리소스 정리

Amazon EC2 인스턴스가 활성화된 동안 요금이 부과됩니다. 요금이 발생하지 않도록 해당 인스턴스를 종료할 수 있습니다.

생성한 리소스를 삭제하려면 다음과 같이 합니다.

1. 탐색 창에서 작업 대기열을 선택합니다.
2. 작업 대기열 테이블에서 자습서를 위해 생성한 작업 대기열을 선택합니다.
3. 비활성화를 선택합니다. 작업 대기열 상태가 비활성화가 되면 삭제를 선택할 수 있습니다.
4. 작업 대기열이 삭제되면 탐색 창에서 컴퓨팅 환경을 선택합니다.
5. 이 자습서용으로 생성한 컴퓨팅 환경을 선택한 후 비활성화를 선택합니다. 컴퓨팅 환경 비활성화가 완료되는 데는 1~2분 정도 걸릴 수 있습니다.
6. 컴퓨팅 환경의 상태가 비활성화가 되면 삭제를 선택합니다. 컴퓨팅 환경이 삭제되는 데는 1~2분 정도 걸릴 수 있습니다.

추가 리소스

자습서를 완료한 후 다음과 같은 주제를 살펴보세요.

- AWS Batch 핵심 구성 요소를 살펴봅니다. 자세한 내용은 의 구성 요소 AWS Batch 단원을 참조하십시오.
- AWS Batch에서 사용할 수 있는 다양한 컴퓨팅 환경에 대해 자세히 알아봅니다.
- 작업 대기열 및 다양한 예약 옵션에 대해 자세히 알아봅니다.
- 작업 정의 및 다양한 구성 옵션에 대해 자세히 알아봅니다.
- 다른 작업 유형에 대해 자세하 알아봅니다.

마법사를 사용하여 AWS Batch 및 Fargate 오케스트레이션 시작하기

AWS Fargate는 컨테이너에 대해 지정한 리소스 요구 사항과 거의 일치하도록 컴퓨팅을 시작하고 확장합니다. Fargate를 사용하면 과도하게 프로비저닝하거나 추가 서버를 위해 비용을 지불할 필요가 없습니다. 자세한 내용은 Fargate를 참조하세요.

개요

이 자습서에서는 마법사 AWS Batch 를 사용하여를 설정하여 AWS Fargate를 구성하고를 실행하는 방법을 보여줍니다`Hello World`.

대상

이 자습서는 AWS Batch의 설정, 테스트 및 배포를 담당하는 시스템 관리자 및 개발자를 위해 설계되었습니다.

사용된 기능

이 자습서에서는 AWS Batch 콘솔 마법사를 사용하여 다음을 수행하는 방법을 보여줍니다.

- AWS Fargate 컴퓨팅 환경 생성 및 구성
- 작업 대기열을 생성합니다.
- 작업 정의 생성
- 실행할 작업 생성 및 제출
- CloudWatch에서 작업 출력 보기

필요한 시간

이 자습서를 완료하는 데 약 10~15분이 소요됩니다.

리전별 제한 사항

이 솔루션의 사용과 관련된 국가별 또는 리전별 제한은 없습니다.

리소스 사용 비용

AWS 계정 생성에는 요금이 부과되지 않습니다. 그러나 이 솔루션을 구현하면 아래 표에 나열된 비용이 일부 또는 전부 발생할 수 있습니다.

설명	비용(USD)
요금은 태스크 또는 포드에 대해 요청된 vCPU, 메모리, 운영 체제, CPU 아키텍처 및 스토리지 리소스를 기반으로 합니다.	요금에 대한 자세한 내용은 Fragate 요금을 참조하세요.

사전 조건

시작하기 전:

- 없는 AWS 계정 경우를 생성합니다.
- 태스크 실행 역할을 생성합니다. 태스크 실행 역할을 아직 생성하지 않은 경우, 이 자습서의 일부로 생성할 수 있습니다.

1단계 컴퓨팅 환경 생성

> **Important**
>
> 이 자습서에는 가능한 한 간단하고 빠르게 시작하기 위해 기본 설정을 사용하는 단계가 포함되어 있습니다. 프로덕션 용도로 생성하기 전에 모든 설정을 숙지하고 요구 사항을 충족하는 설정으로 배포하는 것이 좋습니다.

Fargate 오케스트레이션을 위한 컴퓨팅 환경을 만들려면

1. AWS Batch 콘솔 최초 실행 마법사를 엽니다.
2. 작업 및 오케스트레이션 유형 구성에서 Fargate를 선택합니다.
3. 다음을 선택합니다.
4. 이름 컴퓨팅 환경 구성에서 해당 컴퓨팅 환경의 고유한 이름을 지정합니다. 각 이름의 최대 길이는 128자입니다. 대문자 및 소문자, 숫자, 하이픈(-) 및 밑줄(_)을 포함할 수 있습니다.
5. 다른 모든 구성 옵션에는 기본값을 그대로 둘 수 있습니다.
6. 다음을 선택합니다.

2단계: 작업 대기열 생성

작업 대기열은 AWS Batch 스케줄러가 컴퓨팅 환경의 리소스에서 작업을 실행할 때까지 제출된 작업을 저장합니다. 작업 대기열 생성

Fargate 오케스트레이션을 위한 작업 대기열을 만들려면

1. 이름에 대한 작업 대기열 구성 섹션에서 작업 대기열의 고유한 이름을 지정합니다. 각 이름의 최대 길이는 128자입니다. 대문자 및 소문자, 숫자, 하이픈(-) 및 밑줄(_)을 포함할 수 있습니다.
2. 우선 순위에서 작업 대기열에 900을 입력합니다.
3. 다른 모든 구성 옵션에는 기본값을 그대로 둘 수 있습니다.
4. 다음을 선택합니다.

3단계: 작업 정의 생성

작업 정의를 생성하려면

1. 일반 구성 섹션에서:

   - 이름에 대한 일반 구성 섹션에서 작업 정의의 고유한 이름을 지정합니다. 각 이름의 최대 길이는 128자입니다. 대문자 및 소문자, 숫자, 하이픈(-) 및 밑줄(_)을 포함할 수 있습니다.

2. Fargate 플랫폼 구성 섹션에서:

   a. 퍼블릭 IP 할당을 켜서 퍼블릭 IP 주소를 할당합니다. 프라이빗 이미지 리포지토리를 설정하지 않은 한 컨테이너 이미지를 다운로드하려면 퍼블릭 IP가 필요합니다.

   b. 실행 역할에서 Amazon Elastic Container Service(Amazon ECS) 에이전트가 사용자를 대신하여 AWS 호출할 수 있는 작업 실행 역할을 선택합니다. ecsTaskExecutionRole 또는 BatchEcsTaskExecutionRole을 선택합니다.

      실행 역할을 생성하려면 실행 역할 생성을 선택합니다. IAM 역할 생성 모달에서 IAM 역할 생성을 선택합니다.

      i. IAM 콘솔에는 실행 역할 생성을 위해 이미 구성된 권한 설정이 있습니다.

      ii. 신뢰할 수 있는 엔터티 유형에 AWS 서비스를 선택되어 있는지 학인합니다.

      iii. 서비스 또는 사용자 사례에 Elastic Container Service가 선택되어 있는지 확인합니다.

      iv. 다음을 선택합니다.

      v. 권한 정책에 AmazonECSTaskExecutionRolePolicy 정책이 선택되어 있는지 확인합니다.

      vi. 다음을 선택합니다.

      vii. 이름, 검토 및 생성에서 역할 이름이 BatchEcsTaskExecutionRole인지 확인합니다.

      viii. 역할 생성을 선택합니다.

      ix. AWS Batch 콘솔에서 실행 역할 옆의 새로 고침 버튼을 선택합니다. BatchEcsTaskExecutionRole 실행 역할을 선택합니다.

3. 컨테이너 구성 섹션에서:

   - 명령에서 `hello world`를 사용자 지정 메시지로 변경하거나 그대로 둘 수 있습니다.

4. 다른 모든 구성 옵션에는 기본값을 그대로 둘 수 있습니다.

5. 다음을 선택합니다.

4단계: 작업 생성

Fargate 작업을 생성하려면 다음을 수행합니다.

1. 이름의 작업 구성 섹션에서 작업의 고유한 이름을 지정합니다. 각 이름의 최대 길이는 128자입니다. 대문자 및 소문자, 숫자, 하이픈(-) 및 밑줄(_)을 포함할 수 있습니다.
2. 다른 모든 구성 옵션에는 기본값을 그대로 둘 수 있습니다.
3. 다음을 선택합니다.

5단계: 검토 및 생성

검토 및 생성 페이지에서 구성 과정을 검토합니다. 변경해야 하는 경우 편집을 선택합니다 입력이 끝나면 리소스 생성을 선택합니다.

6단계: 작업의 출력 보기

작업의 출력을 보려면 다음을 수행합니다.

1. 탐색 창에서 작업을 선택합니다.
2. 작업 대기열 드롭다운에서 자습서를 위해 생성한 작업 대기열을 선택합니다.
3. 작업 테이블에는 모든 작업과 현재 상태가 나열됩니다. 작업의 상태가 성공이 되면 작업 이름을 선택하여 작업의 세부 정보를 봅니다.
4. 세부 정보 창에서 로그 스트림 이름을 선택합니다. 작업에 대한 CloudWatch 콘솔이 열리고 `hello world` 메시지 또는 사용자 지정 메시지가 포함된 이벤트가 하나 있을 것입니다.

7단계: 자습서 리소스 정리

Amazon EC2 인스턴스가 활성화된 동안 요금이 부과됩니다. 요금이 발생하지 않도록 해당 인스턴스를 종료할 수 있습니다.

생성한 리소스를 삭제하려면 다음과 같이 합니다.

1. 탐색 창에서 작업 대기열을 선택합니다.
2. 작업 대기열 테이블에서 자습서를 위해 생성한 작업 대기열을 선택합니다.
3. 비활성화를 선택합니다. 작업 대기열 상태가 비활성화가 되면 삭제를 선택할 수 있습니다.

4. 작업 대기열이 삭제되면 탐색 창에서 컴퓨팅 환경을 선택합니다.
5. 이 자습서용으로 생성한 컴퓨팅 환경을 선택한 후 비활성화를 선택합니다. 컴퓨팅 환경 비활성화가 완료되는 데는 1~2분 정도 걸릴 수 있습니다.
6. 컴퓨팅 환경의 상태가 비활성화가 되면 삭제를 선택합니다. 컴퓨팅 환경이 삭제되는 데는 1~2분 정도 걸릴 수 있습니다.

추가 리소스

자습서를 완료한 후 다음과 같은 주제를 살펴보세요.

- 모범 사례에 대해 자세히 알아봅니다.
- AWS Batch 핵심 구성 요소를 살펴봅니다. 자세한 내용은 의 구성 요소 AWS Batch 단원을 참조하십시오.
- AWS Batch에서 사용할 수 있는 다양한 컴퓨팅 환경에 대해 자세히 알아봅니다.
- 작업 대기열 및 다양한 예약 옵션에 대해 자세히 알아봅니다.
- 작업 정의 및 다양한 구성 옵션에 대해 자세히 알아봅니다.
- 다른 작업 유형에 대해 자세하 알아봅니다.

를 사용하여 AWS Batch 및 Fargate 시작하기 AWS CLI

이 자습서에서는 AWS Batch AWS Fargate 오케스트레이션으로 설정하고 AWS Command Line Interface ()를 사용하여 간단한 "Hello World" 작업을 실행하는 방법을 보여줍니다AWS CLI. 컴퓨팅 환경, 작업 대기열, 작업 정의를 생성하고 AWS Batch에 작업을 제출하는 방법을 알아봅니다.

주제

- 사전 조건
- IAM 실행 역할 생성
- 컴퓨팅 환경 생성
- 작업 대기열 생성
- 작업 정의 생성
- 작업 제출 및 모니터링
- 작업 출력 보기
- 리소스 정리

- 프로덕션으로 이동
- 다음 단계

사전 조건

이 튜토리얼을 시작하기 전에 다음 사항을 확인해야 합니다.

1. AWS CLI. 설치해야 하는 경우 AWS CLI 설치 안내서를 따르세요. 가 포함된를 사용할 AWS CloudShell 수도 있습니다 AWS CLI.
2. 적절한 자격 증명 AWS CLI 으로를 구성했습니다. 자격 증명을 아직 설정하지 않은 경우 `aws configure`를 실행합니다.
3. 명령줄 인터페이스 및 컨테이너화 개념에 대한 기본 지식.
4. AWS Batch 에서 IAM을 사용하는 방법에서 AWS Batch 리소스, IAM 역할 및 VPC 리소스를 생성하고 관리합니다 AWS 계정.
5. 에 있는 VPC의 서브넷 ID 및 보안 그룹 ID입니다 AWS 계정. VPC 인증서가 없는 경우 하나를 생성할 수 있습니다. 를 사용하여 이러한 리소스 IDs를 검색 AWS CLI 하는 방법에 대한 자세한 내용은 AWS CLI 명령 참조의 describe-subnets 및 describe-security-groups를 참조하세요.

소요 시간: 이 자습서를 완료하는 데 약 15~20분이 소요됩니다.

비용: 이 자습서에서는 Fargate 컴퓨팅 리소스를 사용합니다. 정리 지침에 따라 완료 직후 리소스를 삭제한다고 가정했을 때, 이 자습서를 완료하는 데 대한 예상 비용은 0.01 USD 미만입니다. Fargate 요금은 vCPU 및 사용된 메모리 리소스를 기준으로 하며, 초당 요금(최소 1분 동안)이 부과됩니다. 요금에 대한 자세한 내용은 AWS Fargate 요금을 참조하세요.

IAM 실행 역할 생성

AWS Batch 에는 Amazon Elastic Container Service(Amazon ECS) 에이전트가 사용자를 대신하여 AWS API를 호출할 수 있는 실행 역할이 필요합니다. 이 역할은 Fargate 태스크가 컨테이너 이미지를 가져오고 Amazon CloudWatch에 로그를 쓰는 데 필요합니다.

신뢰 정책 문서 생성

먼저, Amazon EKS 태스크 서비스가 이 역할을 수임하도록 허용하는 신뢰 정책을 생성합니다.

```
cat > batch-execution-role-trust-policy.json << EOF
```

```
{
  "Version":"2012-10-17",
  "Statement": [
    {
      "Sid": "",
      "Effect": "Allow",
      "Principal": {
        "Service": "ecs-tasks.amazonaws.com"
      },
      "Action": "sts:AssumeRole"
    }
  ]
}
EOF
```

실행 역할 생성

다음 명령은 방금 생성한 신뢰 정책을 사용해 BatchEcsTaskExecutionRoleTutorial이라는 이름의 IAM 역할을 생성합니다.

```
aws iam create-role \
    --role-name BatchEcsTaskExecutionRoleTutorial \
    --assume-role-policy-document file://batch-execution-role-trust-policy.json
```

필수 정책 연결

Amazon ECS 작업 실행에 필요한 권한을 제공하는 AWS 관리형 정책을 연결합니다.

```
aws iam attach-role-policy \
    --role-name BatchEcsTaskExecutionRoleTutorial \
    --policy-arn arn:aws:iam::aws:policy/service-role/AmazonECSTaskExecutionRolePolicy
```

이제이 역할을 AWS Batch Fargate 작업 실행에 사용할 준비가 되었습니다.

컴퓨팅 환경 생성

컴퓨팅 환경은 배치 작업이 실행될 컴퓨팅 리소스를 정의합니다. 이 자습서에서는 작업 요구 사항을 기반으로 리소스를 자동으로 프로비저닝하고 확장하는 관리형 Fargate 컴퓨팅 환경을 생성합니다.

컴퓨팅 환경 생성

다음 명령은 Fargate 컴퓨팅 환경을 생성합니다. 사전 조건 섹션에 따라 예시의 서브넷 및 보안 그룹 ID를 실제 값으로 바꿉니다.

```
aws batch create-compute-environment \
    --compute-environment-name my-fargate-compute-env \
    --type MANAGED \
    --state ENABLED \
    --compute-resources type=FARGATE,maxvCpus=128,subnets=subnet-
a123456b,securityGroupIds=sg-a12b3456
```

다음은 명령이 성공적으로 실행되었을 때 출력이 어떻게 표시되는지 보여줍니다.

```
{
    "computeEnvironmentName": "my-fargate-compute-env",
    "computeEnvironmentArn": "arn:aws:batch:us-west-2:123456789012:compute-environment/
my-fargate-compute-env"
}
```

컴퓨팅 환경이 준비될 때까지 대기

계속하기 전에 컴퓨팅 환경의 상태를 확인하여 준비가 되었는지 확인합니다.

```
aws batch describe-compute-environments \
    --compute-environments my-fargate-compute-env \
    --query 'computeEnvironments[0].status'
```

```
"VALID"
```

상태가 VALID로 표시되면 컴퓨팅 환경이 작업을 수락할 준비가 된 것입니다.

작업 대기열 생성

작업 대기열은 AWS Batch 스케줄러가 컴퓨팅 환경의 리소스에서 작업을 실행할 때까지 제출된 작업을 저장합니다. 작업은 대기열 내에서 우선 순위에 따라 처리됩니다.

작업 대기열 생성

다음 명령은 Fargate 컴퓨팅 환경을 사용하는, 우선 순위 900의 작업 대기열을 생성합니다.

```
aws batch create-job-queue \
    --job-queue-name my-fargate-job-queue \
    --state ENABLED \
    --priority 900 \
    --compute-environment-order order=1,computeEnvironment=my-fargate-compute-env
```

다음은 명령이 성공적으로 실행되었을 때 출력이 어떻게 표시되는지 보여줍니다.

```
{
    "jobQueueName": "my-fargate-job-queue",
    "jobQueueArn": "arn:aws:batch:us-west-2:123456789012:job-queue/my-fargate-job-
queue"
}
```

작업 대기열이 준비되었는지 확인

작업 대기열이 ENABLED 상태이고 작업을 수락할 준비가 되었는지 확인합니다.

```
aws batch describe-job-queues \
    --job-queues my-fargate-job-queue \
    --query 'jobQueues[0].state' "ENABLED"
```

작업 정의 생성

작업 정의는 사용할 Docker 이미지, 리소스 요구 사항 및 기타 파라미터를 포함하여 작업을 실행하는 방법을 지정합니다. Fargate에는 기존 vCPU 및 메모리 파라미터 대신 리소스 요구 사항을 사용합니다.

작업 정의 생성

다음 명령은 busybox 컨테이너 이미지를 사용하여 간단한 'Hello World' 명령을 실행하는 작업 정의를 생성합니다. 를 실제 AWS 계정 ID123456789012로 바꾸고 예제를 자신의 AWS 리전 것으로 바꿉니다.

```
aws batch register-job-definition \
    --job-definition-name my-fargate-job-def \
    --type container \
    --platform-capabilities FARGATE \
    --container-properties '{
        "image": "busybox",
        "resourceRequirements": [
```

```
            {"type": "VCPU", "value": "0.25"},
            {"type": "MEMORY", "value": "512"}
        ],
        "command": ["echo", "hello world"],
        "networkConfiguration": {
            "assignPublicIp": "ENABLED"
        },
        "executionRoleArn": "arn:aws:iam::123456789012:role/
BatchEcsTaskExecutionRoleTutorial"
    },
{
    "jobDefinitionName": "my-fargate-job-def",
    "jobDefinitionArn": "arn:aws:batch:us-west-2:123456789012:job-definition/my-
fargate-job-def:1",
    "revision": 1
}'
```

작업 정의는 Fargate 태스크의 최소 리소스인 0.25 vCPU 및 512MB 메모리를 지정합니다. 컨테이너가 Docker Hub에서 busybox 이미지를 가져올 수 있도록 `assignPublicIp` 설정이 활성화됩니다.

작업 제출 및 모니터링

이제 필요한 구성 요소가 모두 있으므로 대기열에 작업을 제출하고 진행 상황을 모니터링할 수 있습니다.

작업 제출

다음 명령은 생성한 작업 정의를 사용하여 작업을 대기열에 제출합니다.

```
aws batch submit-job \
    --job-name my-hello-world-job \
    --job-queue my-fargate-job-queue \
    --job-definition my-fargate-job-def
```

다음은 명령이 성공적으로 실행되었을 때 출력이 어떻게 표시되는지 보여줍니다.

```
{
    "jobArn": "arn:aws:batch:us-west-2:123456789012:job/my-hello-world-job",
    "jobName": "my-hello-world-job",
    "jobId": "1509xmpl-4224-4da6-9ba9-1d1acc96431a"
}
```

응답에서 반환된 jobId를 기록해 두세요. 이는 작업 진행 상황을 모니터링하는 데 사용됩니다.

작업 상태 모니터링

작업 ID를 사용하여 작업의 상태를 확인합니다. 작업 진행 상황은 SUBMITTED, PENDING, RUNNABLE, STARTING, RUNNING 상태를 거쳐 마지막으로 SUCCEEDED 또는 FAILED 상태에 도달합니다.

```
aws batch describe-jobs --jobs 1509xmpl-4224-4da6-9ba9-1d1acc96431a
```

다음은 명령이 성공적으로 실행되었을 때 출력이 어떻게 표시되는지 보여줍니다.

```
{
    "jobs": [
        {
            "jobArn": "arn:aws:batch:us-west-2:123456789012:job/my-hello-world-job",
            "jobName": "my-hello-world-job",
            "jobId": "1509xmpl-4224-4da6-9ba9-1d1acc96431a",
            "jobQueue": "arn:aws:batch:us-west-2:123456789012:job-queue/my-fargate-job-
queue",
            "status": "SUCCEEDED",
            "createdAt": 1705161908000,
            "jobDefinition": "arn:aws:batch:us-west-2:123456789012:job-definition/my-
fargate-job-def:1"
        }
    ]
}
```

상태가 SUCCEEDED로 표시되면 작업이 성공적으로 완료된 것입니다.

작업 출력 보기

작업이 완료되면 Amazon CloudWatch Logs에서 출력을 볼 수 있습니다.

로그 스트림 이름 찾기

먼저 작업 세부 정보에서 로그 스트림 이름을 검색합니다. 예시 작업 ID를 실제 ID로 바꿉니다.

```
aws batch describe-jobs --jobs 1509xmpl-4224-4da6-9ba9-1d1acc96431a \
    --query 'jobs[0].attempts[0].containers[0].logStreamName' \
    --output text
```

```
my-fargate-job-def/default/1509xmpl-4224-4da6-9ba9-1d1acc96431a
```

작업 로그 보기

로그 스트림 이름을 사용하여 CloudWatch Logs에서 작업의 출력을 검색합니다.

```
aws logs get-log-events \
    --log-group-name /aws/batch/job \
    --log-stream-name my-fargate-job-def/default/1509xmpl-4224-4da6-9ba9-1d1acc96431a \
    --query 'events[*].message' \
    --output text
```

출력에는 작업이 성공적으로 실행되었음을 확인해 주는 'Hello World'가 표시됩니다.

리소스 정리

요금이 계속 청구되지 않게 하려면 이 자습서에서 생성한 리소스를 정리합니다. 종속성으로 인해 리소스를 올바른 순서로 삭제해야 합니다.

작업 대기열 비활성화 및 삭제

먼저 작업 대기열을 비활성화한 다음 삭제합니다.

```
aws batch update-job-queue \
    --job-queue my-fargate-job-queue \
    --state DISABLED
```

```
aws batch delete-job-queue \
    --job-queue my-fargate-job-queue
```

컴퓨팅 환경 비활성화 및 삭제

작업 대기열이 삭제된 후 컴퓨팅 환경을 비활성화하고 삭제합니다.

```
aws batch update-compute-environment \
    --compute-environment my-fargate-compute-env \
    --state DISABLED
```

```
aws batch delete-compute-environment \
```

```
    --compute-environment my-fargate-compute-env
```

IAM 역할 정리

정책 연결을 제거하고 IAM 역할을 삭제합니다.

```
aws iam detach-role-policy \
    --role-name BatchEcsTaskExecutionRoleTutorial \
    --policy-arn arn:aws:iam::aws:policy/service-role/AmazonECSTaskExecutionRolePolicy
```

```
aws iam delete-role \
    --role-name BatchEcsTaskExecutionRoleTutorial
```

임시 파일 제거

생성한 신뢰 정책 파일을 삭제합니다.

```
rm batch-execution-role-trust-policy.json
```

모든 리소스가 성공적으로 정리되었습니다.

프로덕션으로 이동

이 자습서는가 Fargate에서 AWS Batch 작동하는 방식을 이해하는 데 도움이 되도록 설계되었습니다. 프로덕션 배포의 경우 다음과 같은 추가 요구 사항을 고려하세요.

보안 고려 사항:

- 기본 보안 그룹을 사용하는 대신, 최소화된 필수 액세스 권한을 가진 전용 보안 그룹을 생성합니다.
- 컨테이너에 대한 퍼블릭 IP 할당 대신 NAT 게이트웨이와 함께 프라이빗 서브넷을 사용합니다.
- 퍼블릭 리포지토리를 사용하는 대신 Amazon ECR에 컨테이너 이미지를 저장합니다.
- AWS 서비스 통신을 위한 VPC 엔드포인트를 구현하여 인터넷 트래픽 방지

아키텍처 고려 사항

- 고가용성을 위해 여러 가용 영역에 배포합니다.
- 오류 처리를 위해 작업 재시도 전략 및 Dead Letter Queue(DLQ)를 구현합니다.

- 워크로드 관리를 위해 우선순위가 다른 여러 작업 대기열을 사용합니다.
- 대기열 깊이 및 리소스 사용률을 기반으로 오토 스케일링 정책을 구성합니다.
- 작업 실패 및 리소스 사용률에 대한 모니터링 및 알림을 구현합니다.

운영 고려 사항:

- 모니터링을 위한 CloudWatch 대시보드 및 경보를 설정합니다.
- 적절한 로깅 및 감사 추적을 구현합니다.
- 코드형 인프라 AWS CDK 에 CloudFormation 또는 사용
- 백업 및 재해 복구 절차를 수립합니다.

프로덕션에서 바로 사용할 수 있는 아키텍처에 대한 포괄적인 지침은 AWS Well-Architected 프레임워크 및 AWS 보안 모범 사례를 참조하세요.

다음 단계

이제이 자습서를 완료했으므로 고급 AWS Batch 기능을 살펴볼 수 있습니다.

- 작업 대기열 - 작업 대기열 예약 및 우선 순위 관리에 대해 알아봅니다.
- 작업 정의 - 환경 변수, 볼륨 및 재시도 전략을 포함한 고급 작업 정의 구성을 살펴봅니다.
- 컴퓨팅 환경 AWS Batch - 다양한 컴퓨팅 환경 유형과 규모 조정 옵션에 대해 이해합니다.
- 다중 노드 병렬 작업 - 여러 컴퓨팅 노드에 걸쳐 작업을 실행합니다.
- 배열 작업 - 많은 수의 유사한 작업을 효율적으로 제출합니다.
- 에 대한 모범 사례 AWS Batch - 프로덕션 워크로드를 위한 최적화 기법에 대해 알아봅니다.

Amazon EKS AWS Batch 에서 시작하기

AWS Batch Amazon EKS의는 배치 워크로드를 기존 Amazon EKS 클러스터로 예약 및 확장하는 관리형 서비스입니다. AWS Batch 는 사용자를 대신하여 Amazon EKS 클러스터의 수명 주기 작업을 생성, 관리 또는 수행하지 않습니다. AWS Batch 오케스트레이션은에서 관리하는 노드를 확장 및 축소 AWS Batch 하고 해당 노드에서 포드를 실행합니다.

AWS Batch 는 Amazon EKS 클러스터 내의 AWS Batch 컴퓨팅 환경과 연결되지 않은 노드, 오토 스케일링 노드 그룹 또는 포드 수명 주기를 터치하지 않습니다. 가 효과적으로 작동 AWS Batch 하려면 서

비스 연결 역할에 기존 Amazon EKS 클러스터의 Kubernetes 역할 기반 액세스 제어(RBAC) 권한이 필요합니다. 자세한 내용은 Kubernetes 설명서의 RBAC 승인 사용을 참조하세요.

AWS Batch 에는 포드의 범위를 AWS Batch 작업으로 지정할 수 있는 Kubernetes 네임스페이스가 필요합니다. 포 AWS Batch 드를 다른 클러스터 워크로드와 격리하려면 전용 네임스페이스를 사용하는 것이 좋습니다.

AWS Batch 에 RBAC 액세스 권한이 부여되고 네임스페이스가 설정되면 CreateComputeEnvironment API 작업을 사용하여 해당 Amazon EKS 클러스터를 AWS Batch 컴퓨팅 환경에 연결할 수 있습니다. 작업 대기열은이 새로운 Amazon EKS 컴퓨팅 환경과 연결할 수 있습니다. AWS Batch 작업은 SubmitJob API 작업을 사용하여 Amazon EKS 작업 정의를 기반으로 작업 대기열에 제출됩니다. AWS Batch 그런 다음 Kubernetes는 AWS Batch 관리형 노드를 시작하고 작업 대기열의 작업을 포드로 AWS Batch 컴퓨팅 환경과 연결된 EKS 클러스터에 배치합니다.

다음 섹션에서는 Amazon EKS AWS Batch 에서를 설정하는 방법을 다룹니다.

목차

개요

이 자습서에서는 AWS CLIkubectl 및를 사용하여 Amazon EKS AWS Batch 로를 설정하는 방법을 보여줍니다eksctl.

대상

이 자습서는 AWS Batch의 설정, 테스트 및 배포를 담당하는 시스템 관리자 및 개발자를 위해 설계되었습니다.

사용된 기능

이 자습서에서는를 사용하여 다음을 AWS CLI수행하는 방법을 보여줍니다.

- Amazon EKS 컴퓨팅 환경 생성 및 구성
- 작업 대기열을 생성합니다.
- 작업 정의 생성
- 실행할 작업 생성 및 제출
- 재정의를 통한 작업 제출

필요한 시간

이 자습서를 완료하는 데 약 30~40분이 소요됩니다.

리전별 제한 사항

이 솔루션의 사용과 관련된 국가별 또는 리전별 제한은 없습니다.

리소스 사용 비용

AWS 계정 생성에는 요금이 부과되지 않습니다. 그러나 이 솔루션을 구현하면 아래 표에 나열된 비용이 일부 또는 전부 발생할 수 있습니다.

설명	비용(USD)
클러스터 시간에 따라 요금이 청구됩니다.	인스턴스에 따라 다릅니다. Amazon EKS 요금을 참조하세요.

사전 조건

이 자습서를 시작하기 전에 및 AWS Batch Amazon EKS 리소스를 모두 생성하고 관리하는 데 필요한 다음 도구와 리소스를 설치하고 구성해야 합니다.

- AWS CLI - Amazon EKS를 비롯한 AWS 서비스를 사용한 작업을 위한 명령줄 도구입니다. 이 가이드에서는 버전 2.8.6 이상 또는 1.26.0 이상을 사용해야 합니다. 자세한 내용은 AWS Command

Line Interface 사용 설명서의 AWS CLI의 설치, 업데이트, 제거를 참조하세요. 설치 후 도 구성하는 것이 AWS CLI좋습니다. 자세한 내용은 AWS Command Line Interface 사용 설명서에서 aws configure를 사용한 빠른 구성을 참조하세요.

- **kubectl** - Kubernetes 클러스터 작업을 위한 명령줄 도구입니다. 이 가이드에서는 버전 `1.23` 이상을 사용해야 합니다. 자세한 내용은 Amazon EKS 사용 설명서의 kubectl 설치 또는 업데이트를 참조하세요.
- **eksctl** - 많은 개별 태스크를 자동화하는 Amazon EKS 클러스터를 사용하기 위한 명령줄 도구입니다. 이 가이드에서는 버전 `0.115.0` 이상을 사용해야 합니다. 자세한 내용은 Amazon EKS 사용 설명서의 eksctl 설치 또는 업데이트를 참조하세요.
- 필수 IAM 권한 - 사용 중인 IAM 보안 주체는 Amazon EKS IAM 역할 및 서비스 연결 역할 CloudFormation, VPC 및 관련 리소스를 사용할 수 있는 권한이 있어야 합니다. 자세한 내용은 IAM 사용 설명서의 Amazon Elastic Kubernetes Service에 사용되는 작업, 리소스 및 조건 키와 서비스 연결 역할 사용을 참조하세요. 이 가이드의 모든 단계를 동일한 사용자로 완료해야 합니다.
- 권한 - Amazon EKS 리소스를 사용하는 컴퓨팅 환경을 생성하기 위해 CreateComputeEnvironment API 작업을 호출하는 사용자는 `eks:DescribeCluster` API 작업에 대한 권한이 필요합니다.
- AWS 계정 number - AWS 계정 ID를 알아야 합니다. AWS 계정 ID 보기의 지침을 따릅니다.
- (선택 사항) CloudWatch - (선택 사항) 재정의로 작업 제출의 세부 정보를 검사하려면 로깅을 구성해야 합니다. 자세한 내용은 CloudWatch Logs를 사용하여 Amazon EKS 작업 AWS Batch 에서 모니터링 단원을 참조하십시오.

1단계: 용 Amazon EKS 클러스터 생성 AWS Batch

> **Important**
>
> 이 자습서에는 가능한 한 간단하고 빠르게 시작하기 위해 기본 설정을 사용하는 단계가 포함되어 있습니다. 프로덕션 용도로 생성하기 전에 모든 설정을 숙지하고 요구 사항을 충족하는 설정으로 배포하는 것이 좋습니다.

사전 조건을 설치한 후에는 `eksctl`을 사용하여 클러스터를 생성해야 합니다. 클러스터 생성에는 10~15분이 소요될 수 있습니다.

```
$  eksctl create cluster --name my-cluster-name --region region-code
```

앞의 명령에서 다음 항목을 교체합니다.

- *new-cluster-name*을 새 클러스터에 사용할 이름으로 바꿉니다.
- *region-code*를 로 바꾸 AWS 리전 어 예를 들어에서 클러스터를 생성합니다us-west-2.

클러스터 이름과 리전은 이 자습서의 뒷부분에서 필요합니다.

2단계: Amazon EKS 클러스터 준비 AWS Batch

필요한 단계는 다음과 같습니다.

1. AWS Batch 작업을 위한 전용 네임스페이스 생성

   kubectl을 사용하여 새 네임스페이스를 생성합니다.

   ```
   $ namespace=my-aws-batch-namespace
   ```

   ```
   $ cat - <<EOF | kubectl create -f -
   {
     "apiVersion": "v1",
     "kind": "Namespace",
     "metadata": {
       "name": "${namespace}",
       "labels": {
         "name": "${namespace}"
       }
     }
   }
   EOF
   ```

   출력:

   ```
   namespace/my-aws-batch-namespace created
   ```

2. 역할 기반 액세스 제어(RBAC)를 통한 액세스 활성화

   kubectl를 사용하여가 노드와 포드를 감시하고 Kubernetes 역할을 바인딩할 수 AWS Batch 있도록 클러스터에 대한 역할을 생성합니다. 이 작업은 각 EKS 클러스터마다 한 번씩 수행해야 합니다.

   ```
   $ cat - <<EOF | kubectl apply -f -
   apiVersion: rbac.authorization.k8s.io/v1
   ```

```
kind: ClusterRole
metadata:
  name: aws-batch-cluster-role
rules:
  - apiGroups: [""]
    resources: ["namespaces"]
    verbs: ["get"]
  - apiGroups: [""]
    resources: ["nodes"]
    verbs: ["get", "list", "watch"]
  - apiGroups: [""]
    resources: ["pods"]
    verbs: ["get", "list", "watch"]
  - apiGroups: [""]
    resources: ["events"]
    verbs: ["list"]
  - apiGroups: [""]
    resources: ["configmaps"]
    verbs: ["get", "list", "watch"]
  - apiGroups: ["apps"]
    resources: ["daemonsets", "deployments", "statefulsets", "replicasets"]
    verbs: ["get", "list", "watch"]
  - apiGroups: ["rbac.authorization.k8s.io"]
    resources: ["clusterroles", "clusterrolebindings"]
    verbs: ["get", "list"]
---
apiVersion: rbac.authorization.k8s.io/v1
kind: ClusterRoleBinding
metadata:
  name: aws-batch-cluster-role-binding
subjects:
- kind: User
  name: aws-batch
  apiGroup: rbac.authorization.k8s.io
roleRef:
  kind: ClusterRole
  name: aws-batch-cluster-role
  apiGroup: rbac.authorization.k8s.io
EOF
```

출력:

```
clusterrole.rbac.authorization.k8s.io/aws-batch-cluster-role created
```

```
clusterrolebinding.rbac.authorization.k8s.io/aws-batch-cluster-role-binding created
```

3. 에 대한 네임스페이스 범위 Kubernetes 역할을 생성 AWS Batch 하여 포드를 관리하고 수명 주기로 바인딩합니다. 이 작업은 각 고유 네임스페이스마다 한 번씩 수행해야 합니다.

```
$ namespace=my-aws-batch-namespace
```

```
$ cat - <<EOF | kubectl apply -f - --namespace "${namespace}"
apiVersion: rbac.authorization.k8s.io/v1
kind: Role
metadata:
  name: aws-batch-compute-environment-role
  namespace: ${namespace}
rules:
  - apiGroups: [""]
    resources: ["pods"]
    verbs: ["create", "get", "list", "watch", "delete", "patch"]
  - apiGroups: [""]
    resources: ["serviceaccounts"]
    verbs: ["get", "list"]
  - apiGroups: ["rbac.authorization.k8s.io"]
    resources: ["roles", "rolebindings"]
    verbs: ["get", "list"]
---
apiVersion: rbac.authorization.k8s.io/v1
kind: RoleBinding
metadata:
  name: aws-batch-compute-environment-role-binding
  namespace: ${namespace}
subjects:
- kind: User
  name: aws-batch
  apiGroup: rbac.authorization.k8s.io
roleRef:
  kind: Role
  name: aws-batch-compute-environment-role
  apiGroup: rbac.authorization.k8s.io
EOF
```

출력:

```
role.rbac.authorization.k8s.io/aws-batch-compute-environment-role created
```

```
rolebinding.rbac.authorization.k8s.io/aws-batch-compute-environment-role-binding
 created
```

4. Kubernetes aws-auth 구성 맵을 업데이트하여 이전 RBAC 권한을 AWS Batch 서비스 연결 역할에 매핑합니다.

   다음 명령에서 다음 항목을 교체합니다.

   - *<your-account-number>*를 AWS 계정 번호로 바꿉니다.

```
$ eksctl create iamidentitymapping \
    --cluster my-cluster-name \
    --arn "arn:aws:iam::<your-account-number>:role/AWSServiceRoleForBatch" \
    --username aws-batch
```

   출력:

```
2022-10-25 20:19:57 [#]  adding identity "arn:aws:iam::<your-account-number>:role/
AWSServiceRoleForBatch" to auth ConfigMap
```

> **Note**
>
> 서비스 연결 역할의 ARN에서 aws-service-role/batch.amazonaws.com/ 경로가 제거되었습니다. 이는 구성 맵에 aws-auth 문제가 있기 때문입니다. 자세한 내용은 경로가 aws-authconfigmap의 ARN에 포함될 때 경로가 있는 역할이 작동하지 않는 경우를 참조하세요.

3단계: Amazon EKS 컴퓨팅 환경 생성

AWS Batch 컴퓨팅 환경은 배치 워크로드 요구 사항에 맞게 컴퓨팅 리소스 파라미터를 정의합니다. 관리형 컴퓨팅 환경에서 AWS Batch 를 사용하면 Amazon EKS 클러스터 내에서 컴퓨팅 리소스(Kubernetes 노드)의 용량 및 인스턴스 유형을 관리할 수 있습니다. 이는 컴퓨팅 환경을 생성할 때 사용자가 정의한 컴퓨팅 리소스 사양을 기반으로 합니다. 사용자는 EC2 온디맨드 인스턴스 또는 EC2 스팟 인스턴스를 선택할 수 있습니다.

이제 AWSServiceRoleForBatch 서비스 연결 역할이 Amazon EKS 클러스터에 액세스할 수 있으므로 AWS Batch 리소스를 생성할 수 있습니다. 먼저 Amazon EKS 클러스터를 가리키는 컴퓨팅 환경을 생성합니다.

- subnets에는 `eksctl get cluster` *my-cluster-name*를 실행하여 클러스터에서 사용하는 서브넷을 가져옵니다.
- securityGroupIds 파라미터에는 Amazon EKS 클러스터와 동일한 보안 그룹을 사용할 수 있습니다. 이 명령은 클러스터의 보안 그룹 ID를 검색합니다.

```
$ aws eks describe-cluster \
    --name my-cluster-name \
    --query cluster.resourcesVpcConfig.clusterSecurityGroupId
```

- instanceRole은 클러스터를 생성할 때 생성됩니다. AmazonEKSWorkerNodePolicy을 찾으려면 instanceRole 정책을 사용하는 모든 엔터티를 나열합니다.

```
$  aws iam list-entities-for-policy --policy-arn arn:aws:iam::aws:policy/
AmazonEKSWorkerNodePolicy
```

정책 역할의 이름에는 eksctl-*my-cluster-name*-nodegroup-example을 생성한 클러스터의 이름이 포함됩니다.

instanceRole arn을 찾으려면 다음 명령을 실행합니다.

```
$  aws iam list-instance-profiles-for-role --role-name eksctl-my-cluster-name-
nodegroup-example
```

출력:

```
INSTANCEPROFILES        arn:aws:iam::<your-account-number>:instance-profile/
eks-04cb2200-94b9-c297-8dbe-87f12example
```

자세한 내용은 Amazon EKS 사용 설명서의 Amazon EKS 노드 IAM 역할 생성 및 클러스터에 대한 IAM 엔터티 액세스 활성화를 참조하세요. 포드 네트워킹을 사용하는 경우 Amazon EKS 사용 설명서의 서비스 계정에 IAM 역할을 사용하도록 Kubernetes에 Amazon VPC CNI 플러그인 구성을 참조하세요.

```
$ cat <<EOF > ./batch-eks-compute-environment.json
```

```
{
  "computeEnvironmentName": "My-Eks-CE1",
  "type": "MANAGED",
  "state": "ENABLED",
  "eksConfiguration": {
    "eksClusterArn": "arn:aws:eks:region-code:your-account-number:cluster/my-cluster-name",
    "kubernetesNamespace": "my-aws-batch-namespace"
  },
  "computeResources": {
    "type": "EC2",
    "allocationStrategy": "BEST_FIT_PROGRESSIVE",
    "minvCpus": 0,
    "maxvCpus": 128,
    "instanceTypes": [
        "m5"
    ],
    "subnets": [
        "<eks-cluster-subnets-with-access-to-internet-for-image-pull>"
    ],
    "securityGroupIds": [
        "<eks-cluster-sg>"
    ],
    "instanceRole": "<eks-instance-profile>"
  }
}
EOF
```

```
$ aws batch create-compute-environment --cli-input-json file://./batch-eks-compute-environment.json
```

참고

- Amazon EKS 컴퓨팅 환경의 유지 관리는 공동 책임입니다. 자세한 내용은 Kubernetes 노드에 대한 공동 책임 단원을 참조하십시오.

4단계: 작업 대기열 생성 및 컴퓨팅 환경 연결

> **Important**
>
> 진행하기 전에 컴퓨팅 환경이 정상인지 확인하는 것이 중요합니다. DescribeComputeEnvironments API 작업을 사용하여 이 작업을 수행할 수 있습니다.
>
> ```
> $ aws batch describe-compute-environments --compute-environments My-Eks-CE1
> ```
>
> status 파라미터가 INVALID로 되어 있지 않은지 확인합니다. 그럴 경우 statusReason 파라미터에서 원인을 확인합니다. 자세한 내용은 문제 해결 AWS Batch 단원을 참조하십시오.

이 새 작업 대기열에 제출된 작업은 컴퓨팅 환경과 연결된 Amazon EKS 클러스터에 조인된 AWS Batch 관리형 노드에서 포드로 실행됩니다.

```
$ cat <<EOF > ./batch-eks-job-queue.json
 {
    "jobQueueName": "My-Eks-JQ1",
    "priority": 10,
    "computeEnvironmentOrder": [
      {
        "order": 1,
        "computeEnvironment": "My-Eks-CE1"
      }
    ]
  }
EOF
```

```
$ aws batch create-job-queue --cli-input-json file://./batch-eks-job-queue.json
```

5단계: 작업 정의 생성

다음 작업 정의는 포드가 60초 동안 절전 모드로 전환하도록 지시합니다.

```
$ cat <<EOF > ./batch-eks-job-definition.json
{
  "jobDefinitionName": "MyJobOnEks_Sleep",
  "type": "container",
  "eksProperties": {
```

```
    "podProperties": {
      "hostNetwork": true,
      "containers": [
        {
          "image": "public.ecr.aws/amazonlinux/amazonlinux:2",
          "command": [
            "sleep",
            "60"
          ],
          "resources": {
            "limits": {
              "cpu": "1",
              "memory": "1024Mi"
            }
          }
        }
      ],
      "metadata": {
        "labels": {
          "environment": "test"
        }
      }
    }
  }
}
EOF
```

```
$ aws batch register-job-definition --cli-input-json file://./batch-eks-job-
definition.json
```

참고

- cpu및 memory 파라미터에 대한 고려 사항이 있습니다. 자세한 내용은 Amazon EKS의 AWS Batch에 대한 메모리 및 vCPU 고려 사항 단원을 참조하십시오.

6단계: 작업 제출

다음 AWS CLI 명령을 실행하여 새 작업을 제출합니다.

```
$ aws batch submit-job --job-queue My-Eks-JQ1 \
    --job-definition MyJobOnEks_Sleep --job-name My-Eks-Job1
```

작업의 상태를 확인하려면:

```
$ aws batch describe-jobs --job <jobId-from-submit-response>
```

참고

- Amazon EKS 리소스에서 작업을 실행하는 방법에 대한 자세한 내용은 Amazon EKS 작업 섹션을 참조하세요.

7단계: 작업 출력 보기

작업의 출력을 보려면 다음을 수행합니다.

1. https://console.aws.amazon.com/batch/ AWS Batch 콘솔을 엽니다.
2. 탐색 창에서 작업을 선택합니다.
3. 작업 대기열 드롭다운에서 자습서를 위해 생성한 작업 대기열을 선택합니다.
4. 작업 테이블에는 모든 작업과 현재 상태가 나열됩니다. 작업 상태가 성공이 되면 작업의 이름에 *My-Eks-JQ1*을 선택하여 작업의 세부 정보를 봅니다.
5. 세부 정보 창에서 시작 시간과 중지 시간의 간격은 1분이어야 합니다.

8단계: (선택 사항) 재정의를 통한 작업 제출

이 작업은 container. AWS Batch aggressively에 전달된 명령을 재정의합니다. 작업이 완료된 후 포드를 정리하여에 대한 부하를 줄입니다Kubernetes. 작업의 세부 정보를 검사하려면 로깅을 구성해야 합니다. 자세한 내용은 CloudWatch Logs를 사용하여 Amazon EKS 작업 AWS Batch 에서 모니터링 단원을 참조하십시오.

```
$ cat <<EOF > ./submit-job-override.json
{
  "jobName": "EksWithOverrides",
  "jobQueue": "My-Eks-JQ1",
  "jobDefinition": "MyJobOnEks_Sleep",
  "eksPropertiesOverride": {
    "podProperties": {
      "containers": [
        {
          "command": [
            "/bin/sh"
```

```
          ],
          "args": [
            "-c",
            "echo hello world"
          ]
        }
      ]
    }
  }
}
EOF
```

```
$ aws batch submit-job --cli-input-json file://./submit-job-override.json
```

참고

- 작업 세부 정보에 대한 가시성을 높이려면 Amazon EKS 컨트롤 플레인 로깅을 활성화합니다. 자세한 내용을 알아보려면 Amazon EKS 사용 설명서의 Amazon EKS 클러스터 컨트롤 플레인 로깅을 참조하세요.
- Daemonsets 및 kubelets 오버헤드는 사용 가능한 vCPU 및 메모리 리소스, 특히 규모 조정 및 작업 배치에 영향을 줍니다. 자세한 내용은 Amazon EKS의 AWS Batch에 대한 메모리 및 vCPU 고려 사항 단원을 참조하십시오.

작업의 출력을 보려면 다음을 수행합니다.

1. https://console.aws.amazon.com/batch/ AWS Batch 콘솔을 엽니다.
2. 탐색 창에서 작업을 선택합니다.
3. 작업 대기열 드롭다운에서 자습서를 위해 생성한 작업 대기열을 선택합니다.
4. 작업 테이블에는 모든 작업과 현재 상태가 나열됩니다. 작업의 상태가 성공이 되면 작업 이름을 선택하여 작업의 세부 정보를 봅니다.
5. 세부 정보 창에서 로그 스트림 이름을 선택합니다. 작업에 대한 CloudWatch 콘솔이 열리고 `hello world` 메시지 또는 사용자 지정 메시지가 포함된 이벤트가 하나 있을 것입니다.

9단계: 자습서 리소스 정리

Amazon EC2 인스턴스가 활성화된 동안 요금이 부과됩니다. 요금이 발생하지 않도록 해당 인스턴스를 종료할 수 있습니다.

생성한 리소스를 삭제하려면 다음과 같이 합니다.

1. https://console.aws.amazon.com/batch/ AWS Batch 콘솔을 엽니다.
2. 탐색 창에서 작업 대기열을 선택합니다.
3. 작업 대기열 테이블에서 자습서를 위해 생성한 작업 대기열을 선택합니다.
4. 비활성화를 선택합니다. 작업 대기열 상태가 비활성화가 되면 삭제를 선택할 수 있습니다.
5. 작업 대기열이 삭제되면 탐색 창에서 컴퓨팅 환경을 선택합니다.
6. 이 자습서용으로 생성한 컴퓨팅 환경을 선택한 후 비활성화를 선택합니다. 컴퓨팅 환경 비활성화가 완료되는 데는 1~2분 정도 걸릴 수 있습니다.
7. 컴퓨팅 환경의 상태가 비활성화가 되면 삭제를 선택합니다. 컴퓨팅 환경이 삭제되는 데는 1~2분 정도 걸릴 수 있습니다.

추가 리소스

자습서를 완료한 후 다음과 같은 주제를 살펴보세요.

- 모범 사례에 대해 자세히 알아봅니다.
- AWS Batch 핵심 구성 요소를 살펴봅니다. 자세한 내용은 의 구성 요소 AWS Batch 단원을 참조하십시오.
- AWS Batch에서 사용할 수 있는 다양한 컴퓨팅 환경에 대해 자세히 알아봅니다.
- 작업 대기열 및 다양한 예약 옵션에 대해 자세히 알아봅니다.
- 작업 정의 및 다양한 구성 옵션에 대해 자세히 알아봅니다.
- 다른 작업 유형에 대해 자세하 알아봅니다.

Amazon EKS 프라이빗 클러스터 AWS Batch 에서 시작하기

AWS Batch 는 Amazon Elastic Kubernetes Service(Amazon EKS) 클러스터에서 배치 워크로드를 오케스트레이션하는 관리형 서비스입니다. 여기에는 대기열, 종속성 추적, 관리되는 작업 재시도 및 우선 순위 관리, 포드 관리, 노드 조정 등이 포함됩니다. 이 기능은 기존 프라이빗 Amazon EKS 클러스터 AWS Batch 를와 연결하여 대규모로 작업을 실행합니다. `eksctl` (Amazon EKS용 명령줄 인터페이스), AWS 콘솔 또는를 사용하여 다른 모든 필수 리소스와 함께 프라이빗 Amazon EKS 클러스터를 AWS Command Line Interface 생성할 수 있습니다.

기본적으로 Amazon EKS 프라이빗 전용 클러스터는 인바운드/아웃바운드 인터넷 액세스 권한이 없으며 VPC 또는 연결된 네트워크 내에서만 API 서버에 액세스할 수 있습니다. Amazon VPC 엔드포인트

는 다른 AWS 서비스에 대한 프라이빗 액세스를 활성화하는 데 사용됩니다. `eksctl`은 기존 Amazon VPC 및 서브넷을 사용하여 완전한 프라이빗 클러스터 생성을 지원합니다. 또한 `eksctl`은 제공된 Amazon VPC에서 Amazon VPC 엔드포인트를 생성하고 제공된 서브넷의 라우팅 테이블을 수정합니다.

`eksctl`은 기본 라우팅 테이블을 수정하지 않으므로 각 서브넷에는 연결된 명시적 라우팅 테이블이 있어야 합니다. 클러스터에서는 Amazon VPC에 있는 컨테이너 레지스트리에서 이미지를 가져와야 합니다. 또한 Amazon VPC에서 Amazon Elastic Container Registry를 생성하고 노드에서 가져올 컨테이너 이미지를 복사할 수 있습니다. 자세한 내용은 한 리포지토리에서 다른 리포지토리로 컨테이너 이미지 복사를 참조하세요. Amazon ECR 프라이빗 리포지토리를 시작하려면 Amazon ECR 프라이빗 리포지토리를 참조하세요.

선택적으로 Amazon ECR을 사용하여 풀스루 캐시 규칙을 생성할 수 있습니다. 외부 퍼블릭 레지스트리에 대한 풀스루 캐시 규칙이 생성되면 Amazon ECR 프라이빗 레지스트리 URI(Uniform Resource Identifier)를 사용하여 해당 외부 퍼블릭 레지스트리에서 이미지를 가져올 수 있습니다. 그러면 Amazon ECR이 리포지토리를 생성하고 해당 이미지를 캐시합니다. Amazon ECR 프라이빗 레지스트리 URI를 사용하여 캐시된 이미지를 가져오면 Amazon ECR은 원격 레지스트리를 점검하여 이미지의 새 버전이 있는지 확인하며, 최대 24시간마다 한 번씩 프라이빗 레지스트리를 업데이트합니다.

목차

개요

이 자습서에서는 AWS CloudShell, `kubectl` 및 AWS Batch 를 사용하여 프라이빗 Amazon EKS로 설정하는 방법을 보여줍니다`eksctl`.

대상

이 자습서는 AWS Batch의 설정, 테스트 및 배포를 담당하는 시스템 관리자 및 개발자를 위해 설계되었습니다.

사용된 기능

이 자습서에서는를 사용하여 다음을 AWS CLI수행하는 방법을 보여줍니다.

- Amazon Elastic Container Registry(Amazon ECR)를 사용하여 컨테이너 이미지 저장
- Amazon EKS 컴퓨팅 환경 생성 및 구성
- 작업 대기열을 생성합니다.
- 작업 정의 생성
- 실행할 작업 생성 및 제출
- 재정의를 통한 작업 제출

필요한 시간

이 자습서를 완료하는 데 약 40~50분이 소요됩니다.

리전별 제한 사항

이 솔루션의 사용과 관련된 국가별 또는 리전별 제한은 없습니다.

리소스 사용 비용

AWS 계정 생성에는 요금이 부과되지 않습니다. 그러나 이 솔루션을 구현하면 아래 표에 나열된 비용이 일부 또는 전부 발생할 수 있습니다.

설명	비용(USD)
클러스터 시간에 따라 요금이 청구됩니다.	인스턴스에 따라 다릅니다. Amazon EKS 요금을 참조하세요.
Amazon EC2 인스턴스	생성된 각 Amazon EC2 인스턴스에 대해 비용을 지불합니다. 요금에 대한 자세한 정보는 Amazon EC2 요금을 참조하십시오.

사전 조건

이 자습서에서는에서 직접 시작하는 브라우저 기반 사전 인증된 AWS CloudShell 셸인를 사용합니다 AWS Management Console. 이는 더 이상 퍼블릭 인터넷에 액세스할 수 없게 된 클러스터에 대한 액세스를 허용합니다. AWS CLI, `kubectl`및가 이미의 일부로 설치되어 있을 `eksctl` 수 있습니다 AWS CloudShell. 에 대한 자세한 내용은 AWS CloudShell사용 설명서를 AWS CloudShell참조하세요. 의 대안 AWS CloudShell 은 클러스터의 VPC 또는 연결된 네트워크에 연결하는 것입니다.

kubectl 명령을 실행하려면 Amazon EKS 클러스터에 대한 프라이빗 액세스 권한이 필요합니다. 즉, 클러스터 API 서버에 대한 모든 트래픽은 클러스터의 VPC 또는 연결된 네트워크 내에서 비롯되어야 합니다.

- AWS CLI - Amazon EKS를 포함한 AWS 서비스 작업을 위한 명령줄 도구입니다. 이 가이드에서는 버전 2.8.6 이상 또는 1.26.0 이상을 사용해야 합니다. 자세한 내용은 AWS Command Line Interface 사용 설명서의 AWS CLI의 설치, 업데이트, 제거를 참조하세요. 설치 후 도 구성하는 것이 AWS CLI 좋습니다. 자세한 내용은 AWS Command Line Interface 사용 설명서에서 `aws configure`를 사용한 빠른 구성을 참조하세요.
- **`kubectl`** - Kubernetes 클러스터 작업을 위한 명령줄 도구입니다. 이 가이드에서는 버전 `1.23` 이상을 사용해야 합니다. 자세한 내용은 Amazon EKS 사용 설명서의 `kubectl` 설치 또는 업데이트를 참조하세요.
- **`eksctl`** - 많은 개별 태스크를 자동화하는 Amazon EKS 클러스터를 사용하기 위한 명령줄 도구입니다. 이 가이드에서는 버전 `0.115.0` 이상을 사용해야 합니다. 자세한 내용은 Amazon EKS 사용 설명서의 `eksctl` 설치 또는 업데이트를 참조하세요.
- 권한 - Amazon EKS 리소스를 사용하는 컴퓨팅 환경을 생성하기 위해 CreateComputeEnvironment API 작업을 호출하는 사용자는 `eks:DescribeCluster` 및 `eks:ListClusters` API 작업에 대한 권한이 필요합니다. IAM 사용 설명서의 IAM 자격 증명 권한 추가 및 제거 지침에 따라 AWSBatchFullAccess 관리형 정책을 사용자 계정에 연결할 수 있습니다.
- InstanceRole - `AmazonEKSWorkerNodePolicy` 및 `AmazonEC2ContainerRegistryPullOnly` 정책이 있는 Amazon EKS 노드에 대한 `InstanceRole`을 생성해야 합니다. `InstanceRole`을 생성하는 방법에 대한 지침은 Amazon EKS 노드 IAM 역할 생성을 참조하세요. `InstanceRole`의 ARN이 필요합니다.
- AWS 계정 ID - AWS 계정 ID를 알아야 합니다. AWS 계정 ID 보기의 지침을 따릅니다.
- (선택 사항) CloudWatch - (선택 사항) 재정의로 작업 제출의 세부 정보를 검사하려면 로깅을 구성해야 합니다. 자세한 내용은 CloudWatch Logs를 사용하여 Amazon EKS 작업 AWS Batch 에서 모니터링 단원을 참조하십시오.

1단계: 용 EKS 클러스터 생성 AWS Batch

> **Important**
>
> 이 자습서에는 가능한 한 간단하고 빠르게 시작하기 위해 기본 설정을 사용하는 단계가 포함되어 있습니다. 프로덕션 용도로 생성하기 전에 모든 설정을 숙지하고 요구 사항을 충족하는 설정으로 배포하는 것이 좋습니다.

`eksctl` 및 다음 구성 파일을 사용하여 클러스터를 생성하는 것이 좋습니다. 수동으로 클러스터를 설정하려면 Amazon EKS 사용 설명서의 인터넷 액세스가 제한된 프라이빗 클러스터 배포의 지침을 따르세요.

1. AWS CloudShell 콘솔을 열고 리전을 `us-east-1`으로 설정합니다. 자습서의 나머지 부분에서는 `us-east-1`을 사용하고 있는지 확인하세요.
2. 샘플 `eksctl` 구성 파일을 사용하여 `us-east-1` 리전에 프라이빗 EKS 클러스터를 생성합니다. yaml 파일을 AWS CloudShell 환경에 저장하고 이름을 `clusterConfig.yaml`로 지정합니다. *`my-test-cluster`*는 클러스터에 사용할 이름으로 변경할 수 있습니다.

   ```
   kind: ClusterConfig
   apiVersion: eksctl.io/v1alpha5
   metadata:
       name: my-test-cluster
       region: us-east-1
   availabilityZones:
       - us-east-1a
       - us-east-1b
       - us-east-1c
   managedNodeGroups:
       - name: ng-1
         privateNetworking: true
   privateCluster:
       enabled: true
       skipEndpointCreation: false
   ```

3. `eksctl create cluster -f clusterConfig.yaml` 명령을 사용하여 리소스를 생성합니다. 스택 생성에는 10~15분이 소요될 수 있습니다.
4. 클러스터 생성이 완료되면 허용 목록에 AWS CloudShell IP 주소를 추가해야 합니다. AWS CloudShell IP 주소를 찾으려면 다음 명령을 실행합니다.

```
curl http://checkip.amazonaws.com
```

퍼블릭 IP 주소를 찾았으면 허용 목록 규칙을 생성해야 합니다.

```
aws eks update-cluster-config \
  --name my-test-cluster \
  --region us-east-1 \
  --resources-vpc-config
 endpointPublicAccess=true,endpointPrivateAccess=true,publicAccessCidrs=["<Public
 IP>/32"]
```

그런 다음 kubectl 구성 파일에 업데이트를 적용합니다.

```
aws eks update-kubeconfig --name my-test-cluster --region us-east-1
```

5. 노드에 액세스할 수 있는지 테스트하려면 다음 명령을 실행합니다.

```
kubectl get nodes
```

명령의 출력을 다음과 같습니다.

```
NAME                                STATUS   ROLES    AGE     VERSION
ip-192-168-107-235.ec2.internal   Ready    none     1h   v1.32.3-eks-473151a
ip-192-168-165-40.ec2.internal    Ready    none     1h   v1.32.3-eks-473151a
ip-192-168-98-54.ec2.internal     Ready    none     1h   v1.32.1-eks-5d632ec
```

2단계: 용 EKS 클러스터 준비 AWS Batch

모든 단계가 필요하며에서 수행해야 합니다 AWS CloudShell.

1. AWS Batch 작업을 위한 전용 네임스페이스 생성

   kubectl을 사용하여 새 네임스페이스를 생성합니다.

```
$ namespace=my-aws-batch-namespace
```

```
$ cat - <<EOF | kubectl create -f -
{
```

```
  "apiVersion": "v1",
  "kind": "Namespace",
  "metadata": {
    "name": "${namespace}",
    "labels": {
      "name": "${namespace}"
    }
  }
}
EOF
```

출력:

```
namespace/my-aws-batch-namespace created
```

2. 역할 기반 액세스 제어(RBAC)를 통한 액세스 활성화

kubectl을 사용하여 AWS Batch 가 노드와 포드를 감시하도록 하는 클러스터에 대한 Kubernetes 역할을 생성하고 역할을 바인딩합니다. 이 작업은 각 Amazon EKS 클러스터마다 한 번씩 수행해야 합니다.

```
$ cat - <<EOF | kubectl apply -f -
apiVersion: rbac.authorization.k8s.io/v1
kind: ClusterRole
metadata:
  name: aws-batch-cluster-role
rules:
  - apiGroups: [""]
    resources: ["namespaces"]
    verbs: ["get"]
  - apiGroups: [""]
    resources: ["nodes"]
    verbs: ["get", "list", "watch"]
  - apiGroups: [""]
    resources: ["pods"]
    verbs: ["get", "list", "watch"]
  - apiGroups: [""]
    resources: ["events"]
    verbs: ["list"]
  - apiGroups: [""]
    resources: ["configmaps"]
    verbs: ["get", "list", "watch"]
```

```
  - apiGroups: ["apps"]
    resources: ["daemonsets", "deployments", "statefulsets", "replicasets"]
    verbs: ["get", "list", "watch"]
  - apiGroups: ["rbac.authorization.k8s.io"]
    resources: ["clusterroles", "clusterrolebindings"]
    verbs: ["get", "list"]
---
apiVersion: rbac.authorization.k8s.io/v1
kind: ClusterRoleBinding
metadata:
  name: aws-batch-cluster-role-binding
subjects:
- kind: User
  name: aws-batch
  apiGroup: rbac.authorization.k8s.io
roleRef:
  kind: ClusterRole
  name: aws-batch-cluster-role
  apiGroup: rbac.authorization.k8s.io
EOF
```

출력:

```
clusterrole.rbac.authorization.k8s.io/aws-batch-cluster-role created
clusterrolebinding.rbac.authorization.k8s.io/aws-batch-cluster-role-binding created
```

에 대한 네임스페이스 범위 Kubernetes 역할을 생성 AWS Batch 하여 포드를 관리하고 수명 주기로 바인딩합니다. 이 작업은 각 고유 네임스페이스마다 한 번씩 수행해야 합니다.

```
$ namespace=my-aws-batch-namespace
```

```
$ cat - <<EOF | kubectl apply -f - --namespace "${namespace}"
apiVersion: rbac.authorization.k8s.io/v1
kind: Role
metadata:
  name: aws-batch-compute-environment-role
  namespace: ${namespace}
rules:
  - apiGroups: [""]
    resources: ["pods"]
    verbs: ["create", "get", "list", "watch", "delete", "patch"]
```

```
  - apiGroups: [""]
    resources: ["serviceaccounts"]
    verbs: ["get", "list"]
  - apiGroups: ["rbac.authorization.k8s.io"]
    resources: ["roles", "rolebindings"]
    verbs: ["get", "list"]
---
apiVersion: rbac.authorization.k8s.io/v1
kind: RoleBinding
metadata:
  name: aws-batch-compute-environment-role-binding
  namespace: ${namespace}
subjects:
- kind: User
  name: aws-batch
  apiGroup: rbac.authorization.k8s.io
roleRef:
  kind: Role
  name: aws-batch-compute-environment-role
  apiGroup: rbac.authorization.k8s.io
EOF
```

출력:

```
role.rbac.authorization.k8s.io/aws-batch-compute-environment-role created
rolebinding.rbac.authorization.k8s.io/aws-batch-compute-environment-role-binding
 created
```

Kubernetes `aws-auth` 구성 맵을 업데이트하여 이전 RBAC 권한을 AWS Batch 서비스 연결 역할에 매핑합니다.

```
$ eksctl create iamidentitymapping \
    --cluster my-test-cluster \
    --arn "arn:aws:iam::<your-account-ID>:role/AWSServiceRoleForBatch" \
    --username aws-batch
```

출력:

```
2022-10-25 20:19:57 [#]  adding identity "arn:aws:iam::<your-account-ID>:role/
AWSServiceRoleForBatch" to auth ConfigMap
```

> **Note**
>
> 서비스 연결 역할의 ARN에서 `aws-service-role/batch.amazonaws.com/` 경로가 제거되었습니다. 이는 구성 맵에 `aws-auth` 문제가 있기 때문입니다. 자세한 내용은 경로가 aws-authconfigmap의 ARN에 포함될 때 경로가 있는 역할이 작동하지 않는 경우를 참조하세요.

3단계: Amazon EKS 컴퓨팅 환경 생성

AWS Batch 컴퓨팅 환경은 배치 워크로드 요구 사항에 맞게 컴퓨팅 리소스 파라미터를 정의합니다. 관리형 컴퓨팅 환경에서 AWS Batch 를 사용하면 Amazon EKS 클러스터 내에서 컴퓨팅 리소스(Kubernetes 노드)의 용량 및 인스턴스 유형을 관리할 수 있습니다. 이는 컴퓨팅 환경을 생성할 때 사용자가 정의한 컴퓨팅 리소스 사양을 기반으로 합니다. 사용자는 EC2 온디맨드 인스턴스 또는 EC2 스팟 인스턴스를 선택할 수 있습니다.

이제 AWSServiceRoleForBatch 서비스 연결 역할이 Amazon EKS 클러스터에 액세스할 수 있으므로 AWS Batch 리소스를 생성할 수 있습니다. 먼저 Amazon EKS 클러스터를 가리키는 컴퓨팅 환경을 생성합니다.

- `subnets`에는 `eksctl get cluster` *`my-test-cluster`*를 실행하여 클러스터에서 사용하는 서브넷을 가져옵니다.
- `securityGroupIds` 파라미터에는 Amazon EKS 클러스터와 동일한 보안 그룹을 사용할 수 있습니다. 이 명령은 클러스터의 보안 그룹 ID를 검색합니다.

```
$ aws eks describe-cluster \
    --name my-test-cluster \
    --query cluster.resourcesVpcConfig.clusterSecurityGroupId
```

- 사전 조건에서 생성한 `instanceRole`의 ARN을 사용합니다.

```
$ cat <<EOF > ./batch-eks-compute-environment.json
{
  "computeEnvironmentName": "My-Eks-CE1",
  "type": "MANAGED",
  "state": "ENABLED",
  "eksConfiguration": {
    "eksClusterArn": "arn:aws:eks:us-east-1:<your-account-ID>:cluster/my-test-cluster",
```

```
    "kubernetesNamespace": "my-aws-batch-namespace"
  },
  "computeResources": {
    "type": "EC2",
    "allocationStrategy": "BEST_FIT_PROGRESSIVE",
    "minvCpus": 0,
    "maxvCpus": 128,
    "instanceTypes": [
        "m5"
    ],
    "subnets": [
        "<eks-cluster-subnets-with-access-to-the-image-for-image-pull>"
    ],
    "securityGroupIds": [
        "<eks-cluster-sg>"
    ],
    "instanceRole": "<eks-instance-profile>"
  }
}
EOF
```

```
$ aws batch create-compute-environment --cli-input-json file://./batch-eks-compute-environment.json
```

참고

- Amazon EKS 컴퓨팅 환경의 유지 관리는 공동 책임입니다. 자세한 내용은 Amazon EKS의 보안을 참조하세요.

4단계: 작업 대기열 생성 및 컴퓨팅 환경 연결

> **Important**
>
> 진행하기 전에 컴퓨팅 환경이 정상인지 확인하는 것이 중요합니다.
> DescribeComputeEnvironments API 작업을 사용하여 이 작업을 수행할 수 있습니다.
>
> ```
> $ aws batch describe-compute-environments --compute-environments My-Eks-CE1
> ```

`status` 파라미터가 `INVALID`로 되어 있지 않은지 확인합니다. 그럴 경우 `statusReason` 파라미터에서 원인을 확인합니다. 자세한 내용은 문제 해결 AWS Batch 단원을 참조하십시오.

이 새 작업 대기열에 제출된 작업은 컴퓨팅 환경과 연결된 Amazon EKS 클러스터에 조인된 AWS Batch 관리형 노드에서 포드로 실행됩니다.

```
$ cat <<EOF > ./batch-eks-job-queue.json
 {
    "jobQueueName": "My-Eks-JQ1",
    "priority": 10,
    "computeEnvironmentOrder": [
      {
        "order": 1,
        "computeEnvironment": "My-Eks-CE1"
      }
    ]
  }
EOF
```

```
$ aws batch create-job-queue --cli-input-json file://./batch-eks-job-queue.json
```

5단계: 풀스루 캐시를 사용하여 Amazon ECR 생성

클러스터에는 퍼블릭 인터넷 액세스 권한이 없으므로 컨테이너 이미지를 위한 Amazon ECR을 생성해야 합니다. 다음 지침은 풀스루 캐시 규칙을 사용하여 이미지를 저장하는 Amazon ECR을 생성합니다.

1. 다음 명령은 풀스루 캐시 규칙을 생성합니다. *tutorial-prefix*는 다른 접두사로 바꿀 수 있습니다.

   ```
   aws ecr create-pull-through-cache-rule \
       --ecr-repository-prefix "my-prefix" \
       --upstream-registry-url "public.ecr.aws" \
       --region us-east-1
   ```

2. 퍼블릭 ECR로 인증합니다.

   ```
   aws ecr get-login-password --region us-east-1 | docker login --username AWS --
   password-stdin <your-account-ID>.dkr.ecr.us-east-1.amazonaws.com
   ```

이제 이미지를 가져올 수 있습니다.

```
docker pull <your-account-ID>.dkr.ecr.us-east-1.amazonaws.com/my-prefix/
amazonlinux/amazonlinux:2
```

3. 다음 명령을 실행하여 리포지토리와 이미지를 확인할 수 있습니다.

```
aws ecr describe-repositories
```

```
aws ecr describe-images --repository-name my-prefix/amazonlinux/amazonlinux
```

4. 컨테이너를 가져오는 데 사용할 이미지 문자열의 형식은 다음과 같습니다.

```
<your-account-ID>.dkr.ecr.us-east-1.amazonaws.com/my-prefix/amazonlinux/
amazonlinux:2
```

6단계: 작업 정의 등록

다음 작업 정의는 포드가 60초 동안 절전 모드로 전환하도록 지시합니다.

작업 정의의 이미지 필드에서 퍼블릭 ECR 리포지토리의 이미지에 대한 링크를 제공하는 대신 프라이빗 ECR 리포지토리에 저장된 이미지에 대한 링크를 제공합니다. 다음 샘플 작업 정의를 참조하세요.

```
$ cat <<EOF > ./batch-eks-job-definition.json
{
  "jobDefinitionName": "MyJobOnEks_Sleep",
  "type": "container",
  "eksProperties": {
    "podProperties": {
      "hostNetwork": true,
      "containers": [
        {
          "image": "<your-account-ID>.dkr.ecr.us-east-1.amazonaws.com/my-prefix/
amazonlinux/amazonlinux:2",
          "command": [
            "sleep",
            "60"
          ],
          "resources": {
            "limits": {
```

```
              "cpu": "1",
              "memory": "1024Mi"
            }
          }
        }
      ],
      "metadata": {
        "labels": {
          "environment": "test"
        }
      }
    }
  }
}
EOF
```

```
$ aws batch register-job-definition --cli-input-json file://./batch-eks-job-definition.json
```

참고

- cpu및 memory 파라미터에 대한 고려 사항이 있습니다. 자세한 내용은 Amazon EKS의 AWS Batch에 대한 메모리 및 vCPU 고려 사항 단원을 참조하십시오.

7단계: 실행할 작업 제출

에서 다음 AWS CLI 명령을 실행 AWS CloudShell 하여 새 작업을 제출하고 고유한 JobID를 반환합니다.

```
$ aws batch submit-job --job-queue My-Eks-JQ1 \
    --job-definition MyJobOnEks_Sleep - -job-name My-Eks-Job1
```

참고

- Amazon EKS 리소스에서 작업을 실행하는 방법에 대한 자세한 내용은 Amazon EKS 작업 섹션을 참조하세요.

8단계: 작업의 출력 보기

작업의 상태를 확인하려면:

```
$ aws batch describe-jobs --job <JobID-from-submit-response>
```

`startedAt`과 `stoppedAt`은 1분 간격이어야 합니다.

9단계: (선택 사항) 재정의를 통한 작업 제출

이 작업은 컨테이너로 전달되는 명령을 재정의합니다.

```
$ cat <<EOF > ./submit-job-override.json
{
  "jobName": "EksWithOverrides",
  "jobQueue": "My-Eks-JQ1",
  "jobDefinition": "MyJobOnEks_Sleep",
  "eksPropertiesOverride": {
    "podProperties": {
      "containers": [
        {
          "command": [
            "/bin/sh"
          ],
          "args": [
            "-c",
            "echo hello world"
          ]
        }
      ]
    }
  }
}
EOF
```

```
$ aws batch submit-job - -cli-input-json file://./submit-job-override.json
```

참고

- 작업 세부 정보에 대한 가시성을 높이려면 Amazon EKS 컨트롤 플레인 로깅을 활성화합니다. 자세한 내용을 알아보려면 Amazon EKS 사용 설명서의 Amazon EKS 클러스터 컨트롤 플레인 로깅을 참조하세요.
- Daemonsets 및 kubelets 오버헤드는 사용 가능한 vCPU 및 메모리 리소스, 특히 규모 조정 및 작업 배치에 영향을 줍니다. 자세한 내용은 Amazon EKS의 AWS Batch에 대한 메모리 및 vCPU 고려 사항 단원을 참조하십시오.

10단계: 자습서 리소스 정리

Amazon EC2 인스턴스가 활성화된 동안 요금이 부과됩니다. 요금이 발생하지 않도록 해당 인스턴스를 종료할 수 있습니다.

생성한 리소스를 삭제하려면 다음과 같이 합니다.

1. https://console.aws.amazon.com/batch/ AWS Batch 콘솔을 엽니다.
2. 탐색 창에서 작업 대기열을 선택합니다.
3. 작업 대기열 테이블에서 자습서를 위해 생성한 작업 대기열을 선택합니다.
4. 작업에서 비활성화를 선택합니다. 작업 대기열 상태가 비활성화가 되면 삭제를 선택할 수 있습니다.
5. 작업 대기열이 삭제되면 탐색 창에서 컴퓨팅 환경을 선택합니다.
6. 이 자습서용으로 생성한 컴퓨팅 환경을 선택한 다음 작업에서 비활성화를 선택합니다. 컴퓨팅 환경 비활성화가 완료되는 데는 1~2분 정도 걸릴 수 있습니다.
7. 컴퓨팅 환경의 상태가 비활성화가 되면 삭제를 선택합니다. 컴퓨팅 환경이 삭제되는 데는 1~2분 정도 걸릴 수 있습니다.

추가 리소스

자습서를 완료한 후 다음과 같은 주제를 살펴보세요.

- 모범 사례에 대해 자세히 알아봅니다.
- AWS Batch 핵심 구성 요소를 살펴봅니다. 자세한 내용은 의 구성 요소 AWS Batch 단원을 참조하십시오.
- AWS Batch에서 사용할 수 있는 다양한 컴퓨팅 환경에 대해 자세히 알아봅니다.
- 작업 대기열 및 다양한 예약 옵션에 대해 자세히 알아봅니다.

- 작업 정의 및 다양한 구성 옵션에 대해 자세히 알아봅니다.
- 다른 작업 유형에 대해 자세하 알아봅니다.

문제 해결

에서 시작된 노드가 이미지를 저장하는 Amazon ECR 리포지토리(또는 기타 리포지토리)에 액세스할 수 AWS Batch 없는 경우 작업은 시작 상태로 유지될 수 있습니다. 포드가 이미지를 다운로드하고 AWS Batch 작업을 실행할 수 없기 때문입니다. 에서 시작한 포드 이름을 클릭하면 오류 메시지를 보고 문제를 확인할 AWS Batch 수 있습니다. 오류 메시지의 내용은 다음과 유사합니다.

```
Failed to pull image "public.ecr.aws/amazonlinux/amazonlinux:2": rpc error: code =
Unknown desc = failed to pull and unpack image
"public.ecr.aws/amazonlinux/amazonlinux:2": failed to resolve reference
"public.ecr.aws/amazonlinux/amazonlinux:2": failed to do request: Head
"https://public.ecr.aws/v2/amazonlinux/amazonlinux/manifests/2": dial tcp: i/o timeout
```

기타 일반적인 문제 해결 시나리오는 AWS Batch문제 해결을 참조하세요. 포드 상태를 기준으로 문제를 해결하려면 Amazon EKS에서 포드 상태 문제를 해결하려면 어떻게 해야 하나요?를 참조하세요.

SageMaker AI AWS Batch 에서 시작하기

AWS Batch 서비스 작업을 사용하면 예약, 우선 순위 지정 및 대기열 기능을 사용하여 AWS Batch 작업 대기열을 통해 SageMaker 훈련 작업을 제출할 수 있습니다. 이 자습서에서는 AWS Batch 서비스 작업을 사용하여 간단한 SageMaker 훈련 작업을 설정하고 실행하는 방법을 보여줍니다.

목차

개요

이 자습서에서는를 사용하여 SageMaker 훈련 작업에 대한 AWS Batch 서비스 작업을 설정하는 방법을 보여줍니다 AWS CLI.

대상

이 자습서는 대규모 기계 학습 훈련 작업의 설정 및 실행을 담당하는 데이터 과학자 및 개발자를 위해 설계되었습니다.

사용된 기능

이 자습서에서는를 사용하여 다음을 AWS CLI 수행하는 방법을 보여줍니다.

- SageMaker 훈련 작업을 위한 서비스 환경 생성
- SageMaker 훈련 작업 대기열 생성
- `SubmitServiceJob` API를 사용하여 서비스 작업 제출
- 작업 상태 모니터링 및 출력 보기
- 훈련 작업에 대한 CloudWatch 로그 액세스

필요한 시간

이 자습서를 완료하려면 약 15분 정도 걸립니다.

리전별 제한 사항

이 자습서는 AWS Batch 및 SageMaker AI를 모두 사용할 수 있는 모든 AWS 리전에서 완료할 수 있습니다.

리소스 사용 비용

AWS 계정 생성에는 요금이 부과되지 않습니다. 하지만 이 솔루션을 구현하면 다음 리소스에 대한 비용이 발생할 수 있습니다.

설명	비용(USD)
SageMaker AI 훈련 인스턴스	사용한 각 SageMaker AI 훈련 인스턴스에 대해 비용을 지불합니다. 요금에 대한 자세한 정보는 SageMaker AI 요금을 참조하세요.
Amazon S3 스토리지	훈련 작업 출력을 저장하는 데 드는 최소한의 비용. 자세한 내용은 Amazon S3 요금을 참조하세요.

사전 조건

이 자습서를 시작하기 전에 및 AWS Batch SageMaker AI 리소스를 모두 생성하고 관리하는 데 필요한 다음 도구와 리소스를 설치하고 구성해야 합니다.

- AWS CLI - AWS Batch 및 SageMaker AI를 포함한 AWS 서비스 작업을 위한 명령줄 도구입니다. 이 가이드에서는 버전 2.8.6 이상을 사용해야 합니다. 자세한 내용은 AWS Command Line Interface 사용 설명서의 AWS CLI의 설치, 업데이트, 제거를 참조하세요. 설치 후 도 구성하는 것이 AWS CLI좋습니다. 자세한 내용은 AWS Command Line Interface 사용 설명서에서 `aws configure`를 사용한 빠른 구성을 참조하세요.

1단계: SageMaker AI 실행 역할 생성

Amazon SageMaker AI는 실행 역할을 통해 사용자 대신 다른 AWS 서비스를 사용하여 작업을 수행합니다. 실행 역할을 생성하고 훈련 작업에 필요한 서비스 및 리소스를 사용할 수 있는 권한을 SageMaker AI에 부여해야 합니다. `AmazonSageMakerFullAccess` 관리형 정책에는 Amazon S3에 대한 권한이 포함되어 있으므로 이 정책을 사용합니다.

> **Note**
>
> 다음 지침에 따라 이 자습서를 위한 SageMaker AI 실행 역할을 생성합니다.
> 프로덕션 환경을 위한 실행 역할을 생성하기 전에 SageMaker AI 개발자 안내서의 SageMaker AI 실행 역할을 사용하는 방법을 검토할 것이 권장됩니다.

1. IAM 역할 생성

   다음 신뢰 정책을 통해 `sagemaker-trust-policy.json`이라는 이름의 JSON 파일을 생성합니다.

   JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Principal": {
```

```
                "Service": "sagemaker.amazonaws.com"
            },
            "Action": "sts:AssumeRole"
        }
    ]
}
```

신뢰 정책을 사용하여 IAM 역할을 생성합니다.

```
aws iam create-role \
    --role-name SageMakerExecutionRole \
    --assume-role-policy-document file://sagemaker-trust-policy.json \
    --description "Execution role for SageMaker training jobs"
```

2. 관리형 정책 연결

필요한 관리형 정책을 역할에 연결합니다.

```
aws iam attach-role-policy \
    --role-name SageMakerExecutionRole \
    --policy-arn arn:aws:iam::aws:policy/AmazonSageMakerFullAccess
```

```
aws iam attach-role-policy \
    --role-name SageMakerExecutionRole \
    --policy-arn arn:aws:iam::aws:policy/AmazonS3FullAccess
```

3. 역할 ARN 기록

역할 ARN을 가져옵니다.이 ARN은 후속 단계에서 필요합니다.

```
aws iam get-role --role-name SageMakerExecutionRole --query 'Role.Arn' --output
 text
```

이 ARN을 저장합니다. 이 ARN은 훈련 작업 페이로드를 생성할 때 사용됩니다.

2단계: 테스트 환경 생성

서비스 환경은 SageMaker 훈련 작업에 대한 용량 제약 조건을 정의합니다. 서비스 환경은 동시에 실행할 수 있는 최대 훈련 인스턴스 수를 캡슐화합니다.

> **Important**
>
> SageMaker 훈련을 위한 첫 번째 서비스 환경을 생성하면 AWS Batch 는 계정에서 `AWSServiceRoleForAWSBatchWithSagemaker`라는 서비스 연결 역할을 자동 생성합니다. 이 역할을 통해 AWS Batch 는 사용자를 대신하여 SageMaker 훈련 작업을 대기열에 넣고 관리할 수 있습니다. 서비스 연결 역할 및 해당 권한에 대한 자세한 내용은 the section called "SageMaker 통합 역할" 섹션을 참조하세요.

최대 5개의 인스턴스를 처리할 수 있는 서비스 환경을 생성합니다.

```
aws batch create-service-environment \
    --service-environment-name TutorialServiceEnvironment \
    --service-environment-type SAGEMAKER_TRAINING \
    --capacity-limits capacityUnit=NUM_INSTANCES,maxCapacity=5
```

출력:

```
{
    "serviceEnvironmentName": "TutorialServiceEnvironment",
    "serviceEnvironmentArn": "arn:aws:batch:your-region:your-account-id:service-
environment/TutorialServiceEnvironment"
}
```

서비스 환경이 성공적으로 생성되었는지 확인합니다.

```
aws batch describe-service-environments --service-
environments TutorialServiceEnvironment
```

출력:

```
{
    "serviceEnvironments": [
        {
            "serviceEnvironmentName": "TutorialServiceEnvironment",
            "serviceEnvironmentArn": "arn:aws:batch:your-region:your-account-
id:service-environment/TutorialServiceEnvironment",
            "serviceEnvironmentType": "SAGEMAKER_TRAINING",
            "state": "ENABLED",
```

```
            "status": "VALID",
            "capacityLimits": [
                {
                    "maxCapacity": 5,
                    "capacityUnit": "NUM_INSTANCES"
                }
            ],
            "tags": {}
        }
    ]
}
```

서비스 환경에 대한 자세한 내용은 서비스 환경 섹션을 참조하세요.

3단계: SageMaker 작업 대기열 생성

SageMaker 작업 대기열은 서비스 작업의 예약 및 실행을 관리합니다. 이 대기열에 제출된 작업은 사용 가능한 용량을 기반으로 서비스 환경으로 디스패치됩니다.

SageMaker 훈련 작업 대기열을 생성합니다.

```
aws batch create-job-queue \
    --job-queue-name my-sm-training-fifo-jq \
    --job-queue-type SAGEMAKER_TRAINING \
    --priority 1 \
    --service-environment-order order=1,serviceEnvironment=TutorialServiceEnvironment
```

출력:

```
{
    "jobQueueName": "my-sm-training-fifo-jq",
    "jobQueueArn": "arn:aws:batch:your-region:your-account-id:job-queue/my-sm-training-fifo-jq"
}
```

작업 대기열이 성공적으로 생성되었는지 확인합니다.

```
aws batch describe-job-queues --job-queues my-sm-training-fifo-jq
```

출력:

```
{
    "jobQueues": [
        {
            "jobQueueName": "my-sm-training-fifo-jq",
            "jobQueueArn": "arn:aws:batch:your-region:your-account-id:job-queue/my-sm-
training-fifo-jq",
            "state": "ENABLED",
            "status": "VALID",
            "statusReason": "JobQueue Healthy",
            "priority": 1,
            "computeEnvironmentOrder": [],
            "serviceEnvironmentOrder": [
                {
                    "order": 1,
                    "serviceEnvironment": "arn:aws:batch:your-region:your-account-
id:service-environment/TutorialServiceEnvironment"
                }
            ],
            "jobQueueType": "SAGEMAKER_TRAINING",
            "tags": {}
        }
    ]
}
```

SageMaker 작업 대기열에 대한 자세한 내용은 the section called “SageMaker 작업 대기열 생성” 섹션을 참조하세요.

4단계: 훈련 작업 생성 및 제출

이제 간단한 훈련 작업을 생성하여 작업 대기열에 제출합니다. 이 예제에서는 서비스 작업 기능을 보여주는 간단한 'hello world' 훈련 작업을 사용합니다.

다음 콘텐츠가 포함된 *my_training_job.json*이라는 파일을 생성합니다. *your-account-id*를 AWS 계정 ID로 바꿉니다.

> **Note**
>
> S3OutputPath는 SageMaker 훈련 작업을 생성하는 데 필요하지만 이 자습서의 결과는 Amazon S3 버킷에 저장되지 않으므로 다음 JSON의 경로를 사용할 수 있습니다. 프로덕션 환경에서는 선택한 경우 출력을 저장할 유효한 Amazon S3 버킷이 필요합니다.

```
{
    "TrainingJobName": "my-simple-training-job",
    "RoleArn": "arn:aws:iam::your-account-id:role/SageMakerExecutionRole",
    "AlgorithmSpecification": {
        "TrainingInputMode": "File",
        "TrainingImage": "763104351884.dkr.ecr.us-west-2.amazonaws.com/pytorch-
training:2.0.0-cpu-py310",
        "ContainerEntrypoint": [
            "echo",
            "hello world"
        ]
    },
    "ResourceConfig": {
        "InstanceType": "ml.c5.xlarge",
        "InstanceCount": 1,
        "VolumeSizeInGB": 1
    },
    "OutputDataConfig": {
        "S3OutputPath": "s3://your-s3-bucket/output"
    },
    "StoppingCondition": {
        "MaxRuntimeInSeconds": 30
    }
}
```

SubmitServiceJob API를 사용하여 훈련 작업을 제출합니다.

```
aws batch submit-service-job \
    --job-queue my-sm-training-fifo-jq \
    --job-name my-batch-sm-job \
    --service-job-type SAGEMAKER_TRAINING \
    --retry-strategy attempts=1 \
    --timeout-config attemptDurationSeconds=60 \
    --service-request-payload file://my_training_job.json
```

출력:

```
{
    "jobArn": "arn:aws:batch:your-region:your-account-id:service-job/your-job-id",
    "jobName": "my-batch-sm-job",
    "jobId": "your-job-id"
}
```

서비스 작업 페이로드에 대한 자세한 내용은 the section called “서비스 작업 페이로드” 섹션을 참조하세요. 서비스 작업 제출에 대한 자세한 내용은 the section called “서비스 작업 제출” 섹션을 참조하세요.

5단계: 작업 상태 모니터링

DescribeServiceJob, ListServiceJobs 및 GetJobQueueSnapshot AWS Batch APIs를 사용하여 훈련 작업을 모니터링할 수 있습니다. 이 섹션에서는 작업 상태 및 대기열 정보를 확인하는 다양한 방법을 보여줍니다.

대기열에서 실행 중인 작업을 봅니다.

```
aws batch list-service-jobs \
    --job-queue my-sm-training-fifo-jq --job-status RUNNING
```

출력:

```
{
    "jobSummaryList": [
        {
            "latestAttempt": {
                "serviceResourceId": {
                    "name": "TrainingJobArn",
                    "value": "arn:aws:sagemaker:your-region:your-account-id:training-
job/AWSBatch<my-simple-training-job><your-attempt-id>"
                }
            },
            "createdAt": 1753718760,
            "jobArn": "arn:aws:batch:your-region:your-account-id:service-job/your-job-
id",
            "jobId": "your-job-id",
            "jobName": "my-batch-sm-job",
            "serviceJobType": "SAGEMAKER_TRAINING",
            "status": "RUNNING",
            "startedAt": 1753718820
        }
    ]
}
```

RUNNABLE 상태에 있는 작업을 봅니다.

```
aws batch list-service-jobs \
```

```
    --job-queue my-sm-training-fifo-jq --job-status RUNNABLE
```

대기열에서 예정된 작업의 스냅샷을 가져옵니다.

```
aws batch get-job-queue-snapshot --job-queue my-sm-training-fifo-jq
```

출력:

```
{
    "frontOfQueue": {
        "jobs": [
            {
                "jobArn": "arn:aws:batch:your-region:your-account-id:service-job/your-job-id",
                "earliestTimeAtPosition": 1753718880
            },
            {
                "jobArn": "arn:aws:batch:your-region:your-account-id:service-job/your-job-id-2",
                "earliestTimeAtPosition": 1753718940
            }
        ],
        "lastUpdatedAt": 1753718970
    }
}
```

이름으로 작업을 검색합니다.

```
aws batch list-service-jobs \
    --job-queue my-sm-training-fifo-jq \
    --filters name=JOB_NAME,values="my-batch-sm-job"
```

출력:

```
{
    "jobSummaryList": [
        {
            "latestAttempt": {
                "serviceResourceId": {
                    "name": "TrainingJobArn",
                    "value": "arn:aws:sagemaker:your-region:your-account-id:training-job/AWSBatch<my-simple-training-job><your-attempt-id>"
```

```
                }
            },
            "createdAt": 1753718760,
            "jobArn": "arn:aws:batch:your-region:your-account-id:service-job/your-job-
id",
            "jobId": "your-job-id",
            "jobName": "my-batch-sm-job",
            "serviceJobType": "SAGEMAKER_TRAINING",
            "status": "RUNNING"
        }
    ]
}
```

작업 상태 매핑에 대한 자세한 내용은 the section called “서비스 작업 상태” 섹션을 참조하세요.

6단계: 작업 출력 보기

작업이 완료되면 AWS Batch 및 SageMaker AI APIs.

AWS Batch다음에서 작업에 대한 자세한 정보를 가져옵니다.

```
aws batch describe-service-job \
    --job-id your-job-id
```

출력:

```
{
    "attempts": [
        {
            "serviceResourceId": {
                "name": "TrainingJobArn",
                "value": "arn:aws:sagemaker:your-region:your-account-id:training-job/
AWSBatch<my-simple-training-job><your-attempt-id>"
            },
            "startedAt": 1753718820,
            "stoppedAt": 1753718880,
            "statusReason": "Received status from SageMaker: Training job completed"
        }
    ],
    "createdAt": 1753718760,
    "jobArn": "arn:aws:batch:your-region:your-account-id:service-job/your-job-id",
    "jobId": "your-job-id",
    "jobName": "my-batch-sm-job",
```

```
    "jobQueue": "arn:aws:batch:your-region:your-account-id:job-queue/my-sm-training-
fifo-jq",
    "latestAttempt": {
        "serviceResourceId": {
            "name": "TrainingJobArn",
            "value": "arn:aws:sagemaker:your-region:your-account-id:training-job/
AWSBatch<my-simple-training-job><your-attempt-id>"
        }
    },
    "retryStrategy": {
        "attempts": 1,
        "evaluateOnExit": []
    },
    "serviceRequestPayload": "your-training-job-request-json",
    "serviceJobType": "SAGEMAKER_TRAINING",
    "startedAt": 1753718820,
    "status": "SUCCEEDED",
    "statusReason": "Received status from SageMaker: Training job completed",
    "stoppedAt": 1753718880,
    "tags": {},
    "timeoutConfig": {
        "attemptDurationSeconds": 60
    }
}
```

이 명령은 SageMaker 훈련 작업 ARN 등 SageMaker AI를 통해 작업에 직접 액세스하는 데 사용할 수 있는 포괄적인 작업 정보를 반환합니다.

```
aws sagemaker describe-training-job \
    --training-job-name AWSBatch<my-simple-training-job><your-attempt-id>
```

훈련 작업에 대한 CloudWatch 로그를 보려면 먼저 로그 스트림 이름을 가져옵니다.

```
aws logs describe-log-streams \
    --log-group-name /aws/sagemaker/TrainingJobs \
    --log-stream-name-prefix AWSBatchmy-simple-training-job
```

출력:

```
{
    "logStreams": [
        {
```

```
            "logStreamName": "your-log-stream-name",
            "creationTime": 1753718830,
            "firstEventTimestamp": 1753718840,
            "lastEventTimestamp": 1753718850,
            "lastIngestionTime": 1753718860,
            "uploadSequenceToken": upload-sequence-token,
            "arn": "arn:aws:logs:your-region:your-account-id:log-group:/aws/sagemaker/
TrainingJobs:log-stream:AWSBatch<my-simple-training-job><your-attempt-id>/algo-1-algo-
id",
            "storedBytes": 0
        }
    ]
}
```

그런 다음 이전 응답의 로그 스트림 이름을 사용하여 로그를 검색합니다.

```
aws logs get-log-events \
    --log-group-name /aws/sagemaker/TrainingJobs \
    --log-stream-name your-log-stream-name
```

출력:

```
{
    "events": [
        {
            "timestamp": 1753718845,
            "message": "hello world",
            "ingestionTime": 1753718865
        }
    ],
    "nextForwardToken": "next-forward-token",
    "nextBackwardToken": "next-backward-token"
}
```

로그 출력에는 훈련 작업의 'hello world' 메시지가 표시되어 작업이 성공적으로 실행되었음을 확인해 줍니다.

7단계: 자습서 리소스 정리

이 자습서를 완료했으면 지속적인 요금이 발생하지 않도록 생성한 리소스를 정리하세요.

먼저 작업 대기열을 비활성화하고 삭제합니다.

```
aws batch update-job-queue \
    --job-queue my-sm-training-fifo-jq \
    --state DISABLED
```

작업 대기열이 비활성화될 때까지 기다린 다음 삭제합니다.

```
aws batch delete-job-queue \
    --job-queue my-sm-training-fifo-jq
```

그런 다음 서비스 환경을 비활성화하고 삭제합니다.

```
aws batch update-service-environment \
    --service-environment TutorialServiceEnvironment \
    --state DISABLED
```

서비스 환경이 비활성화될 때까지 기다린 다음 삭제합니다.

```
aws batch delete-service-environment \
    --service-environment TutorialServiceEnvironment
```

추가 리소스

자습서를 완료한 후 다음과 같은 주제를 살펴볼 수 있습니다.

- PySDK에는 헬퍼 클래스와 유틸리티가 있으므로 서비스 작업 생성 및 작업 대기열 제출에 PySDK를 사용할 것이 권장됩니다. PySDK 사용에 대한 예제는 GitHub의 SageMaker AI 예제를 참조하세요.
- the section called “서비스 작업”에 대해 자세히 알아보세요.
- 보다 복잡한 훈련 작업 구성에 대해서는 의 서비스 작업 페이로드 AWS Batch 섹션을 살펴보세요.
- 에서 서비스 작업 제출 AWS Batch 및 `SubmitServiceJob` API에 대해 알아봅니다.
- AWS Batch 서비스 작업 상태를 SageMaker AI 상태로 매핑 섹션을 검토하여 작업 상태 전환에 대해 이해합니다.
- Python을 사용하여 다양한 기능과 함께 SageMaker 훈련 작업을 생성하고 제출하는 방법은 SageMaker AI Python SDK 설명서를 참조하세요.
- 보다 복잡한 기계 학습 워크플로에 대해서는 SageMaker 예제 노트북을 살펴보세요.

AWS Batch 위젯 대시보드

AWS Batch 대시보드를 통해 최근 작업, 작업 대기열 및 컴퓨팅 환경을 모니터링할 수 있습니다. 기본적으로 다음과 같은 대시보드 위젯이 표시됩니다.

- 작업 개요- AWS Batch 작업에 대한 자세한 정보는 작업 섹션을 참조하세요.
- 작업 대기열 개요- AWS Batch 작업 대기열에 대한 자세한 정보는 작업 대기열 섹션을 참조하세요.
- 컴퓨팅 환경 개요 - AWS Batch 컴퓨팅 환경에 대한 자세한 내용은 컴퓨팅 환경 AWS Batch 섹션을 참조하세요.

사용자는 대시보드 페이지에 표시되는 위젯을 사용자 지정할 수 있습니다. 다음 섹션은 추가할 수 있는 추가 위젯에 대해 설명합니다.

주제

- AWS Batch 대시보드에 단일 작업 대기열 위젯을 추가하는 방법
- AWS Batch 대시보드에 CloudWatch Container Insights 위젯을 추가하는 방법
- AWS Batch 대시보드에 작업 로그 위젯을 추가하는 방법

AWS Batch 대시보드에 단일 작업 대기열 위젯을 추가하는 방법

단일 작업 대기열 위젯은 단일 작업 대기열에 대한 세부 정보를 표시합니다.

이 위젯을 추가하려면 다음 단계를 따르세요.

1. AWS Batch 콘솔을 엽니다.
2. 탐색 표시줄에서 원하는 AWS 리전을 선택합니다.
3. 탐색 창에서 대시보드를 선택합니다.
4. 위젯 추가를 선택합니다.
5. 단일 작업 대기열의 경우 위젯 추가를 선택합니다.
6. 작업 대기열에서 원하는 작업 대기열을 선택합니다.
7. 작업 상태에서 표시할 작업 상태를 선택합니다.
8. (선택 사항) 컴퓨팅 환경의 속성을 표시하지 않으려면 연결형 컴퓨팅 환경을 표시합니다를 끕니다.
9. 컴퓨팅 환경 속성에서 원하는 속성을 선택합니다.
10. 추가를 선택합니다.

AWS Batch 대시보드에 CloudWatch Container Insights 위젯을 추가하는 방법

이 위젯은 AWS Batch 컴퓨팅 환경 및 작업에 대한 집계된 지표를 표시합니다. Container Insights에 대한 자세한 정보는 the section called “CloudWatch Container Insights” 섹션을 참조하세요.

이 위젯을 추가하려면 다음 단계를 따르세요.

1. AWS Batch 콘솔을 엽니다.
2. 탐색 표시줄에서 원하는 AWS 리전을 선택합니다.
3. 탐색 창에서 대시보드를 선택합니다.
4. 위젯 추가를 선택합니다.
5. Container Insights에서 위젯 추가를 선택합니다.
6. 컴퓨팅 환경에서 원하는 컴퓨팅 환경을 선택합니다.
7. 추가를 선택합니다.

AWS Batch 대시보드에 작업 로그 위젯을 추가하는 방법

이 위젯은 작업의 다양한 로그를 한 곳에서 편리하게 보여줍니다. 작업 로그에 대한 자세한 내용은 the section called “CloudWatch Logs에서 작업 로그 보기” 섹션을 참조하세요.

이 위젯을 추가하려면 다음 단계를 따르세요.

1. AWS Batch 콘솔을 엽니다.
2. 탐색 표시줄에서 원하는 AWS 리전을 선택합니다.
3. 탐색 창에서 대시보드를 선택합니다.
4. 위젯 추가를 선택합니다.
5. 작업 로그에서 위젯 추가를 선택합니다.
6. 작업 ID에서 원하는 작업의 작업 ID를 입력합니다.
7. 추가를 선택합니다.

컴퓨팅 환경 AWS Batch

작업 대기열은 하나 이상의 컴퓨팅 환경에 매핑됩니다. 컴퓨팅 환경에는 컨테이너화된 배치 작업을 실행하는 데 사용되는 Amazon ECS 컨테이너 인스턴스가 있습니다. 하나의 특정 컴퓨팅 환경은 한하나 이상의 작업 대기열에 매핑될 수도 있습니다. 작업 대기열 내에서, 각각의 연결된 컴퓨팅 환경은 실행 준비가 완료된 작업을 스케줄러가 어디에 배치할지를 결정하는 데 사용하는 순서를 갖습니다. 첫 번째 컴퓨팅 환경이 `VALID`의 상태이고 사용 가능한 리소스가 있으면 작업은 해당 컴퓨팅 환경 내의 컨테이너 인스턴스에 스케줄 됩니다. 첫 번째 컴퓨팅 환경이 `INVALID`의 상태이고 컴퓨팅 환경에서 적절한 컴퓨팅 리소스를 제공할 수 없는 경우 스케줄러는 다음 컴퓨팅 환경에서 작업을 실행하기 위해 시도합니다.

주제

- 관리형 컴퓨팅 환경
- 비 관리형 컴퓨팅 환경
- 컴퓨팅 환경 생성
- 에서 컴퓨팅 환경 업데이트 AWS Batch
- 컴퓨팅 리소스 AMI
- 에서 Amazon EC2 시작 템플릿 사용 AWS Batch
- 인스턴스 메타데이터 서비스(IMDS) 구성
- EC2 구성
- 에 대한 인스턴스 유형 할당 전략 AWS Batch
- 컴퓨팅 리소스 메모리 관리
- Fargate 컴퓨팅 환경
- Amazon EKS 컴퓨팅 환경

관리형 컴퓨팅 환경

관리형 컴퓨팅 환경을 사용하여 환경 내 컴퓨팅 리소스의 용량 및 인스턴스 유형을 AWS Batch 관리할 수 있습니다. 컴퓨팅 환경을 생성할 때 사용자가 정의한 컴퓨팅 리소스 사양을 기반으로 합니다. 사용자는 Amazon EC2 온디맨드 인스턴스와 Amazon EC2 스팟 인스턴스 사용을 선택할 수 있습니다. 또는 관리형 컴퓨팅 환경에서 Fargate 및 Fargate 스팟 용량을 대신 사용할 수 있습니다. 스팟 인스턴스를 사용할 때 사용자는 선택적으로 최고 가격을 설정할 수 있습니다. 이렇게 하면 스팟 인스턴스는 스팟 인스턴스 가격이 온디맨드 가격에 대해 지정한 비율(%) 이하일 경우에만 시작합니다.

> **Important**
>
> Fargate 스팟 인스턴스는에서 지원되지 않습니다 Windows containers on AWS Fargate. Fargate Spot 컴퓨팅 환경만 사용하는 작업 대기열에 FargateWindows 작업이 제출되는 경우 작업 대기열이 차단됩니다.

> **Important**
>
> AWS Batch 는 Amazon EC2 시작 템플릿, Amazon EC2 Auto Scaling 그룹, Amazon EC2 스팟 플릿 및 Amazon Amazon EC2 클러스터를 포함하여 사용자를 대신하여 계정 내에서 여러 AWS 리소스를 생성하고 관리합니다. 이러한 관리형 리소스는 최적의 AWS Batch 작동을 보장하도록 특별히 구성됩니다. 설명서에 명시적으로 명시되지 AWS Batch 않는 한 이러한 AWS Batch관리형 리소스를 수동으로 수정하면 `INVALID` 컴퓨팅 환경, 최적화되지 않은 인스턴스 조정 동작, 워크로드 처리 지연 또는 예상치 못한 비용 등 예상치 못한 동작이 발생할 수 있습니다. 이러한 수동 수정은 AWS Batch 서비스에서 결정론적으로 지원할 수 없습니다. 항상 지원되는 AWS Batch APIs 또는 AWS Batch 콘솔을 사용하여 컴퓨팅 환경을 관리합니다.
> 지원되지 않는 수동 수정에는 자체 Amazon ECS 태스크 또는 서비스 AWS Batch관리형 Amazon ECS 클러스터 실행 또는 추가 프로세스, 데몬 또는 서비스 직접 AWS Batch관리형 인스턴스 시작이 포함됩니다. AWS Batch 는 관리형 컴퓨팅 환경에서 컴퓨팅 리소스를 완전히 제어하고 언제든지 인스턴스를 종료하거나 태스크를 중지하거나 클러스터를 확장할 수 있습니다. 이러한 관리형 리소스에서 AWS Batch 작업 제출 외부에서 실행하는 모든 워크로드는 경고 없이 중단될 수 있습니다. AWS Batch관리형 클러스터 및 인스턴스에서 워크로드가AWS Batch 아닌 클러스터 및 인스턴스를 실행하면 AWS Batch 작업 예약 및 인스턴스 조정에 방해가 될 수도 있습니다.
> Amazon ECS를 사용하는 컴퓨팅 환경의 경우 시작 템플릿 사용자 데이터에서가 AWS Batch 관리하는 Amazon ECS 에이전트 구성 값을 설정하지 마십시오. 예약 값 목록은 섹션을 참조하세요예약 Amazon ECS 에이전트 구성 값.

관리형 컴퓨팅 환경은 지정한 VPC와 서브넷에서 Amazon EC2 인스턴스를 시작한 다음 Amazon ECS 클러스터에 등록합니다. Amazon EC2 인스턴스는 Amazon ECS 서비스 엔드포인트와 통신하기 위해 외부 네트워크에 액세스해야 합니다. 일부 서브넷은 Amazon EC2 인스턴스에 퍼블릭 IP 주소를 제공하지 않습니다. Amazon EC2 인스턴스 컴퓨팅 리소스에 퍼블릭 IP 주소가 없는 경우, Network Address Translation(NAT)를 사용해야 합니다. 자세한 내용은 Amazon VPC 사용 설명서의 NAT 게이트웨이 섹션을 참조하세요. VPC 생성 방법에 대한 자세한 내용은 Virtual Private Cloud 생성 섹션을 참조하세요.

기본적으로 AWS Batch 관리형 컴퓨팅 환경은 컴퓨팅 리소스에 대해 승인된 최신 버전의 Amazon ECS 최적화 AMI를 사용합니다. 그러나 여러 가지 이유로 관리형 컴퓨팅 환경에 사용할 자체 AMI 생성을 원할 수 있습니다. 자세한 내용은 컴퓨팅 리소스 AMI 단원을 참조하십시오.

> **Note**
>
> AWS Batch 는 생성된 AMIs를 컴퓨팅 환경에서 자동으로 업그레이드하지 않습니다. 예를 들어 Amazon ECS 최적화 AMI의 최신 버전이 릴리스된 경우 컴퓨팅 환경에서 AMI를 업데이트 하지 않습니다. 게스트 운영 체제의 관리는 사용자의 책임입니다. 여기에는 모든 업데이트 및 보안 패치 적용이 포함됩니다. 또한 사용자는 컴퓨팅 리소스에 설치하는 추가 애플리케이션 소프트웨어 또는 유틸리티에 대해서도 책임이 있습니다. AWS Batch 작업에 새 AMI를 사용하는 방법에는 두 가지가 있습니다. 원래 방법은 다음 단계를 완료하는 것입니다.
>
> 1. 새 AMI로 새 컴퓨팅 환경을 생성합니다.
> 2. 기존 작업 대기열에 컴퓨팅 환경을 추가합니다.
> 3. 작업 대기열에서 이전 컴퓨팅 환경을 제거합니다.
> 4. 이전 컴퓨팅 환경을 삭제합니다.
>
> 2022년 4월에 컴퓨팅 환경 업데이트에 대한 향상된 지원이 AWS Batch 추가되었습니다. 자세한 내용은 에서 컴퓨팅 환경 업데이트 AWS Batch 단원을 참조하십시오. 컴퓨팅 환경의 향상된 업데이트를 사용하여 AMI를 업데이트하려면 다음 규칙을 따르세요.
>
> - 서비스 역할(`serviceRole`) 파라미터를 설정하지 않거나 AWSServiceRoleForBatch 서비스 연결 역할로 설정하세요.
> - 할당 전략(`allocationStrategy`) 파라미터를 `BEST_FIT_PROGRESSIVE`, `SPOT_CAPACITY_OPTIMIZED`, 또는 `SPOT_PRICE_CAPACITY_OPTIMIZED`로 설정합니다.
> - 최신 이미지 버전으로 업데이트(`updateToLatestImageVersion`) 파라미터를 `true`으로 설정합니다.
> - `imageId`, `imageIdOverride`(`ec2Configuration`) 또는 시작 템플릿(`launchTemplate`)에 AMI ID를 지정하지 마세요. 이 경우는 인프라 업데이트가 시작될 AWS Batch 때에서 지원하는 최신 Amazon ECS 최적화 AMI를 AWS Batch 선택합니다. `imageId` 또는 `imageIdOverride` 파라미터에 AMI ID를 지정하거나 `LaunchTemplate` 속성으로 식별되는 시작 템플릿을 지정할 수도 있습니다. 이러한 속성을 변경하면 인프라 업데이트가 시작됩니다. 시작 템플릿에 AMI ID가 지정되면 `imageId` 또는 `imageIdOverride` 파라미터에 AMI ID를 지정한다고 이를 바꿀 수 없습니다. 다른 시

작 템플릿을 지정해야만 교체할 수 있습니다. 또는, 시작 템플릿 버전이 `$Default` 혹은 `$Latest`로 설정되면 시작 템플릿의 새 기본 버전을 설정하거나(`$Default`인 경우), 새로운 버전의 시작 템플릿을 추가하여(`$Latest`인 경우) 대체할 수 있습니다.

이러한 규칙을 준수하면 인프라 업데이트를 시작하는 업데이트 시 AMI ID가 다시 선택됩니다. 시작 템플릿(`launchTemplate`)의 `version` 설정이 `$Latest` 또는 `$Default`으로 설정된 경우, `launchTemplate`이 업데이트되어 있지 않더라도 인프라 업데이트 시 시작 템플릿의 최신 버전 또는 기본 버전이 평가됩니다.

다중 노드 병렬 작업 생성 시 고려 사항

AWS Batch 에서는 다중 노드 병렬(MNP) 작업 및 비 MNP 작업을 실행하기 위한 전용 컴퓨팅 환경을 생성할 것을 권장합니다. 이는 관리형 컴퓨팅 환경에서 컴퓨팅 용량이 생성되는 방식 때문입니다. 새 관리형 컴퓨팅 환경을 생성할 때 0보다 큰 `minvCpu` 값을 지정하면는 비 MNP 작업에만 사용할 인스턴스 풀을 AWS Batch 생성합니다. 다중 노드 병렬 작업이 제출되면는 다중 노드 병렬 작업을 실행할 새 인스턴스 용량을 AWS Batch 생성합니다. `minvCpus` 또는 `maxvCpus` 값이 설정된 동일한 컴퓨팅 환경에서 단일 노드 및 다중 노드 병렬 작업이 모두 실행되는 경우 필요한 컴퓨팅 리소스를 사용할 수 없는 경우 새 작업을 실행하는 데 필요한 컴퓨팅 리소스를 생성하기 전에 현재 작업이 완료될 AWS Batch 때까지 기다립니다.

비 관리형 컴퓨팅 환경

비관리형 컴퓨팅 환경에서는 자체 컴퓨팅 리소스를 관리합니다.는 Amazon ECS와 Amazon EKS 모두에 대해 비관리형 컴퓨팅 환경을 AWS Batch 지원하므로 배치의 작업 예약 기능을 활용하면서 인프라를 제어할 수 있습니다.

> **Note**
>
> AWS Fargate 리소스는 비관리형 컴퓨팅 환경에서 지원되지 않습니다.

비관리형 Amazon ECS 컴퓨팅 환경

비관리형 Amazon ECS 컴퓨팅 환경의 경우 컴퓨팅 리소스에 사용하는 AMI가 Amazon ECS 컨테이너 인스턴스 AMI 사양을 충족하는지 확인해야 합니다. 자세한 내용은 컴퓨팅 리소스 AMI 사양 및 자습서: 컴퓨팅 리소스 AMI 생성 섹션을 참조하세요.

비관리형 컴퓨팅 환경을 생성한 후에는 DescribeComputeEnvironments API 작업을 사용하여 컴퓨팅 환경 세부 정보를 봅니다. 환경과 연결된 Amazon ECS 클러스터를 찾은 후 해당 Amazon ECS 클러스터에서 컨테이너 인스턴스를 수동으로 시작합니다.

다음 AWS CLI 명령은 Amazon ECS 클러스터 ARN도 제공합니다.

```
$ aws batch describe-compute-environments \
    --compute-environments unmanagedCE \
    --query "computeEnvironments[].ecsClusterArn"
```

자세한 내용을 알아보려면 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS 컨테이너 인스턴스 시작을 참조하세요. 사용자 컴퓨팅 리소스를 시작할 때 리소스가 다음 Amazon EC2 사용자 데이터에 등록해야 하는 Amazon ECS 클러스터 ARN을 지정합니다. *ecsClusterArn*을 이전 명령에서 받은 클러스터 ARN으로 바꿉니다.

```
#!/bin/bash
echo "ECS_CLUSTER=ecsClusterArn" >> /etc/ecs/ecs.config
```

비관리형 Amazon EKS 컴퓨팅 환경

비관리형 Amazon EKS 컴퓨팅 환경에서는가 작업 예약 및 배치를 AWS Batch 처리하는 동안 자체 Kubernetes 노드를 관리합니다. 이를 통해 보안, 규정 준수 또는 운영 요구 사항을 위해 Kubernetes 인프라를 직접 제어할 수 있습니다. Amazon EKS 노드를 프로비저닝하고 구성하는 것은 사용자의 책임이며,는 기존 Amazon EKS 클러스터와 AWS Batch 통합되어 작업을 예약하고 실행합니다.

자세한 내용은 자습서: Amazon EKS 리소스를 사용하여 비관리형 컴퓨팅 환경 생성을 참조하세요.

> **Amazon EKS Auto Mode 호환성**
>
> AWS Batch 는 현재 Amazon EKS Auto Mode 작업자 노드에서 작업을 실행하지 않습니다. 비 AWS Batch관리형 Amazon EKS 컴퓨팅 환경에는 지속적인 고객 레이블이 지정된 노드가 필요한 반면, Auto Mode는 보류 중인 포드 압력에 따라 Karpenter를 통해 노드를 동적으로 프로비저닝합니다.
>
> 비관리형 Amazon EKS 컴퓨팅 환경은 AWS Batch 컴퓨팅 환경이 Auto Mode에서 관리하지 않는 전용 노드 그룹을 가리키는 한 다른 워크로드에 대해 Auto Mode가 활성화된 Amazon EKS 클러스터와 공존할 수 있습니다. Auto Mode는 AWS Batch 노드 그룹을 방해하지 않고 워크로드가 아닌를AWS Batch 독립적으로 계속 관리합니다.

컴퓨팅 환경 생성

에서 작업을 실행하려면 먼저 컴퓨팅 환경을 생성 AWS Batch해야 합니다. 가 사양에 따라 환경 내에서 Amazon EC2 인스턴스 또는 AWS Fargate 리소스를 AWS Batch 관리하는 관리형 컴퓨팅 환경을 생성할 수 있습니다. 또는 환경 내에서 사용자가 Amazon EC2 인스턴스 구성을 처리하는 비 관리형 컴퓨팅 환경을 만들 수도 있습니다.

> **Important**
>
> Fargate 스팟 인스턴스는 다음 시나리오에서 지원되지 않습니다.
>
> - Windows containers on AWS Fargate
>
> 이러한 시나리오에서 Fargate Spot 컴퓨팅 환경만 사용하는 작업 대기열에 작업이 제출되는 경우 작업 대기열이 차단됩니다.

주제

- 자습서: Fargate 리소스를 사용하여 관리형 컴퓨팅 환경 생성
- 자습서: Amazon EC2 리소스를 사용하여 관리형 컴퓨팅 환경 생성
- 자습서: Amazon EC2 리소스를 사용하여 비관리형 컴퓨팅 환경 생성
- 자습서: Amazon EKS 리소스를 사용하여 관리형 컴퓨팅 환경 생성
- 자습서: Amazon EKS 리소스를 사용하여 비관리형 컴퓨팅 환경 생성
- 리소스: 컴퓨팅 환경 템플릿
- 인스턴스 유형 컴퓨팅 표

자습서: Fargate 리소스를 사용하여 관리형 컴퓨팅 환경 생성

AWS Fargate 리소스를 사용하여 관리형 컴퓨팅 환경을 생성하려면 다음 단계를 완료합니다.

1. https://console.aws.amazon.com/batch/ AWS Batch 콘솔을 엽니다.
2. 탐색 모음에서 사용할 AWS 리전 를 선택합니다.
3. 탐색 창에서 컴퓨팅 환경를 선택합니다.
4. 생성(Create)을 선택합니다.

5. 환경을 구성합니다.

> **Note**
>
> Windows containers on AWS Fargate 작업을 위한 컴퓨팅 환경은 하나 이상의 vCPU여야 합니다.

   a. 컴퓨팅 환경 구성에서 Fargate를 선택합니다.

   b. 이름에 해당 컴퓨팅 환경의 고유한 이름을 지정합니다. 이름은 최대 128자를 포함할 수 있습니다. 대문자 및 소문자, 숫자, 하이픈(-) 및 밑줄(_)을 포함할 수 있습니다.

   c. 서비스 역할에서 서비스가 사용자를 대신하여 필요한 AWS API 작업을 호출 AWS Batch 할 수 있도록 서비스 연결 역할을 선택합니다. 예를 들어, AWSServiceRoleForBatch을 선택합니다. 자세한 내용은 에 대한 서비스 연결 역할 사용 AWS Batch 단원을 참조하십시오.

   d. (선택 사항) 태그를 확장하세요. 태그를 추가하려면 태그 추가를 선택합니다. 그런 다음 키 이름과 선택 사항인 값을 입력합니다. 태그 추가를 선택합니다.

   e. 다음 페이지를 선택합니다.

6. 인스턴스 구성 섹션에서:

   a. (선택 사항) Fargate 스팟 용량 사용에서 Fargate 스팟을 켜세요. Fargate 스팟에 대한 자세한 내용은 Amazon EC2 스팟 및 Fargate_SPOT 사용을 참조하세요.

   b. 최대 vCPUs의 경우, 작업 대기열 수요와 상관없이 사용자 컴퓨팅 환경이 스케일 아웃 할 수 있는 최대 vCPU 수를 선택합니다.

   c. 다음 페이지를 선택합니다.

7. 네트워킹을 구성합니다.

> **Important**
>
> 컴퓨팅 리소스는 Amazon ECS 서비스 엔드포인트와 통신하기 위한 액세스 권한이 필요합니다. 이는 인터페이스 VPC 엔드포인트를 통하거나 퍼블릭 IP 주소가 있는 컴퓨팅 리소스를 통해 이루어질 수 있습니다.
>
> 인터페이스 VPC 엔드포인트에 대한 자세한 정보는 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS 인터페이스 VPC 엔드포인트(AWS PrivateLink)를 참조하세요.
>
> 인터페이스 VPC 엔드포인트가 구성되어 있지 않고 컴퓨팅 리소스에 퍼블릭 IP 주소가 없는 경우 NAT(Network Address Translation)를 사용하여 이 액세스 권한을 제공해야 합니

다. 자세한 내용은 Amazon VPC 사용 설명서의 NAT 게이트웨이 섹션을 참조하세요. 자세한 내용은 the section called “VPC 생성” 단원을 참조하십시오.

a. Virtual Private Cloud(VPC) ID에서 인스턴스를 시작할 VPC를 선택합니다.

b. 서브넷에서 사용할 서브넷을 선택합니다. 기본적으로 선택된 VPC의 모든 서브넷은 사용 가능합니다.

> **Note**
>
> AWS Batch Fargate의는 현재 로컬 영역을 지원하지 않습니다. 자세한 내용은 Amazon Elastic Container 서비스 개발자 안내서의 Amazon ECS clusters in Local Zones, Wavelength Zones, AWS Outposts을 참조하세요

c. Security groups(보안 그룹)에서 인스턴스에 연결할 보안 그룹을 선택합니다. 기본적으로 VPC의 기본 보안 그룹이 선택됩니다.

d. 다음 페이지를 선택합니다.

8. 검토에서 구성 과정을 검토하세요. 변경해야 하는 경우 편집을 선택합니다 작업을 마쳤으면 컴퓨팅 환경 생성을 선택합니다.

자습서: Amazon EC2 리소스를 사용하여 관리형 컴퓨팅 환경 생성

Amazon Elastic Compute Cloud(Amazon EC2) 리소스를 사용하여 관리형 컴퓨팅 환경을 생성하려면 다음 단계를 완료합니다.

1. https://console.aws.amazon.com/batch/ AWS Batch 콘솔을 엽니다.
2. 탐색 모음에서 사용할 AWS 리전 를 선택합니다.
3. 탐색 창에서 환경을 선택합니다.
4. 사용자 환경을 선택한 다음 컴퓨팅 환경을 선택합니다.
5. 환경을 구성합니다.

a. 컴퓨팅 환경에서 Amazon Elastic Compute Cloud(Amazon EC2)를 선택합니다.

b. 오케스트레이션 유형은 관리형을 선택합니다.

c. 이름에 해당 컴퓨팅 환경의 고유한 이름을 지정합니다. 이름은 최대 128자를 포함할 수 있습니다. 대문자 및 소문자, 숫자, 하이픈(-) 및 밑줄(_)을 포함할 수 있습니다.

d. 서비스 역할에서 서비스가 사용자를 대신하여 필요한 AWS API 작업을 AWS Batch 호출할 수 있도록 서비스 연결 역할을 선택합니다. 예를 들어, AWSServiceRoleForBatch을 선택합니다. 자세한 내용은 에 대한 서비스 연결 역할 사용 AWS Batch 단원을 참조하십시오.

e. Instance role(인스턴스 역할)에서 새 인스턴스 프로파일을 생성하도록 선택하거나, 필요한 IAM 권한이 연결되어 있는 기존 인스턴스 프로파일을 사용합니다. 이 인스턴스 프로파일을 사용하면 컴퓨팅 환경에 대해 생성된 Amazon ECS 컨테이너 인스턴스가 사용자를 대신하여 필요한 AWS API 작업을 호출할 수 있습니다. 자세한 내용은 Amazon ECS 인스턴스 역할 단원을 참조하십시오. 새 인스턴스 프로파일을 생성하도록 선택하면 필수 역할(`ecsInstanceRole`)이 생성됩니다.

f. (선택 사항) 태그를 확장하세요.

   i. (선택 사항) EC2 태그의 경우 태그 추가를 선택하여 컴퓨팅 환경에서 시작되는 리소스에 태그를 추가합니다. 그런 다음 키 이름과 선택 사항인 값을 입력합니다. 태그 추가를 선택합니다.

   ii. (선택 사항) 태그의 경우태그 추가를 선택합니다. 그런 다음 키 이름과 선택 사항인 값을 입력합니다. 태그 추가를 선택합니다.

   자세한 내용은 AWS Batch 리소스 태깅 단원을 참조하십시오.

g. 다음을 선택합니다.

6. 인스턴스 구성 섹션에서:

a. (선택 사항) 스팟 인스턴스 사용 활성화를 켭니다. 자세한 내용은 스팟 인스턴스를 참조하세요.

b. (스팟 전용) 온디맨드 가격 최대 %에서 인스턴스를 시작하기 전에 해당 인스턴스 유형에 대한 온디맨드 가격과 비교하여 스팟 인스턴스 가격에 대해 설정할 수 있는 최대 비율(%)을 선택합니다. 예를 들어 최고 가격이 20%인 경우 스팟 가격은 현행 EC2 인스턴스에 대한 온디맨드 가격 보다 20% 이상 낮아야 합니다. 항상 최저 (시장) 가격을 지불하고 최대 비율을 넘지 않도록 할 수 있습니다. 이 필드를 비워두면 기본값은 온디맨드 가격의 100%입니다.

c. (스팟 전용) 스팟 플릿 역할에서 스팟 컴퓨팅 환경에 적용할 기존 Amazon EC2 스팟 플릿 IAM 역할을 선택합니다. 기존 Amazon EC2 스팟 플릿 IAM 역할이 없으면 먼저 역할을 생성해야 합니다. 자세한 내용은 Amazon EC2 스팟 플릿 역할 단원을 참조하십시오.

> **Important**
> 스팟 인스턴스를 생성에 태그하려면 Amazon EC2 스팟 플릿 IAM 역할은 최신 AmazonEC2SpotFleetTaggingRole 관리형 정책을 사용해야 합니다.

AmazonEC2SpotFleetRole 관리형 정책에는 스팟 인스턴스에 태그를 지정에 권한을 요구하지 않습니다. 자세한 내용은 생성 시 태그가 지정되지 않은 스팟 인스턴스 및 the section called “리소스에 태그 지정” 섹션을 참조하세요.

d. 최소 vCPUs에서는 작업 대기열 수요와 상관없이 사용자 컴퓨팅 환경에서 유지할 수 있는 최소 vCPU 수를 선택합니다.

e. 바람직한 vCPU에서는 사용자 컴퓨팅 환경이 시작할 수 있는 vCPU 수를 선택합니다. 작업 대기열 수요가 증가하면 AWS Batch 는 컴퓨팅 환경에서 원하는 vCPU 수를 최대 vCPU 수까지 늘리고 EC2 인스턴스를 추가할 수 있습니다. 수요가 감소하면 AWS Batch 는 컴퓨팅 환경에서 원하는 vCPU 수를 최소 vCPU 수까지 줄이고 인스턴스를 제거할 수 있습니다.

f. 최대 vCPUs의 경우, 작업 대기열 수요와 상관없이 사용자 컴퓨팅 환경이 스케일 아웃 할 수 있는 최대 vCPU 수를 선택합니다.

g. (선택 사항) 축소 지연(분)에서 작업이 완료된 후 컴퓨팅 환경에서 인스턴스를 AWS Batch 계속 실행하는 최소 시간(분)을 선택합니다.

h. 허용된 인스턴스 유형에서 시작할 수 있는 Amazon EC2 인스턴스 유형을 선택합니다. 사용자는 특정 인스턴스 패밀리 (예: `c5`, `c5n`, 혹은 `p3`) 내에서 모든 인스턴스 유형을 시작하기 위해 인스턴스 패밀리를 지정할 수 있습니다. 또는 제품군 내에서 특정 크기(예: `c5.8xlarge`)를 지정할 수 있습니다. 메탈 인스턴스 유형은 인스턴스 패밀리에 없습니다. 예를 들어, `c5`에는 `c5.metal`가 포함되어 있지 않습니다.

   AWS Batch 다음 중 하나를 선택하면에서 인스턴스 유형을 선택할 수 있습니다.

   - `optimal`는 리전 가용성에 따라 최신 m, c 및 r 인스턴스 패밀리에서 인스턴스 유형을 선택합니다. AWS Batch 는 이러한 패밀리 내의 최신 세대로 풀을 정기적으로 업데이트합니다.
   - `default_x86_64` - 작업 대기열의 리소스 수요와 일치하는 x86 기반 인스턴스 유형(m6i, c6i, r6i 및 c7i 인스턴스 패밀리)을 선택합니다.
   - `default_arm64` 작업 대기열의 리소스 수요와 일치하는 Arm 기반 인스턴스 유형(m6g, r6g, c6g및 c7g 인스턴스 패밀리)을 선택합니다.

> **Note**
>
> - 인스턴스 패밀리 가용성은 AWS 리전에 따라 달라집니다. 예를 들어 일부 AWS 리전에는 4세대 인스턴스 패밀리가 없지만 5세대 및 6세대 인스턴스 패밀리가 있을 수 있습니다.

- `default_x86_64` 또는 `default_arm64` 인스턴스 번들을 사용할 때 AWS Batch는 비용 효율과 성능의 균형을 기반으로 인스턴스 패밀리를 선택합니다. 최신 세대 인스턴스는 더 나은 가격 대비 성능을 제공하는 경우가 많지만 워크로드에 최적의 가용성, 비용 및 성능 조합을 제공하는 경우 이전 세대 인스턴스 패밀리를 AWS Batch 선택할 수 있습니다. 예를 들어 c6i 인스턴스와 c7i 인스턴스를 모두 사용할 수 AWS 리전 있는 에서는 특정 작업 요구 사항에 더 나은 비용 효율성을 제공하는 경우 c6i 인스턴스를 AWS Batch 선택할 수 있습니다. AWS Batch 인스턴스 유형 및 AWS 리전 가용성에 대한 자세한 내용은 인스턴스 유형 컴퓨팅 표를 참조하세요.
- AWS Batch 는 기본 번들의 인스턴스를 보다 새롭고 비용 효율적인 옵션으로 주기적으로 업데이트합니다. 업데이트는 사용자의 조치 없이 자동으로 수행됩니다. 워크로드는 업데이트 동안 중단 없이 계속 실행됩니다.

> **Note**
>
> 컴퓨팅 환경을 생성할 때 컴퓨팅 환경에 대해 선택한 인스턴스 유형은 동일한 아키텍처를 공유해야 합니다. 예를 들어, 동일한 컴퓨팅 환경에서 x86 및 ARM 인스턴스를 함께 사용할 수 없습니다.

> **Note**
>
> AWS Batch 는 작업 대기열에 필요한 양에 따라 GPUs를 조정합니다. GPU 일정 설정을 사용하려면 컴퓨팅 환경이 p3, p4, p5, p6, g3, g3s, g4, g5 또는 g6 패밀리의 인스턴스 유형을 포함해야 합니다.

i. 할당 전략의 경우 허용되는 인스턴스 유형 목록에서 인스턴스 유형을 선택할 때 사용할 할당 전략을 선택합니다. EC2 온디맨드 컴퓨팅 환경에는 일반적으로 BEST_FIT_PROGRESSIVE, EC2 스팟 컴퓨팅 환경에는 SPOT_CAPACITY_OPTIMIZED와 SPOT_PRICE_CAPACITY_OPTIMIZED가 더 적합합니다. 자세한 내용은 the section called "인스턴스 유형 할당 전략" 단원을 참조하십시오.

j. 추가 구성을 확장합니다.

   i. (선택 사항) 배치 그룹에는 컴퓨팅 환경의 리소스를 그룹화할 배치 그룹 이름을 입력합니다.

ii. (선택 사항) EC2 키 페어에서는 인스턴스에 연결할 때 퍼블릭 및 프라이빗 키 페어를 보안 자격 증명으로 선택합니다. Amazon EC2 키 페어에 대한 자세한 내용은 Amazon EC2 키 페어 및 Linux 인스턴스를 참조하세요.

iii. (선택 사항) EC2 구성에서 이미지 유형 및 이미지 ID 재정의 값을 선택하여 AWS Batch에 대한 정보를 제공하여 컴퓨팅 환경의 인스턴스에 대해 Amazon Machine Image(AMIs)를 선택합니다. 각 이미지 유형에 대해 이미지 ID 재정의가 지정되지 않은 경우는 최근 Amazon ECS 최적화 AMI를 AWS Batch 선택합니다. 이미지 유형을 지정하지 않으면 기본값은 비 GPU, 비 AWS Graviton 인스턴스의 경우 Amazon Linux 2입니다.

> **Important**
>
> 사용자 지정 AMI를 사용하려면 이미지 유형을 선택한 다음 이미지 ID 재정의 상자에 사용자 지정 AMI ID를 입력합니다.

Amazon Linux 2

모든 AWS Graviton 기반 인스턴스 패밀리(예: , C6gM6gR6g, 및 T4g)의 기본값이며 GPU가 아닌 모든 인스턴스 유형에 사용할 수 있습니다.

Amazon Linux 2(GPU)

모든 GPU 인스턴스 패밀리(예: P4 및 G4)의 기본값이며 AWS Graviton 기반이 아닌 모든 인스턴스 유형에 사용할 수 있습니다.

Amazon Linux 2023

AWS Batch 는 Amazon Linux 2023을 지원합니다.

> **Note**
>
> Amazon Linux 2023은 A1 인스턴스를 지원하지 않습니다.

Amazon Linux 2023(GPU)

모든 GPU 인스턴스 패밀리(예: P4 및 G4)의 기본값이며 AWS Graviton 기반이 아닌 모든 인스턴스 유형에 사용할 수 있습니다.

> **Note**
>
> 컴퓨팅 환경에 대해 선택한 AMI는 해당 컴퓨팅 환경에 사용할 인스턴스 유형의 아키텍처와 일치해야 합니다. 예를 들어, 컴퓨팅 환경에서 A1 인스턴스 유형을 사용하는 경우 선택한 컴퓨팅 리소스 AMI는 ARM 인스턴스를 지원해야 합니다. Amazon ECS는 Amazon ECS에 최적화된 Amazon Linux 2 AMI의 x86 버전과 ARM 버전을 모두 제공합니다. 자세한 내용을 알아보려면 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS 최적화 Amazon Linux 2 AMI를 참조하세요.

k. (선택 사항) 시작 템플릿 확장

i. 기본 시작 템플릿에서 기존 Amazon EC2 시작 템플릿을 선택하여 컴퓨팅 리소스를 구성합니다. 템플릿의 기본 버전이 자동으로 생성됩니다. 자세한 내용은 에서 Amazon EC2 시작 템플릿 사용 AWS Batch 단원을 참조하십시오.

> **Note**
>
> 시작 템플릿에서 생성한 사용자 지정 AMI를 지정할 수 있습니다.

ii. (선택 사항) 기본 버전에 `$Default`, `$Latest` 또는 사용할 특정 버전 번호를 입력합니다.

> **Note**
>
> 참고: 대체 변수(($Default 또는 $Latest))를 사용하는 경우 이 구성이 저장될 때 현재 기본 또는 최신 버전 번호가 적용됩니다. 나중에 기본 버전이나 최신 버전이 변경될 경우 정보를 업데이트해야 합니다. 정보는 자동으로 업데이트되지 않습니다.

> **Important**
>
> 시작 템플릿의 버전 파라미터가 `$Default` 혹은 `$Latest`인 경우 인프라 업데이트 중에 지정된 시작 템플릿의 기본 또는 최신 버전이 평가됩니다. 기본적으로 다른 AMI ID를 선택하거나 시작 템플릿의 최신 버전을 선택한 경우 해당 AMI ID가

업데이트에 사용됩니다. 자세한 내용은 the section called “인프라 업데이트 중 AMI 선택” 단원을 참조하십시오.

iii. (선택 사항) 재정의 시작 템플릿에서 재정의 시작 템플릿을 선택합니다.

A. (선택 사항) 시작 템플릿에서 특정 인스턴스 유형 및 패밀리에 사용할 기존 Amazon EC2 시작 템플릿을 선택합니다.

B. (선택 사항) 기본 버전에서 사용할 특정 버전 번호, `$Default` 또는 `$Latest`를 입력합니다.

> **Note**
>
> `$Default` 또는 `$Latest` 변수를 사용하는 경우 AWS Batch 는 컴퓨팅 환경이 생성된 시점의 최신 정보를 적용합니다. 향후 기본 또는 최신 버전이 변경되는 경우 UpdateComputeEnvironment 또는 AWS Management Console -를 통해 정보를 업데이트해야 합니다 AWS Batch.

C. (선택 사항) 대상 인스턴스 유형에서 재정의 시작 템플릿을 적용할 인스턴스 유형 또는 패밀리를 선택합니다.

> **Note**
>
> 재정의 시작 템플릿을 지정하는 경우 대상 인스턴스 유형이 필요합니다. 자세한 내용은 LaunchTemplateSpecificationOverride.targetInstanceTypes를 참조하세요.

> **Note**
>
> 선택하려는 인스턴스 유형 또는 패밀리가 이 목록에 표시되지 않는 경우, `Allowed instance types`에서 선택한 항목을 검토합니다.

I. 다음을 선택합니다.

7. 네트워크 구성 섹션에서:

> **Important**
>
> 컴퓨팅 리소스는 Amazon ECS 서비스 엔드포인트와 통신하기 위한 액세스 권한이 필요합니다. 이는 인터페이스 VPC 엔드포인트를 통하거나 퍼블릭 IP 주소가 있는 컴퓨팅 리소스를 통해 이루어질 수 있습니다.
> 인터페이스 VPC 엔드포인트에 대한 자세한 정보는 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS 인터페이스 VPC 엔드포인트(AWS PrivateLink)를 참조하세요.
> 인터페이스 VPC 엔드포인트가 구성되어 있지 않고 컴퓨팅 리소스에 퍼블릭 IP 주소가 없는 경우 NAT(Network Address Translation)를 사용하여 이 액세스 권한을 제공해야 합니다. 자세한 내용은 Amazon VPC 사용 설명서의 NAT 게이트웨이 섹션을 참조하세요. 자세한 내용은 the section called “VPC 생성” 단원을 참조하십시오.

a. Virtual Private Cloud(VPC) ID에서 인스턴스를 시작할 VPC를 선택합니다.

b. 서브넷에서 사용할 서브넷을 선택합니다. 기본적으로 선택된 VPC의 모든 서브넷은 사용 가능합니다.

> **Note**
>
> AWS Batch Amazon EC2의는 로컬 영역을 지원합니다. 자세한 내용은 Amazon EC2 사용 설명서의 로컬 영역 및 Amazon Elastic Container Service 개발자 안내서의 로컬 영역, Wavelength 영역 및 AWS Outposts의 Amazon ECS 클러스터를 참조하세요.

c. (선택 사항) 보안 그룹에서 인스턴스에 연결할 보안 그룹을 선택합니다. 기본적으로 VPC의 기본 보안 그룹이 선택됩니다.

> **Note**
>
> 참고: 대체 변수(($Default 또는 $Latest))를 사용하는 경우 이 구성이 저장될 때 현재 기본 또는 최신 버전 번호가 적용됩니다. 나중에 기본 버전이나 최신 버전이 변경될 경우 정보를 업데이트해야 합니다. 정보는 자동으로 업데이트되지 않습니다.

8. 다음 페이지를 선택합니다.

9. 검토에서 구성 과정을 검토하세요. 변경해야 하는 경우 편집을 선택합니다 작업을 마쳤으면 컴퓨팅 환경 생성을 선택합니다.

자습서: Amazon EC2 리소스를 사용하여 비관리형 컴퓨팅 환경 생성

Amazon Elastic Compute Cloud(Amazon EC2) 리소스를 사용하여 비관리형 컴퓨팅 환경을 생성하려면 다음 단계를 완료합니다.

1. https://console.aws.amazon.com/batch/ AWS Batch 콘솔을 엽니다.
2. 탐색 모음에서 사용할 AWS 리전 를 선택합니다.
3. 컴퓨팅 환경 페이지에서 생성을 선택합니다.
4. 환경을 구성합니다.
   a. 컴퓨팅 환경에서 Amazon Elastic Compute Cloud(Amazon EC2)를 선택합니다.
   b. 오케스트레이션 유형에서 비 관리형을 선택합니다
5. 이름에 해당 컴퓨팅 환경의 고유한 이름을 지정합니다. 각 이름의 최대 길이는 128자입니다. 대문자 및 소문자, 숫자, 하이픈(-) 및 밑줄(_)을 포함할 수 있습니다.
6. 서비스 역할에서 AWS Batch 서비스가 사용자를 대신하여 필요한 AWS API 작업을 호출할 수 있는 역할을 선택합니다.

   > **Note**
   > 관리되지 않는 컴퓨팅 환경에서는 `AWSServiceRoleForBatch`를 사용할 수 없습니다.

7. 최대 vCPUs의 경우, 작업 대기열 수요와 상관없이 사용자 컴퓨팅 환경이 스케일 아웃 할 수 있는 최대 vCPU 수를 선택합니다.
8. (선택 사항) 태그를 확장하세요. 태그를 추가하려면 태그 추가를 선택합니다. 그런 다음 키 이름과 선택 사항인 값을 입력합니다. 태그 추가를 선택합니다. 자세한 내용은 AWS Batch 리소스 태깅 단원을 참조하십시오.
9. 다음 페이지를 선택합니다.
10. 검토에서 구성 과정을 검토하세요. 변경해야 하는 경우 편집을 선택합니다 작업을 마쳤으면 컴퓨팅 환경 생성을 선택합니다.

자습서: Amazon EKS 리소스를 사용하여 관리형 컴퓨팅 환경 생성

Amazon Elastic Kubernetes Service(Amazon EKS) 리소스를 사용하여 관리형 컴퓨팅 환경을 생성하려면 다음 단계를 완료합니다.

1. https://console.aws.amazon.com/batch/ AWS Batch 콘솔을 엽니다.
2. 탐색 모음에서 사용할 AWS 리전 를 선택합니다.
3. 탐색 창에서 컴퓨팅 환경를 선택합니다.
4. 생성(Create)을 선택합니다.
5. 컴퓨팅 환경 구성에서 Amazon Elastic Kubernetes Service(Amazon EKS)를 선택합니다.
6. 이름에 해당 컴퓨팅 환경의 고유한 이름을 지정합니다. 각 이름의 최대 길이는 128자입니다. 대문자 및 소문자, 숫자, 하이픈(-) 및 밑줄(_)을 포함할 수 있습니다.
7. 인스턴스 역할에서 필요한 IAM 권한이 연결되어 있는 기존 인스턴스 프로파일을 선택합니다.

   > **Note**
   >
   > AWS Batch 콘솔에서 컴퓨팅 환경을 생성하려면 `eks:ListClusters` 및 `eks:DescribeCluster` 권한이 있는 인스턴스 프로파일을 선택합니다.

8. EKS 클러스터에서 기존 Amazon EKS 클러스터를 선택합니다.
9. 네임스페이스에서 클러스터에 AWS Batch 프로세스를 그룹화할 Kubernetes 네임스페이스를 입력합니다.
10. (선택 사항) 태그를 확장하세요. 태그 추가를 선택한 다음 키-값 페어를 입력합니다.
11. 다음 페이지를 선택합니다.
12. (선택 사항) EC2 스팟 인스턴스 사용에서 Amazon EC2 스팟 인스턴스를 사용하려면 스팟 인스턴스 사용 활성화를 켜세요.
13. (스팟 전용) 온디맨드 가격 최대 %에서 인스턴스를 시작하기 전에 해당 인스턴스 유형에 대한 온디맨드 가격과 비교하여 스팟 인스턴스 가격에 대해 설정할 수 있는 최대 비율(%)을 선택합니다. 예를 들어 최고 가격이 20%인 경우 스팟 가격은 현행 EC2 인스턴스에 대한 온디맨드 가격 보다 20% 이상 낮아야 합니다. 항상 최저 (시장) 가격을 지불하고 최대 비율을 넘지 않도록 할 수 있습니다. 이 필드를 비워두면 기본값은 온디맨드 가격의 100%입니다.
14. (스팟 전용) 스팟 플릿 역할에서 SPOT 컴퓨팅 환경에 적용할 Amazon EC2 스팟 플릿 IAM 역할을 선택합니다.

    > **Important**
    >
    > 할당 전략이 `BEST_FIT`로 설정되어 있거나 설정되지 않은 경우 이 역할이 필요합니다.

15. (선택 사항) 최소 vCPU에서 작업 대기열 수요와 상관없이 사용자 컴퓨팅 환경이 유지할 최소 vCPU 수를 선택합니다.

16. (선택 사항) 최대 vCPU에서 작업 대기열 수요와 상관없이 사용자 컴퓨팅 환경이 확장할 수 있는 최대 vCPU 수를 선택합니다.
17. (선택 사항) 축소 지연(분)에서 작업이 완료된 후 컴퓨팅 환경에서 인스턴스를 AWS Batch 계속 실행하는 최소 시간(분)을 선택합니다.
18. 허용된 인스턴스 유형에서 시작할 수 있는 Amazon EC2 인스턴스 유형을 선택합니다. 사용자는 특정 인스턴스 패밀리 (예: `c5`, `c5n`, 혹은 `p3`) 내에서 모든 인스턴스 유형을 시작하기 위해 인스턴스 패밀리를 지정할 수 있습니다. 또는 제품군 내에서 특정 크기(예: `c5.8xlarge`)를 지정할 수 있습니다. 메탈 인스턴스 유형은 인스턴스 패밀리에 없습니다. 예를 들어, `c5`에는 `c5.metal`가 포함되어 있지 않습니다.

    AWS Batch 다음 중 하나를 선택하면에서 인스턴스 유형을 선택할 수 있습니다.

    - `optimal`는 리전 가용성에 따라 최신 m, c 및 r 인스턴스 패밀리에서 인스턴스 유형을 선택합니다. AWS Batch 는 이러한 패밀리 내의 최신 세대로 풀을 정기적으로 업데이트합니다.
    - `default_x86_64` - 작업 대기열의 리소스 수요와 일치하는 x86 기반 인스턴스 유형(m6i, c6i, r6i 및 c7i 인스턴스 패밀리)을 선택합니다.
    - `default_arm64` 작업 대기열의 리소스 수요와 일치하는 Arm 기반 인스턴스 유형(m6g, r6g, c6g및 c7g 인스턴스 패밀리)을 선택합니다.

    > **Note**
    >
    > - 인스턴스 패밀리 가용성은 AWS 리전에 따라 달라집니다. 예를 들어 일부 AWS 리전에는 4세대 인스턴스 패밀리가 없지만 5세대 및 6세대 인스턴스 패밀리가 있을 수 있습니다.
    > - `default_x86_64` 또는 `default_arm64` 인스턴스 번들을 사용하는 경우는 비용 효율성과 성능의 균형을 기반으로 인스턴스 패밀리를 AWS Batch 선택합니다. 최신 세대 인스턴스는 더 나은 가격 대비 성능을 제공하는 경우가 많지만 워크로드에 최적의 가용성, 비용 및 성능 조합을 제공하는 경우 이전 세대 인스턴스 패밀리를 AWS Batch 선택할 수 있습니다. 예를 들어 c6i 인스턴스와 c7i 인스턴스를 모두 사용할 수 AWS 리전 있는 에서는 특정 작업 요구 사항에 더 나은 비용 효율성을 제공하는 경우 c6i 인스턴스를 AWS Batch 선택할 수 있습니다. AWS Batch 인스턴스 유형 및 AWS 리전 가용성에 대한 자세한 내용은 인스턴스 유형 컴퓨팅 테이블을 참조하세요.

- AWS Batch 는 기본 번들의 인스턴스를 보다 새롭고 비용 효율적인 옵션으로 주기적으로 업데이트합니다. 업데이트는 사용자의 조치 없이 자동으로 수행됩니다. 워크로드는 업데이트 동안 중단 없이 계속 실행됩니다.

> **Note**
>
> 컴퓨팅 환경을 생성할 때 컴퓨팅 환경에 대해 선택한 인스턴스 유형은 동일한 아키텍처를 공유해야 합니다. 예를 들어, 동일한 컴퓨팅 환경에서 x86 및 ARM 인스턴스를 함께 사용할 수 없습니다.

> **Note**
>
> AWS Batch 는 작업 대기열에 필요한 양에 따라 GPUs를 조정합니다. GPU 일정 설정을 사용하려면 컴퓨팅 환경이 p3, p4, p5, p6, g3, g3s, g4, g5 또는 g6 패밀리의 인스턴스 유형을 포함해야 합니다.

19. (선택 사항) 추가 구성을 확장합니다.

    a. (선택 사항) 배치 그룹에는 컴퓨팅 환경의 리소스를 그룹화할 배치 그룹 이름을 입력합니다.

    b. 할당 전략의 경우 BEST_FIT_PROGRESSIVE를 선택하세요.

    c. (선택 사항) Amazon Machine Image(AMI) 구성의 경우 Amazon Machine Image(AMI) 구성 추가를 선택합니다.

       Amazon EKS 최적화 Amazon Linux AMI 또는 사용자 지정 AMI를 사용할 수 있습니다.

       i. Amazon EKS 최적화 Amazon Linux AMI를 사용하려면

          A. 이미지 유형에서 다음 중 하나를 선택합니다.

             - Amazon Linux 2: 모든 AWS Graviton 기반 인스턴스 패밀리(예: , C6gM6gR6g, 및 T4g)의 기본값이며 GPU가 아닌 모든 인스턴스 유형에 사용할 수 있습니다.
             - Amazon Linux 2(가속): 모든 GPU 인스턴스 패밀리(예: P4 및 G4)의 기본값이며 AWS Graviton 기반이 아닌 모든 인스턴스 유형에 사용할 수 있습니다.
             - Amazon Linux 2023:는 Amazon Linux 2023(AL2023)을 AWS Batch 지원합니다.

- Amazon Linux 2023(가속): GPU 인스턴스 패밀리이며 모든 비 AWS Graviton 기반 인스턴스 유형에 사용할 수 있습니다.

B. Kubernetes 버전에 Kubernetes 버전 번호를 입력합니다.

ii. 사용자 지정 AMI를 사용하려면:

A. 이미지 유형에서 사용자 지정 AMI의 기반이 되는 AMI 유형을 선택합니다.

- Amazon Linux 2: 모든 AWS Graviton 기반 인스턴스 패밀리(예: , C6gM6gR6g, 및 T4g)의 기본값이며 GPU가 아닌 모든 인스턴스 유형에 사용할 수 있습니다.
- Amazon Linux 2(가속): 모든 GPU 인스턴스 패밀리(예: P4 및 G4)의 기본값이며 AWS Graviton 기반이 아닌 모든 인스턴스 유형에 사용할 수 있습니다.
- Amazon Linux 2023: AL2023을 AWS Batch 지원합니다.
- Amazon Linux 2023(가속): GPU 인스턴스 패밀리이며 모든 비 AWS Graviton 기반 인스턴스 유형에 사용할 수 있습니다.

B. 이미지 ID 재정의에 사용자 지정 AMI ID를 입력합니다.

C. Kubernetes 버전에 Kubernetes 버전 번호를 입력합니다.

d. (선택 사항) 시작 템플릿에서 기존 시작 템플릿을 선택합니다.

e. (선택 사항) 템플릿 버전 시작에 **$Default**, **$Latest** 또는 버전 번호를 입력합니다.

f. (선택 사항) 재정의를 추가하려면 재정의 시작 템플릿에서 재정의 시작 템플릿 추가를 선택합니다.

i. (선택 사항) 시작 템플릿에서 재정의를 추가할 시작 템플릿을 선택합니다.

ii. (선택 사항) 시작 템플릿 버전에서 시작 템플릿의 버전 번호, $Default 또는 $Latest를 선택합니다.

iii. (선택 사항) 대상 인스턴스 유형에서 이 재정의가 적용되어야 하는 인스턴스 유형 또는 패밀리를 선택합니다. 이는 허용되는 인스턴스 유형에 포함된 인스턴스 유형 및 패밀리만 대상으로 할 수 있습니다.

iv. (선택 사항) userdataType에서 EKS 노드 초기화를 선택합니다. 시작 템플릿 또는 시작 템플릿 재정의로 지정된 AMI가 있는 경우에만 이 필드를 사용합니다. EKS_AL2023 또는 EKS_AL2023_NVIDIA 기반 사용자 지정 AMI의 경우 EKS_NODEADM을 선택하고 EKS_AL2 및 EKS_AL_NVIDIA의 경우 EKS_BOOSTRAP_SH를 선택합니다. 기본값은 EKS_BOOSTRAP_SH입니다.

동일한 컴퓨팅 환경에서 AL2와 AL2023 기반 사용자 지정 AMI를 모두 사용하는 혼합 환경의 경우 userdataType을 사용합니다.

20. 다음 페이지를 선택합니다.
21. Virtual Private Cloud(VPC) ID에서 인스턴스를 시작할 VPC를 선택합니다.
22. 서브넷에서 사용할 서브넷을 선택합니다. 기본적으로 선택된 VPC의 모든 서브넷은 사용 가능합니다.

> **Note**
>
> AWS Batch Amazon EKS의는 로컬 영역을 지원합니다. 자세한 내용은 Amazon EKS 사용 설명서의 Amazon EKS 및 AWS 로컬 영역을 참조하세요.

23. (선택 사항) 보안 그룹에서 인스턴스에 연결할 보안 그룹을 선택합니다. 사용자 VPC의 기본 보안 그룹이 기본 선택됩니다.
24. 다음 페이지를 선택합니다.
25. 검토에서 구성 과정을 검토하세요. 변경해야 하는 경우 편집을 선택합니다 작업을 마쳤으면 컴퓨팅 환경 생성을 선택합니다.

자습서: Amazon EKS 리소스를 사용하여 비관리형 컴퓨팅 환경 생성

Amazon Elastic Kubernetes Service(Amazon EKS) 리소스를 사용하여 비관리형 컴퓨팅 환경을 생성하려면 다음 단계를 완료하세요.

1. https://console.aws.amazon.com/batch/ AWS Batch 콘솔을 엽니다.
2. 페이지 상단의 탐색 모음에서 사용할 AWS 리전 를 선택합니다.
3. 탐색 창에서 컴퓨팅 환경를 선택합니다.
4. 생성(Create)을 선택합니다.
5. 환경을 구성합니다.
   a. 컴퓨팅 환경 구성에서 Amazon Elastic Kubernetes Service(Amazon EKS)를 선택합니다.
   b. 오케스트레이션 유형에서 비 관리형을 선택합니다
6. 이름에 해당 컴퓨팅 환경의 고유한 이름을 지정합니다. 각 이름의 최대 길이는 128자입니다. 대문자 및 소문자, 숫자, 하이픈(-) 및 밑줄(_)을 포함할 수 있습니다.
7. EKS 클러스터에서 기존 Amazon EKS 클러스터를 선택합니다. 새 EKS 클러스터를 생성하려면 Amazon EKS 클러스터 생성 페이지의 단계를 따릅니다.

> **Note**
>
> AWS Batch 는 현재 Amazon EKS Auto Mode 작업자 노드에서 작업을 실행하지 않습니다. AWS Batch의 비관리형 Amazon EKS 컴퓨팅 환경에는 지속적인 고객 레이블이 지정된 노드가 필요한 반면, Auto Mode는 보류 중인 포드 압력에 따라 Karpenter를 통해 노드를 동적으로 프로비저닝합니다.
> 비관리형 Amazon EKS 컴퓨팅 환경은 AWS Batch 컴퓨팅 환경이 Auto Mode에서 관리하지 않는 전용 노드 그룹을 가리키는 한 다른 워크로드에 대해 Auto Mode가 활성화된 Amazon EKS 클러스터와 공존할 수 있습니다. Auto Mode는 AWS Batch 노드 그룹을 방해하지 않고 워크로드가 아닌를AWS Batch 독립적으로 계속 관리합니다.

8. 네임스페이스에서 클러스터에 AWS Batch 프로세스를 그룹화할 Kubernetes 네임스페이스를 입력합니다.
9. (선택 사항) 최대 vCPUs에서 프로비저닝된 용량에서 작업 예약에 사용할 수 있는 최대 vCPUs 수를 지정합니다.
10. (선택 사항) 태그를 확장하세요. 태그 추가를 선택한 다음 키-값 페어를 입력합니다.
11. 다음 페이지를 선택합니다.
12. 검토에서 구성 과정을 검토하세요. 변경해야 하는 경우 편집을 선택합니다 작업을 마쳤으면 컴퓨팅 환경 생성을 선택합니다.

> **비관리형 컴퓨팅 환경에 Amazon EKS 클러스터 노드 할당**
>
> 비관리형 컴퓨팅 환경을 생성한 후에는 Amazon EKS 노드에 컴퓨팅 환경 UUID로 레이블을 지정해야 합니다.
> 먼저 `DescribeComputeEnvironments` API 결과에서 컴퓨팅 환경 UUID를 가져옵니다.
>
> ```
> $ aws batch describe-compute-environments \
>     --compute-environments unmanagedEksCE \
>     --query "computeEnvironments[].{name: computeEnvironmentName, uuid: uuid}"
> ```
>
> 노드 정보를 가져옵니다.
>
> ```
> kubectl get nodes -o name
> ```
>
> AWS Batch 컴퓨팅 환경 UUID로 노드에 레이블을 지정합니다.

```
kubectl label <node-name> batch.amazonaws.com/compute-environment-uuid=uuid
```

리소스: 컴퓨팅 환경 템플릿

다음 예제는 빈 컴퓨팅 환경 템플릿입니다. 이 템플릿을 사용하여 파일에 저장하고 옵션과 함께 사용할 수 있는 컴퓨팅 환경을 생성할 수 있습니다 AWS CLI `--cli-input-json`. 이러한 파라미터에 대한 자세한 내용은 AWS Batch API 참조에서 CreateComputeEnvironment를 참조하세요.

관리형 Amazon EC2 컴퓨팅 환경을 생성하기 전에 다음 사전 요구 사항이 있는지 확인합니다. 이러한 사전 조건은 `type` 필드가 로 설정된 경우에 적용됩니다`MANAGED`.

- 보안 그룹 - 컴퓨팅 리소스에는 인스턴스가 Amazon ECS 서비스 엔드포인트와 통신하고 컨테이너 이미지를 가져올 수 있도록 아웃바운드 트래픽을 허용하는 보안 그룹이 필요합니다. 자세한 내용은 보안 그룹 생성 단원을 참조하십시오.
- IAM 역할 - 컨테이너 인스턴스가 사용자를 대신하여 AWS API를 호출할 수 있도록 허용하는 Amazon ECS 인스턴스 역할이 AWS Batch 필요합니다. 자세한 내용은 Amazon ECS 인스턴스 역할 및 에 대한 서비스 연결 역할 사용 AWS Batch 섹션을 참조하세요.

  > **Note**
  >
  > `instanceRole` 필드는 역할 ARN이 아닌 인스턴스 프로파일 ARN을 허용합니다. 형식은 `arn:aws:iam::account_id:instance-profile/ecsInstanceRole`입니다.

- 네트워크 액세스 - 컴퓨팅 리소스가 Amazon ECS 서비스 엔드포인트에 도달할 수 있어야 합니다. 인스턴스가 퍼블릭 IP 주소가 없는 프라이빗 서브넷에 있는 경우 NAT 게이트웨이 또는 Amazon VPC 인터페이스 엔드포인트를 사용할 수 있습니다. 자세한 내용은 인터페이스 엔드포인트를 사용하여 액세스 AWS Batch 단원을 참조하십시오.

> **Note**
>
> 다음 AWS CLI 명령을 사용하여 컴퓨팅 환경 템플릿을 생성할 수 있습니다.
>
> ```
> $ aws batch create-compute-environment --generate-cli-skeleton
> ```

> **Important**
>
> 컴퓨팅 환경은 ENABLED 상태에서 생성해야 합니다.

다음 예제에서는 관리형 Amazon EC2 컴퓨팅 환경에 대한 스켈레톤 템플릿을 보여줍니다. 가 type인 경우 computeResources 블록이 필요합니다MANAGED.

```
{
    "computeEnvironmentName": "",
    "type": "MANAGED",
    "state": "ENABLED",
    "computeResources": {
        "type": "EC2",
        "allocationStrategy": "BEST_FIT_PROGRESSIVE",
        "minvCpus": 0,
        "maxvCpus": 16,
        "desiredvCpus": 0,
        "instanceTypes": [
            "default_arm64"
        ],
        "subnets": [
            "subnet-a1b2c3d4"
        ],
        "securityGroupIds": [
            "sg-a1b2c3d4"
        ],
        "instanceRole": "arn:aws:iam::123456789012:instance-profile/ecsInstanceRole",
        "tags": {
            "KeyName": ""
        },
        "launchTemplate": {
            "launchTemplateId": "",
            "version": "$Default"
        },
        "ec2Configuration": [
            {
                "imageType": "ECS_AL2023"
            }
        ]
    },
    "serviceRole": "",
    "tags": {
```

```
        "KeyName": ""
    }
}
```

다음 예제에서는 비관리형 Amazon EC2 컴퓨팅 환경에 대한 스켈레톤 템플릿을 보여줍니다. `computeResources` 블록은 `UNMANAGED` 컴퓨팅 환경에 사용되지 않으므로 생략해야 합니다.

```
{
    "computeEnvironmentName": "",
    "type": "UNMANAGED",
    "state": "ENABLED",
    "unmanagedvCpus": 0,
    "serviceRole": "",
    "tags": {
        "KeyName": ""
    }
}
```

인스턴스 유형 컴퓨팅 표

다음 표에는 AWS 리전인스턴스 패밀리 키워드 및 사용 가능한 인스턴스 패밀리가 나열되어 있습니다. AWS Batch 는 최신 패밀리에서 인스턴스를 할당하려고 시도하지만 인스턴스 패밀리 가용성은 이전 인스턴스 패밀리 생성에 따라 다를 수 AWS 리전 있습니다.

default_x86_64

리전	인스턴스 패밀리
AWS 리전가 지원하는 모든 AWS Batch	m6i, c6i, r6i c7i

default_arm64

리전	인스턴스 패밀리
AWS 리전가 지원하는 모든 AWS Batch	m6g, c6g, r6g c7g

최적

리전	인스턴스 패밀리
AWS 리전가 지원하는 모든 AWS Batch	리전 가용성에 따라 최신 m, c 및 r 인스턴스 패밀리입니다. AWS Batch 는 이러한 패밀리 내의 최신 세대로 풀을 정기적으로 업데이트합니다.

에서 컴퓨팅 환경 업데이트 AWS Batch

AWS Batch 는 컴퓨팅 환경을 업데이트하기 위한 여러 전략을 제공하며, 각 전략은 특정 업데이트 시나리오 및 요구 사항에 맞게 설계되었습니다. 이러한 접근 방식은 동일한 기본 업데이트 API를 사용하지만 업데이트를 효과적으로 관리하기 위한 다양한 규범적 방법을 사용합니다. AWS Batch 콘솔 또는를 사용하여 이러한 업데이트를 관리할 수 있습니다 AWS CLI. 이러한 전략을 이해하면 워크로드 중단을 최소화하면서 필요에 가장 적합한 방법을 선택하는 데 도움이 됩니다.

이 주제에서는 사용 가능한 업데이트 전략의 개요와 언제 각 접근 방식을 사용해야 하는지에 대한 지침을 제공합니다. 자세한 절차는 각 업데이트 전략의 개별 섹션을 참조하세요.

Important

AWS Batch 는 Amazon EC2 시작 템플릿, Amazon EC2 Auto Scaling 그룹, Amazon EC2 스팟 플릿 및 Amazon Amazon EC2 클러스터를 포함하여 사용자를 대신하여 계정 내에서 여러 AWS 리소스를 생성하고 관리합니다. 이러한 관리형 리소스는 최적의 AWS Batch 작동을 보장하도록 특별히 구성됩니다. AWS Batch 설명서에 명시적으로 명시되지 않는 한 이러한 AWS Batch관리형 리소스를 수동으로 수정하면 `INVALID` 컴퓨팅 환경, 최적화되지 않은 인스턴스 조정 동작, 워크로드 처리 지연 또는 예상치 못한 비용 등 예상치 못한 동작이 발생할 수 있습니다. 이러한 수동 수정은 AWS Batch 서비스에서 결정론적으로 지원할 수 없습니다. 항상 지원되는 AWS Batch APIs 또는 AWS Batch 콘솔을 사용하여 컴퓨팅 환경을 관리합니다.
지원되지 않는 수동 수정에는 자체 Amazon ECS 태스크 또는 서비스 AWS Batch관리형 Amazon ECS 클러스터 실행 또는 추가 프로세스, 데몬 또는 서비스 직접 AWS Batch관리형 인스턴스 시작이 포함됩니다.는 관리형 컴퓨팅 환경에서 컴퓨팅 리소스를 완전히 제어할 AWS Batch 수 있으며 언제든지 인스턴스를 종료하거나 태스크를 중지하거나 클러스터를 확장할 수 있습니다. 이러한 관리형 리소스에서 AWS Batch 작업 제출 외부에서 실행하는 모든 워크로드는 경고 없이 중단될 수 있습니다. 관리 AWS Batch형 클러스터 및 인스턴스에서 워크로드가 AWS Batch 아닌 클러스터 및 인스턴스를 실행하면 AWS Batch 작업 예약 및 인스턴스 크기 조정에도 방해가 될 수 있습니다.

주제

컴퓨팅 환경 업데이트 전략

규모 조정 또는 인프라 업데이트를 사용하면 컴퓨팅 환경이 그대로 업데이트됩니다. 블루/그린 업데이트 전략에서는 새 컴퓨팅 환경(그린)을 생성한 다음 이전 컴퓨팅 환경(블루)에서 새 컴퓨팅 환경(그린)으로 워크로드를 마이그레이션합니다.

AWS Batch 는 컴퓨팅 환경 업데이트를 위한 세 가지 전략을 제공합니다.

규모 조정 업데이트

규모 조정 업데이트는 기존 인스턴스를 교체하지 않고 인스턴스를 추가하거나 제거하여 컴퓨팅 환경의 용량을 조정합니다. 이는 가장 빠른 업데이트 시나리오이며 가동 중지 시간이 필요하지 않습니다. 용량 설정(vCPU)을 변경해야 하는 경우 규모 조정 업데이트를 사용합니다. 이러한 업데이트는 일반적으로 몇 분 내에 완료됩니다.

Fargate 업데이트는 규모 조정 업데이트와 동일한 절차를 사용하여 수행됩니다. 자세한 내용은 규모 조정 업데이트 수행 단원을 참조하십시오.

인프라 업데이트

인프라 업데이트는 컴퓨팅 환경의 인스턴스를 설정이 업데이트된 새 인스턴스로 대체합니다. 이러한 업데이트에는 특정 서비스 역할 및 할당 전략 구성이 필요하지만 가동 중지 시간을 최소화할 수 있으며 실행 중인 작업이 중단될 가능성이 있습니다. 인스턴스 유형, AMI 구성, 네트워킹 설정, 서비스 역할, 환경 상태 또는 기타 인프라 구성 요소를 수정해야 하는 경우 인프라 업데이트를 사용합니다. 이러한 업데이트는 일반적으로 작업 완료에 따라 10~30분 이내에 완료됩니다.

자세한 내용은 인프라 업데이트 수행 단원을 참조하십시오.

블루/그린 업데이트

블루/그린 업데이트는 기존 환경 옆에 새로운 컴퓨팅 환경을 생성하여 가동 중지 없이 점진적인 워크로드 전환을 가능하게 합니다. 이 접근 방식은 가장 안전한 업데이트 경로를 제공하지만 일시적

으로 두 개의 환경을 실행해야 합니다. 가동 중지 시간이 없어야 하거나, 전체 배포 전에 변경 사항을 테스트하고자 하거나, 빠른 롤백 기능이 필요하거나, 인프라 업데이트에 대해 지원되지 않는 구성을 사용하려는 경우 블루/그린 업데이트를 사용합니다. 완료 시간은 가변적이며 사용자가 제어합니다.

자세한 내용은 컴퓨팅 환경에 대한 블루/그린 업데이트 수행 단원을 참조하십시오.

올바른 업데이트 전략 선택

이 결정 가이드를 사용하여 필요에 가장 적합한 업데이트 전략을 선택합니다.

다음과 같은 경우 규모 조정 업데이트를 선택합니다.

컴퓨팅 용량(vCPU)만 조정해야 하는 경우 규모 조정 업데이트 전략을 선택합니다. 규모 조정 업데이트는 가동 중지 없이 빠른 업데이트가 필요하고 인프라 구성 변경이 필요하지 않은 경우에 이상적입니다.

자세한 절차는 규모 조정 업데이트 수행 섹션을 참조하세요.

다음과 같은 경우 인프라 업데이트를 선택합니다.

인스턴스 유형, AMI 설정, 서비스 역할, 환경 상태 또는 네트워킹 구성을 수정해야 하는 경우 인프라 업데이트 전략을 선택합니다. 환경은 AWSServiceRoleForBatch 서비스 연결 역할과 `BEST_FIT_PROGRESSIVE`, `SPOT_CAPACITY_OPTIMIZED` 또는 `SPOT_PRICE_CAPACITY_OPTIMIZED`의 할당 전략을 사용해야 합니다. 인프라 업데이트는 업데이트 중 일부 작업 중단이 허용되고 최신 Amazon ECS 최적화 AMI에 대한 자동 업데이트를 원하는 경우 잘 작동합니다.

자세한 절차는 인프라 업데이트 수행 섹션을 참조하세요.

다음과 같은 경우 블루/그린 업데이트를 선택합니다.

워크로드에 가동 중지 시간이 없어야 하거나 프로덕션 워크로드를 전환하기 전에 변경 사항을 테스트해야 하는 경우 블루/그린 업데이트 전략을 선택합니다. 이 접근 방식은 빠른 롤백 기능이 중요하거나, 환경에서 `BEST_FIT` 할당 전략을 사용하거나, 환경에서 AWSServiceRoleForBatch 서비스 연결 역할을 사용하지 않는 경우에 필수적입니다. 블루/그린 업데이트는 수동 업데이트가 필요한 사용자 지정 AMI를 사용하거나 대대적인 구성 변경이 필요할 때도 최고의 선택입니다.

자세한 절차는 컴퓨팅 환경에 대한 블루/그린 업데이트 수행 섹션을 참조하세요.

AMI 업데이트 고려 사항

AMIs 업데이트 방법은 컴퓨팅 환경 구성에 따라 다릅니다.

AWS Batch 제공된 기본 AMI를 최신으로 업데이트

AWS Batch 는 인프라 업데이트 중에 다음 조건이 모두 충족될 때 최신 Amazon ECS 최적화 AMI로 ??? 업데이트할 수 있습니다.

> **Note**
>
> 인프라 업데이트가 완료된 후 `updateToLatestImageVersion`은 `false`로 설정됩니다. 다른 업데이트를 시작하려면 `updateToLatestImageVersion`을 `true`로 설정해야 합니다.

- 컴퓨팅 환경은 AWSServiceRoleForBatch 서비스 연결 역할을 사용합니다.
- 할당 전략은 `BEST_FIT_PROGRESSIVE`, `SPOT_CAPACITY_OPTIMIZED`또는 로 설정됩니다`SPOT_PRICE_CAPACITY_OPTIMIZED`.
- `imageId`, `imageIdOverride`또는 시작 템플릿에는 AMI ID가 명시적으로 지정되지 않습니다.
- `updateToLatestImageVersion`는 `true`으로 설정됩니다.

블루/그린 배포를 사용한 AMI 업데이트

다음 시나리오에서 블루/그린 배포를 사용하여 AMI를 업데이트해야 합니다.

- 특정 버전의 Amazon ECS 최적화 AMI를 사용하는 경우.
- AMI ID가 다음 중 하나에 지정된 경우:
  - 시작 템플릿(템플릿을 업데이트하거나 제거해야 함).
  - `imageId` 파라미터입니다.
  - EC2 구성의 `imageIdOverride` 파라미터입니다.
- `BEST_FIT` 할당 전략을 사용하는 경우(인프라 업데이트를 지원하지 않음).
- AWSServiceRoleForBatch 서비스 연결 역할을 사용하지 않는 경우.

사용자 지정 AMI에 대한 AMI 업데이트

컴퓨팅 환경의 시작 템플릿에서 사용자 지정 AMI를 지정하면 `imageId` 파라미터 또는 EC2 구성의 `imageIdOverride` 파라미터 AWS Batch 가 인프라 업데이트 중에 사용자 지정 AMI를 자동으로 업데이트하지 않습니다. 컴퓨팅 환경 생성 중에 원래 사용된 파라미터에 새 ID를 지정하여 사용자 지정 AMI ID를 업데이트할 수 있습니다. AWS Batch제공 AMI를 사용하여 로 전환하려는 경우 컴퓨팅 환경 업데이트에서 사용자 지정 AMI ID를 제거하여 전환할 수 있습니다.

규모 조정 업데이트 수행

규모 조정 업데이트는 인스턴스를 추가하거나 제거하여 컴퓨팅 환경의 용량을 조정합니다. 이는 가장 빠른 업데이트 전략이며 기존 인스턴스를 교체할 필요가 없습니다. 규모 조정 업데이트는 모든 서비스 역할 유형 및 할당 전략에서 작동하므로 가장 유연한 업데이트 옵션입니다.

규모 조정 업데이트를 트리거하는 변경 사항

다음 설정만 수정하면 AWS Batch 에서 규모 조정 업데이트를 수행합니다. 이러한 설정을 다른 컴퓨팅 환경 설정과 함께 수정하면가 인프라 업데이트를 대신 AWS Batch 수행합니다.

다음 설정은 독점적으로 수정될 때 규모 조정 업데이트를 트리거합니다.

- `desiredvCpus` - 환경의 목표 vCPUs 수를 설정합니다.
- `maxvCpus` - 시작할 수 있는 최대 vCPU 수를 정의합니다.
- `minvCpus` - 유지할 최소 vCPU 수를 지정합니다.
- `minScaleDownDelayMinutes` - 작업이 완료된 후 컴퓨팅 환경에서 인스턴스를 AWS Batch 계속 실행하는 최소 시간(분)을 지정합니다.

> **Note**
>
> `minScaleDownDelayMinutes`는 인프라 업데이트 중에 교체되는 인스턴스에는 적용되지 않습니다.

Fargate 컴퓨팅 환경의 경우, 규모 조정 업데이트를 위해 다음 설정을 수정할 수도 있습니다.

- `securityGroupIds` - 컴퓨팅 환경의 보안 그룹 ID.
- `subnets` - 컴퓨팅 환경의 서브넷.

> **Note**
>
> AWS Batch 가 `desiredvCpus`를 동적으로 조정하므로를 사용하여 조정 업데이트를 시작하지 않는 것이 좋습니다`desiredvCpus`. 대신에 `minvCpus`를 업데이트해야 합니다. `desiredvCpus`를 업데이트할 때 값은 `minvCpus`와 `maxvCpus` 사이여야 합니다. 새 값은 현재 `desiredvCpus`보다 크거나 같아야 합니다. 자세한 내용은 the section called "`desiredvCpus` 설정을 업데이트할 때 나타나는 오류 메시지" 단원을 참조하십시오.

> **Important**
>
> 이러한 조정 설정을 다른 컴퓨팅 환경 설정(예: 인스턴스 유형, AMI IDs 또는 시작 템플릿)과 함께 수정하는 경우는 조정 업데이트 대신 인프라 업데이트를 AWS Batch 수행합니다. 인프라 업데이트는 더 오래 걸리며 기존 인스턴스를 대체할 수 있습니다.

Performing scaling updates using the AWS Management Console

1. https://console.aws.amazon.com/batch/ AWS Batch 콘솔을 엽니다.
2. 탐색 창에서 컴퓨팅 환경을 선택한 다음 컴퓨팅 환경 탭을 선택합니다.
3. 업데이트할 컴퓨팅 환경을 선택합니다.
4. 작업을 선택하고 편집을 선택합니다.
5. 규모 조정 업데이트를 지원하는 설정을 하나 이상 수정합니다. 예제:
   - 최소 vCPU에 최소 vCPU 수를 입력합니다.
   - 원하는 vCPU에 원하는 vCPU 수를 입력합니다.
   - 최대 vCPU에 최대 vCPU 수를 입력합니다.
6. 변경 사항 저장을 선택합니다.
7. 컴퓨팅 환경 상태를 모니터링합니다. 업데이트는 규모 조정 작업만 수반되므로 빠르게 완료됩니다.

Performing scaling updates using the AWS CLI

update-compute-environment 명령을 사용하여 규모 조정 업데이트를 수행합니다. 다음 두 예제는 일반적인 규모 조정 작업을 보여줍니다. 다음과 같은 규모 조정 업데이트를 지원하는 설정 중 하나 이상을 수정할 수 있습니다.

- 이 예제에서는 원하는 vCPU, 최소 vCPU 및 최대 vCPU를 업데이트합니다.

```
aws batch update-compute-environment \
    --compute-environment your-compute-environment-name \
    --compute-resources minvCpus=2,maxvCpus=8
```

규모 조정 업데이트 모니터링

AWS Batch 콘솔을 사용하여 조정 업데이트를 모니터링하여 컴퓨팅 환경 상태를 확인하고 인스턴스 수 및 vCPU 지표를 확인합니다. 명령과 AWS CLI 함께를 사용하여 상태를 describe-compute-environments 확인하고 인스턴스 수와 vCPU 값을 모니터링할 수도 있습니다.

인프라 업데이트 수행

인프라 업데이트는 컴퓨팅 환경의 인스턴스를 설정이 업데이트된 새 인스턴스로 대체합니다. 이 업데이트 전략은 규모 조정 업데이트보다 오래 걸리며 특정 서비스 역할 및 할당 전략 설정이 필요합니다. 인프라 업데이트는 서비스 가용성을 유지하면서 기본 컴퓨팅 환경 구성을 수정할 수 있는 방법을 제공합니다.

> **Important**
>
> 인프라 업데이트에는 AWSServiceRoleForBatch 서비스 연결 역할과 `BEST_FIT_PROGRESSIVE`, `SPOT_CAPACITY_OPTIMIZED` 또는 `SPOT_PRICE_CAPACITY_OPTIMIZED`의 할당 전략이 필요합니다. 환경이 이러한 요구 사항을 충족하지 않는 경우, 대신 블루/그린 업데이트를 사용합니다.

인프라 업데이트를 트리거하는 변경 사항

다음 설정 중 하나를 수정하면가 인프라 업데이트를 AWS Batch 수행합니다. 인프라 업데이트는 규모 조정 업데이트 설정과 함께 이러한 설정을 수정할 때도 발생합니다.

다음 설정은 인프라 업데이트를 트리거합니다.

컴퓨팅 구성

- `allocationStrategy` -가 인스턴스 유형을 AWS Batch 선택하는 방법을 결정합니다.
- `instanceTypes` - 어느 EC2 인스턴스 유형을 사용할지를 지정합니다.

- `bidPercentage` - 스팟 인스턴스에 대한 온디맨드 가격의 최대 백분율.
- `type` - 컴퓨팅 환경 유형(EC2 또는 SPOT).

AMI 및 시작 구성

- `imageId` - 인스턴스에 사용할 특정 AMI.
- `ec2Configuration` - `imageIdOverride`를 포함한 EC2 구성.
- `launchTemplate` - EC2 시작 템플릿 설정.
- `ec2KeyPair` - 인스턴스 액세스를 위한 SSH 키 페어.
- `updateToLatestImageVersion` - 자동 AMI 업데이트 설정.

네트워킹 및 보안

- `subnets` - 인스턴스가 시작되는 VPC 서브넷(EC2 컴퓨팅 환경용).
- `securityGroupIds` - 인스턴스의 보안 그룹(EC2 컴퓨팅 환경용).
- `placementGroup` - EC2 배치 그룹 구성.

기타 설정

- `instanceRole` - EC2 인스턴스에 대한 IAM 역할.
- `tags` - EC2 인스턴스에 적용되는 태그.

> **Important**
>
> 규모 조정 업데이트 설정(예: `desiredvCpus`, `maxvCpus` 또는 `minvCpus`)과 함께 인프라 업데이트 설정을 수정하면 AWS Batch 에서 인프라 업데이트를 수행합니다. 인프라 업데이트는 규모 조정 업데이트보다 더 오래 걸립니다.

인프라 업데이트 중 AMI 선택

인프라 업데이트 중에 AMI가 이 세 가지 설정 중 어디에 설정되었는지에 따라 컴퓨팅 환경의 AMI ID가 변경될 수 있습니다. AMI는 `imageId`(`computeResources`에서), `imageIdOverride`(`ec2Configuration`에서)에서 지정되거나 아니면 시작 템플릿

이 launchTemplate에 지정됩니다. AMI ID가 아무런 설정에도 지정되어 있지 않고 updateToLatestImageVersion 설정이 true라고 가정해 보겠습니다. 그런 다음에서 지원하는 최신 Amazon ECS 최적화 AMI AWS Batch 가 모든 인프라 업데이트에 사용됩니다.

AMI ID가 이러한 설정 중 하나 중에 지정되면 업데이트 전에 사용한 AMI ID 제공 설정에 따라 업데이트가 달라집니다. 컴퓨팅 환경을 생성할 때 AMI ID 선택 우선 순위는 가장 먼저 시작 템플릿, 그리고 imageId 설정, 마지막으로 imageIdOverride 설정입니다. 하지만 사용한 AMI ID를 시작 템플릿에서 가져오면 imageId 또는 imageIdOverride 설정은 AMI ID를 업데이트하지 않습니다. 시작 템플릿에 선택된 AMI ID를 업데이트하는 유일한 방법은 시작 템플릿을 업데이트하는 것입니다. 시작 템플릿의 버전 파라미터가 $Default 혹은 $Latest인 경우 지정된 시작 템플릿의 기본 버전 또는 최신 버전이 평가됩니다. 기본적으로 다른 AMI ID를 선택하거나 시작 템플릿의 최신 버전을 선택한 경우 해당 AMI ID가 업데이트에 사용됩니다.

시작 템플릿에 AMI ID를 선택하지 않은 경우 imageId 또는 imageIdOverride 파라미터에 지정된 AMI ID가 사용됩니다. 둘 다 지정된 경우 imageIdOverride 파라미터에 지정된 AMI ID가 사용됩니다.

컴퓨팅 환경에 또는 imageId, imageIdOverride, 혹은 launchTemplate 파라미터로 지정한 AMI ID를 사용하고 AWS Batch에서 지원하는 최신 Amazon ECS 최적화 AMI를 사용하기를 원한다고 가정해 보겠습니다. 그러면 업데이트는 AMI ID를 제공한 설정을 제거해야 합니다. imageId의 경우, 해당 파라미터에 빈 문자열을 지정해야 합니다. imageIdOverride의 경우 ec2Configuration 파라미터에 빈 문자열을 지정해야 합니다.

AMI ID가 시작 템플릿에서 가져온 경우 다음 방법 중 하나를 사용하여에서 지원하는 최신 Amazon ECS 최적화 AMI AWS Batch 로 변경할 수 있습니다.

- launchTemplateId 또는 launchTemplateName 파라미터에 빈 문자열을 지정하여 시작 템플릿을 제거합니다. 그러면 AMI ID만 제거되는 것이 아니라 전체 시작 템플릿이 제거됩니다.
- 업데이트된 버전의 시작 템플릿에 AMI ID가 지정되지 않은 경우 updateToLatestImageVersion 파라미터를 true로 설정해야 합니.

업데이트 중 작업 처리

업데이트 정책을 사용하여 인프라 업데이트 중에 실행 중인 작업을 처리하는 방법을 구성합니다. terminateJobsOnUpdate=true를 설정하면 실행 중인 작업이 즉시 종료되고 jobExecutionTimeoutMinutes 설정이 무시되며 인스턴스를 교체할 수 있게 되는 즉시 업데이트가 진행됩니다. terminateJobsOnUpdate=false를 설정하면 지정된 제한 시간(기본 제한 시간은 30분) 동안 실행 중인 작업이 계속 진행되고 제한 시간을 초과하면 작업이 종료됩니다.

> **Note**
>
> 업데이트 중에 종료된 작업을 재시도하려면 작업 재시도 전략을 구성합니다. 자세한 내용은 the section called “작업 자동 재시도” 단원을 참조하십시오.

Performing infrastructure updates using the AWS Management Console

> **Note**
>
> 콘솔에서 최신 AMI 버전으로 업데이트하려면 섹션을 참조하세요AMI 버전 업데이트.

1. https://console.aws.amazon.com/batch/ AWS Batch 콘솔을 엽니다.
2. 탐색 창에서 환경을 선택한 다음 컴퓨팅 환경 탭을 선택합니다.
3. 업데이트할 컴퓨팅 환경을 선택합니다.
4. 작업을 선택하고 편집을 선택합니다.
5. 업데이트 동작 섹션에서 실행 중인 작업을 처리하는 방법을 구성합니다.
   - 최신 버전으로 AMI 업데이트를 선택하여 AMI를 최신 버전으로 업데이트합니다.
   - 업데이트 프로세스가 실행된 후 작업을 종료하려면 업데이트 즉시 작업 종료를 선택합니다.
   - 작업 실행 제한 시간에 업데이트 프로세스를 시작하기 전에 대기할 시간(분)을 입력합니다.
6. 인프라 업데이트가 필요한 설정을 하나 이상 수정합니다. 예제:
   - 인스턴스 역할
   - EC2 스팟 인스턴스 사용
   - 허용되는 인스턴스 유형
   - 배치 그룹
   - EC2 키 페어
   - EC2 구성
   - 시작 템플릿
   - 서브넷
   - 보안 그룹
7. 변경 사항 저장을 선택합니다.

8. 컴퓨팅 환경 상태를 모니터링합니다. 업데이트 프로세스 동안 환경은 UPDATING으로 표시됩니다.

Performing infrastructure updates using the AWS CLI

update-compute-environment 명령을 사용하여 인프라 업데이트가 필요한 설정을 하나 이상 변경합니다. 다음 세 가지 예제는 흔히 수행되는 인프라 작업입니다.

- 이 예제는 인스턴스 유형을 업데이트하고 업데이트 정책을 구성합니다.

```
aws batch update-compute-environment \
    --compute-environment your-compute-environment-name \
    --compute-resources instanceTypes=default_x86_64 \
    --update-policy terminateJobsOnUpdate=false,jobExecutionTimeoutMinutes=30
```

- 이 예제는 VPC 서브넷 및 보안 그룹을 업데이트합니다.

```
aws batch update-compute-environment \
    --compute-environment your-compute-environment-name \
    --compute-resources subnets=subnet-abcd1234,subnet-efgh5678
 securityGroupIds=sg-abcd1234 \
    --update-policy terminateJobsOnUpdate=true
```

- 이 예제는 최신 Amazon ECS 최적화 AMI에 대한 자동 업데이트를 활성화합니다.

```
aws batch update-compute-environment \
    --compute-environment your-compute-environment-name \
    --compute-resources updateToLatestImageVersion=true \
    --update-policy terminateJobsOnUpdate=false,jobExecutionTimeoutMinutes=60
```

인프라 업데이트 모니터링

AWS Batch 콘솔을 사용하여 인프라 업데이트를 모니터링하여 로의 컴퓨팅 환경 상태 변경을 모니터링하고UPDATING, 인스턴스 교체 진행 상황을 모니터링하고, 실패한 업데이트가 있는지 확인합니다. 컴퓨팅 환경 상태가 VAILD가 되면 업데이트가 성공한 것입니다. CloudWatch를 사용하여 인스턴스 종료 이벤트를 추적하고 업데이트 동안 작업 상태를 모니터링할 수도 있습니다. 에서 describe-compute-environments 명령을 AWS CLI사용하여 상태를 확인하고 인스턴스 수명 주기 이벤트를 모니터링합니다.

컴퓨팅 환경에 대한 블루/그린 업데이트 수행

블루/그린 업데이트는 기존 컴퓨팅 환경(블루)과 함께 새 컴퓨팅 환경(그린)을 생성하여 가동 중지 시간과 리스크를 줄이는 업데이트 전략입니다. 이 접근 방식은 기존 환경의 작동을 유지하면서 워크로드를 새 환경으로 점진적으로 전환할 수 있게 해 줍니다. 블루/그린 업데이트는 가장 안전한 업데이트 경로를 제공하며 모든 서비스 역할 유형 또는 할당 전략과 작동합니다.

개요

블루/그린 업데이트는 프로덕션 환경에 적합한 몇 가지 이점을 제공합니다. 업데이트 프로세스 중에 워크로드의 지속적인 실행을 유지하여 가동 중지 시간을 0으로 유지합니다. 이 접근 방식을 사용하면 쉬운 롤백이 가능하여 문제가 발생했을 때 원래 환경으로 빠르게 되돌릴 수 있습니다. 점진적 전환 전략을 구현하여 프로덕션 워크로드를 완전히 전환하기 전에 새 환경의 성능을 확인할 수 있습니다. 또한 이 방법은 제거를 선택하기 전까지 원래 환경이 변경 없이 운영되므로 뛰어난 위험 완화 기능을 제공합니다.

블루/그린 업데이트가 필요한 경우

다음과 같은 상황에서는 블루/그린 업데이트를 사용해야 합니다.

- 컴퓨팅 환경에서 BEST_FIT 할당 전략을 사용하는 경우(인프라 업데이트를 지원하지 않음).
- 컴퓨팅 환경에서 AWSServiceRoleForBatch 서비스 연결 역할을 사용하지 않는 경우.
- 서로 다른 서비스 역할 유형 간에 전환해야 하는 경우.

블루/그린 업데이트가 권장되는 경우

워크로드에 가동 중지 시간이 없는 프로덕션 환경에는 블루/그린 업데이트를 사용하는 것이 좋습니다. 이 접근 방식은 프로덕션 워크로드를 전환하기 전에 새 구성을 테스트하여 변경 사항이 성능 및 신뢰성 요구 사항을 충족하는지 확인해야 할 때 효과적입니다. 특히 상당한 변경 사항이 있는 사용자 지정 AMI를 업데이트하는 경우 등 운영에 있어 빠른 롤백 기능이 중요할 때 블루/그린 업데이트를 선택합니다. 이 방법은 변경 사항을 완전히 적용하기 전에 성능 특성 및 동작을 검증하여 업데이트 프로세스에 대한 신뢰도를 제공하려는 경우에도 이상적입니다.

사전 조건

블루/그린 업데이트를 수행하기 전에 다음을 확인합니다.

- 컴퓨팅 환경을 생성하고 관리하기 위한 적절한 IAM 권한.
- 작업 대기열 설정을 보고 수정할 수 있는 액세스 권한.

- 전환 중에 발생할 수 있는 장애를 처리하기 위해 작업 정의에 대해 구성된 작업 재시도 전략 자세한 내용은 작업 자동 재시도 단원을 참조하십시오.
- 새 컴퓨팅 환경의 AMI ID. 이는 다음 중 하나일 수 있습니다.
  - Amazon ECS 최적화 AMI의 승인된 최신 버전(기본적으로 사용됨).
  - Amazon ECS 컨테이너 인스턴스 AMI 사양을 충족하는 사용자 지정 AMI 사용자 지정 AMI를 사용하는 경우 다음 방법 중 하나로 지정할 수 있습니다.
    - EC2 구성에서 이미지 ID 재정의 필드 사용.
    - 시작 템플릿에서 지정합니다.

    자체 사용자 지정 AMI 생성에 대한 자세한 정보는 자습서: 컴퓨팅 리소스 AMI 생성 섹션을 참조하세요.

새 환경을 생성하기 전에 기존 컴퓨팅 환경의 구성을 기록해 두어야 합니다. AWS Management Console 또는를 사용하여이 작업을 수행할 수 있습니다 AWS CLI.

> **Note**
>
> 다음 절차에서는 AMI만 변경하는 블루/그린 업데이트를 수행하는 방법을 자세히 설명합니다. 새 환경에 대한 다른 설정을 업데이트할 수 있습니다.

> **Important**
>
> 이전(블루) 컴퓨팅 환경을 제거하면 인스턴스가 종료되므로 해당 인스턴스에서 현재 실행 중인 모든 작업이 실패합니다. 이러한 실패를 자동 처리하도록 작업 정의에 작업 재시도 전략을 구성하세요. 자세한 내용은 작업 자동 재시도 단원을 참조하십시오.
> 새 환경에 대한 확신이 생기면:
>
> 1. 작업 대기열을 편집하여 이전 컴퓨팅 환경을 제거합니다.
> 2. 이전 환경에서 실행 중인 작업이 완료될 때까지 기다립니다.
> 3. 이전 컴퓨팅 환경을 삭제합니다.

Performing blue/green updates using the AWS Management Console

1. 현재 컴퓨팅 환경 복제

a. https://console.aws.amazon.com/batch/ AWS Batch 콘솔을 엽니다.
b. 기존 컴퓨팅 환경을 선택합니다.
c. 작업을 선택한 다음 복제를 선택합니다.
d. 이름에 해당 컴퓨팅 환경의 고유한 이름을 입력합니다.
e. 다음을 선택합니다.
f. 인스턴스 구성 섹션에서 AMI 설정을 업데이트합니다.
    i. 추가 구성을 확장합니다.
    ii. EC2 구성에서 이미지 유형에 새 AMI 유형을 지정하고 이미지 ID 재정의 필드에 AMI ID를 지정합니다.
g. 다음을 선택합니다.
h. 네트워크 구성에서 다음을 선택합니다.
i. 기존 환경에서 자동으로 복사되는 다른 설정을 검토합니다.
j. 컴퓨팅 환경 생성을 선택합니다.
k. 새 컴퓨팅 환경 상태가 `VALID`가 될 때까지 기다립니다.

2. 작업 대기열 순서 변경
    a. 탐색 창에서 작업 대기열을 선택합니다.
    b. 기존 컴퓨팅 환경과 연결된 작업 대기열을 선택합니다.
    c. 편집을 선택합니다.
    d. 연결된 컴퓨팅 환경에 새 컴퓨팅 환경을 추가합니다.
        - 기존 환경보다 순서 번호가 높게 새 컴퓨팅 환경을 추가하여 워크로드를 전환합니다.
        - 새 환경이 올바르게 작동하는지 확인한 후에는 더 낮은 순서 번호를 지정하여 기본 환경으로 만들 수 있습니다.
    e. 작업 대기열 업데이트를 선택합니다.
3. 정리
    a. 새 환경에서 작업 실행을 모니터링하여 모든 것이 예상대로 작동하는지 확인합니다.
    b. 새 환경에 대한 확신이 생기면:
        1. 작업 대기열을 편집하여 이전 컴퓨팅 환경을 제거합니다.
        2. 이전 환경에서 실행 중인 작업이 완료될 때까지 기다립니다.

3. 이전 컴퓨팅 환경을 삭제합니다.

Performing blue/green updates using the AWS CLI

1. 를 사용하여 구성을 가져오려면 다음 명령을 AWS CLI사용합니다.

```
aws batch describe-compute-environments \
  --compute-environments your-compute-environment-name
```

새 환경을 생성할 때 참조할 수 있도록 출력을 저장해 둡니다.

2. 기존 환경의 구성을 사용하되 새 AMI를 사용하여 새 컴퓨팅 환경을 생성합니다. 다음은 명령 구조의 예입니다.

예제 값을 이전 단계의 실제 구성으로 바꿉니다.

```
cat <<EOF > ./blue-green-compute-environment.json
{
  "computeEnvironmentName": "your-new-compute-environment-name",
  "type": "MANAGED",
  "state": "ENABLED",
  "computeResources": {
    "instanceRole": "arn:aws:iam::012345678901:instance-profile/
ecsInstanceRole",
    "type": "EC2",
    "minvCpus": 2,
    "desiredvCpus": 2,
    "maxvCpus": 256,
    "instanceTypes": [
      "optimal"
    ],
    "allocationStrategy": "BEST_FIT_PROGRESSIVE",
    "ec2Configuration": [
      {
        "imageType": "ECS_AL2023",
        "imageIdOverride": "ami-0abcdef1234567890"
      }
    ],
    "subnets": [,
      "subnet-0abcdef1234567890"
    ],
    "securityGroupIds": [
```

```
      "sg-0abcdef1234567890"
    ]
  }
}
EOF
```

```
$ aws batch create-compute-environment --cli-input-json file://./blue-green-compute-environment.json
```

3. 새 환경을 사용할 수 있을 때까지 기다립니다.

```
aws batch describe-compute-environments \
  --compute-environments your-new-compute-environment-name \
  --query 'computeEnvironments[].status'
```

4. 작업 대기열에 새 컴퓨팅 환경을 추가합니다.

```
aws batch update-job-queue \
  --job-queue your-job-queue \
  --compute-environment-order order=1,computeEnvironment=your-existing-environment \
  order=2,computeEnvironment=your-new-compute-environment-name
```

5. 확인되면 다시 업데이트하여 새 환경을 기본 환경으로 만듭니다.

```
aws batch update-job-queue \
  --job-queue your-job-queue \
  --compute-environment-order order=1,computeEnvironment=your-new-compute-environment-name
```

이전 환경에서 모든 작업이 완료되면 해당 환경을 비활성화한 다음 삭제합니다.

```
aws batch update-compute-environment \
    --compute-environment your-existing-environment \
    --state DISABLED
```

```
aws batch delete-compute-environment \
  --compute-environment your-existing-environment
```

컴퓨팅 리소스 AMI

기본적으로 AWS Batch 관리형 컴퓨팅 환경은 컴퓨팅 리소스에 대해 승인된 최신 버전의 Amazon ECS 최적화 AMI를 사용합니다. 하지만 여러분의 관리형 및 비 관리형 컴퓨팅 환경에 자체 AMI 만들어 사용할 수도 있습니다. 다음 중 해당하는 항목이 있다면 자체 AMI 생성을 권장합니다.

- AMI 루트 또는 데이터 볼륨의 스토리지 크기 늘리기.
- 지원되는 Amazon EC2 인스턴스 유형에 대한 인스턴스 스토리지 볼륨 추가.
- Amazon ECS 컨테이너 에이전트 사용자 지정.
- Docker를 커스터마이징 하고자 하는 경우
- 지원되는 Amazon EC2 인스턴스 유형에서 컨테이너가 GPU 하드웨어를 액세스할 수 있도록 GPU 워크로드 AMI를 구성하고 싶은 경우

> **Note**
>
> 컴퓨팅 환경이 생성된 후 AWS Batch 는 컴퓨팅 환경에서 AMIs를 업그레이드하지 않습니다. AWS Batch 또한 최신 버전의 Amazon ECS 최적화 AMIs를 사용할 수 있는 경우는 컴퓨팅 환경에서 AMI를 업데이트하지 않습니다. 게스트 운영 체제의 관리는 사용자의 책임입니다. 여기에는 모든 업데이트 및 보안 패치 적용이 포함됩니다. 또한 사용자는 컴퓨팅 리소스에 설치하는 추가 애플리케이션 소프트웨어 또는 유틸리티에 대해서도 책임이 있습니다. AWS Batch 작업에 새 AMI를 사용하려면 다음을 수행합니다.
>
> 1. 새 AMI로 새 컴퓨팅 환경을 생성합니다.
> 2. 기존 작업 대기열에 컴퓨팅 환경을 추가합니다.
> 3. 작업 대기열에서 이전 컴퓨팅 환경을 제거합니다.
> 4. 이전 컴퓨팅 환경을 삭제합니다.
>
> 2022년 4월에 컴퓨팅 환경 업데이트에 대한 향상된 지원이 AWS Batch 추가되었습니다. 자세한 내용은 에서 컴퓨팅 환경 업데이트 AWS Batch 단원을 참조하십시오. 컴퓨팅 환경의 향상된 업데이트를 사용하여 AMI를 업데이트하려면 다음 규칙을 따르세요.
>
> - 서비스 역할(`serviceRole`) 파라미터를 설정하지 않거나 AWSServiceRoleForBatch 서비스 연결 역할로 설정하세요.

- 할당 전략(allocationStrategy) 파라미터를 BEST_FIT_PROGRESSIVE, SPOT_CAPACITY_OPTIMIZED, 또는 SPOT_PRICE_CAPACITY_OPTIMIZED로 설정합니다.
- 최신 이미지 버전으로 업데이트(updateToLatestImageVersion) 파라미터를 true으로 설정합니다.
- imageId, imageIdOverride(ec2Configuration) 또는 시작 템플릿(launchTemplate)에 AMI ID를 지정하지 마세요. AMI ID를 지정하지 않으면는 인프라 업데이트가 시작될 때가 AWS Batch 지원하는 최신 Amazon ECS 최적화 AMI를 AWS Batch 선택합니다. 선택적으로 imageId 혹은 imageIdOverride 파라미터에 AMI ID를 지정할 수도 있습니다. 또는 LaunchTemplate 속성으로 식별된 시작 템플릿을 지정할 수 있습니다. 이러한 속성을 변경하면 인프라 업데이트가 시작됩니다. 시작 템플릿에 AMI ID가 지정되면 imageId 또는 imageIdOverride 파라미터에 AMI ID를 지정한다고 AMI ID를 바꿀 수 없습니다. AMI ID는 다른 시작 템플릿을 지정해야만 교체할 수 있습니다. 시작 템플릿 버전이 $Default 혹은 $Latest로 설정되면 시작 템플릿의 새 기본 버전을 설정하거나($Default의 경우), 새로운 버전의 시작 템플릿을 추가하여($Latest의 경우) AMI ID를 대체할 수 있습니다.

이러한 규칙을 준수하면 인프라 업데이트를 시작하는 업데이트 시 AMI ID가 다시 선택됩니다. 시작 템플릿(launchTemplate)의 version 설정이 $Latest 또는 $Default으로 설정된 경우, launchTemplate이 업데이트되어 있지 않더라도 인프라 업데이트 시 시작 템플릿의 최신 버전 또는 기본 버전이 평가됩니다.

주제

- 컴퓨팅 리소스 AMI 사양
- AMI 선택 순서
- 컴퓨팅 환경에서 AMI 버전 관리
- 자습서: 컴퓨팅 리소스 AMI 생성
- GPU 워크로드 AMI 사용
- Amazon Linux 사용 중단
- Amazon EKS Amazon Linux 2 AMI 사용 중단
- Amazon ECS Amazon Linux 2 AMI 사용 중단

컴퓨팅 리소스 AMI 사양

기본 AWS Batch 컴퓨팅 리소스 AMI 사양은 다음으로 구성됩니다.

필수

- HVM 가상화 유형 AMI에서 Linux 커널 버전 3.10 이상을 실행하는 최신 Linux 배포. Windows 컨테이너는 지원되지 않습니다.

  > **Important**
  >
  > 다중 노드 병렬 작업은 `ecs-init` 패키지가 설치되어 있는 Amazon Linux 인스턴스에서 시작된 컴퓨팅 리소스에서만 실행될 수 있습니다. 자신만의 컴퓨팅 환경을 생성할 때는 기본 Amazon ECS 최적화 AMI를 사용하는 것이 좋습니다. 사용자 지정 AMI를 지정하지 않으면 이 작업이 가능합니다. 자세한 내용은 다중 노드 병렬 작업 단원을 참조하십시오.

- Amazon ECS 컨테이너 에이전트. 최신 버전을 사용하는 것이 좋습니다. 자세한 내용은 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS 컨테이너 에이전트 설치를 참조하세요.
- Amazon ECS 컨테이너 에이전트를 시작할 때 `ECS_AVAILABLE_LOGGING_DRIVERS` 환경 변수에 `awslogs` 로그 드라이버를 사용 가능한 로그 드라이버로 지정해야 합니다. 자세한 내용을 알아보려면 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS 컨테이너 에이전트 구성을 참조하세요.
- 버전 1.9 이상에서 실행하는 도커 대몬(daemon) 및 도커 런타임 종속성. 자세한 내용은 Docker 설명서의 런타임 종속성 확인을 참조하세요.

  > **Note**
  >
  > 도커 버전은 사용 중인 해당 Amazon ECS 에이전트 버전으로 테스트해 보는 것이 좋습니다. Amazon ECS는 GitHub에서 Amazon ECS 최적화 AMI의 Linux 변형에 대한 변경 로그를 제공합니다. 자세한 정보는 Changelog를 참조하세요.

권장

- Amazon ECS 에이전트를 실행하고 모니터링할 초기화 및 nanny 프로세스. Amazon ECS 최적화 AMI는 `ecs-init` 업스타트 프로세스를 사용하지만 기타 운영 체제는 `systemd`를 사용할 수 있습니다. 자세한 내용은 Amazon Elastic Container Service 개발자 안내서의 컨테이너 인스턴스 사용자 데이터 구성 스크립트 예제를 참조하세요. `ecs-init`에 대한 자세한 내용은 GitHub에서 `ecs-init`

프로젝트를 참조하세요. 관리형 컴퓨팅 환경에서는 부팅 시, 최소한 Amazon ECS 에이전트를 시작해야 합니다. Amazon ECS 에이전트가 컴퓨팅 리소스에서 실행되지 않는 경우에서 작업을 수락할 수 없습니다 AWS Batch.

Amazon ECS 최적화 AMI는 이러한 요구 사항 및 권장 사항을 미리 구성하여 놓았습니다. Amazon ECS 최적화 AMI 혹은 Amazon Linux AMI는 사용자 컴퓨팅 리소스에 설치된 `ecs-init` 패키지와 함께 운용하는 것이 좋습니다. 사용자 애플리케이션이 특정 운영체제를 요구하거나 해당 AMI에서 아직 사용할 수 없는 도커 버전이 필요하다면 다른 AMI를 선택하세요. 자세한 내용은 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS 최적화 AMI를 참조하세요.

AMI 선택 순서

AWS Batch 는 다음 우선 순위에 따라 컴퓨팅 리소스의 Amazon Machine Image(AMI)를 결정합니다. 이 순서를 이해하면가 컴퓨팅 환경에 특정 AMI를 AWS Batch 선택한 이유를 이해하는 데 도움이 됩니다.

1. 시작 템플릿 재정의 AMI - 시작된 인스턴스에 대한 시작 템플릿 재정의에 이미지가 있는 경우 해당 이미지가 사용됩니다.
2. 컴퓨팅 리소스 이미지 ID(사용되지 않음) - 설정하면이 컴퓨팅 환경 AMI가 사용됩니다. 참고: 더 이상 사용되지 않는 필드입니다. 대신 ec2Configuration.imageIdOverride를 사용합니다.
3. EC2 구성 이미지 ID 재정의 - 지정된 경우이 필드의 이미지가 사용됩니다.
4. 시작 템플릿 AMI - 컴퓨팅 환경에 이미지와 연결된 시작 템플릿이 있는 경우이 이미지가 사용됩니다.
5. AWS 기본 AMI - 위의 항목 중 하나라도 구성되지 않은 경우 ec2Configuration에서 지정된 imageType을 기반으로 기본 AMI를 AWS Batch 선택합니다.

> **Note**
>
> ec2Configuration 파라미터는 선택 사항입니다. 생략하면 컴퓨팅 환경에서 시작된 인스턴스 유형에 따라 적절한 ec2Configuration 및 기본 AMI를 AWS Batch 자동으로 선택합니다.

> **Note**
>
> 이 AMI 선택 순서는 Fargate 컴퓨팅 환경에 적용되지 않습니다.

우선 순위가 가장 높은 AMI 선택 순서부터 가장 낮은 AMI 선택 순서까지

1. 시작 템플릿 재정의 AMI(가장 높은 우선 순위)

   API 필드: 대상 인스턴스 유형 `overrides[].launchTemplateId` 포함

   참조: LaunchTemplateSpecification

   템플릿 재정의는 특정 인스턴스 유형을 대상으로 하며 기본 시작 템플릿보다 더 세분화된 제어를 제공합니다. 일치하는 인스턴스 유형에 대해 다른 모든 AMI 사양보다 우선합니다.

```
{
  "computeResources": {
    "launchTemplate": {
      "launchTemplateId": "lt-default",
      "overrides": [
        {
          "launchTemplateId": "lt-gpu-optimized",
          "targetInstanceTypes": ["p3.2xlarge", "g4dn.xlarge"]
        }
      ]
    }
  }
}
```

2. 컴퓨팅 리소스 이미지 ID

   API 필드: `computeResources.imageId`

   참조: CreateComputeEnvironment

   컴퓨팅 환경 수준에서 직접 AMI를 지정할 수 있습니다. 이는 EC2 구성 재정의 및 시작 템플릿(재정의 템플릿 제외)보다 우선합니다.

   여러 EC2 구성(예: `ECS_AL2023` 및 `ECS_AL2023_NVIDIA`용)이 있는 컴퓨팅 환경에서는 여기에 지정된 AMI ID가 모든 EC2 구성에 사용됩니다.

   > **Important**
   >
   > `imageId` 필드는 더 이상 사용되지 않습니다. `ec2Configuration.imageIdOverride` 대신를 사용하십시오.

```
{
  "computeResources": {
    "imageId": "ami-12345678",
    "instanceTypes": ["m5.large", "m5.xlarge"]
  }
}
```

3. EC2 구성 이미지 ID 재정의

   API 필드: `computeResources.ec2Configuration[].imageIdOverride`

   참조: Ec2Configuration

   EC2 구성은 이미지 유형별 재정의를 제공합니다. 이 설정은 지정된 이미지 유형에 대한 기본 AMI 선택 및 시작 템플릿 AMI를 재정의합니다.

```
{
  "computeResources": {
    "ec2Configuration": [
      {
        "imageType": "ECS_AL2",
        "imageIdOverride": "ami-87654321"
      }
    ]
  }
}
```

4. 시작 템플릿 AMI

   API 필드: Amazon EC2 시작 템플릿`ImageId`의

   참조: 에서 Amazon EC2 시작 템플릿 사용 AWS Batch

   시작 템플릿에서 AMI를 지정하면 기본 AMI 선택보다 우선하지만 더 높은 우선 순위 설정으로 재정의됩니다.

```
// EC2 Launch Template content
{
  "LaunchTemplateName": "my-batch-template",
  "LaunchTemplateData": {
    "ImageId": "ami-12345678"
```

```
  }
}
```

AWS Batch 시작 템플릿에서 참조:

```
// Batch Launch Template content
{
  "computeResources": {
    "launchTemplate": {
      "launchTemplateName": "my-batch-template",
      "version": "$Latest"
    }
  }
}
```

5. AWS 기본 AMI(최저 우선 순위)

   API 필드: 결정자 `computeResources.ec2Configuration[].imageType`

   참조: Ec2Configuration imageType

   사용자 지정 AMI가 지정되지 않은 경우는 이미지 유형에 따라 승인된 최신 Amazon ECS 최적화 AMI를 AWS Batch 자동으로 선택합니다.

   > **Note**
   >
   > 는 선택 사항`ec2Configuration`입니다. AWS Batch 를 지정하지 않으면 `ec2Configuration`가 적절한 기본 AMI를 선택합니다.

```
{
  "computeResources": {
    "ec2Configuration": [
      {
        "imageType": "ECS_AL2023"
      }
    ]
  }
}
```

컴퓨팅 환경에서 AMI 버전 관리

AWS Batch 는 컴퓨팅 환경에서 사용하는 Amazon Machine Image(AMIs)에 대한 가시성을 제공합니다.

AMI 상태 보기

콘솔을 통해 AWS Batch 또는 describe-compute-environments 명령을 사용하여 컴퓨팅 환경에서 사용되는 AMIs의 상태를 볼 수 있습니다.

Console

AWS Batch 콘솔에서 AMI 상태 정보는 다음 상태 값과 함께 두 위치에 표시됩니다.

- 최신 -에서 지원하는 최신 AMI를 사용합니다 AWS Batch.
- 업데이트 사용 가능 - 업데이트를 사용할 수 있습니다.

> **Note**
>
> AMI 상태 정보는 AWS Batch관리형 AMIs에 대해서만 표시됩니다. 이미지가 기본 시작 템플릿에서 `imageId` (사용되지 않음), `imageIdOverride`또는에 지정되어 있으면 상태가 표시되지 않습니다. 컴퓨팅 환경에 시작 템플릿 재정의가 있는 경우 상태가 표시되지 않습니다. AMI 선택에 대한 자세한 내용은 섹션을 참조하세요AMI 선택 순서.

컴퓨팅 환경 페이지

컴퓨팅 환경 페이지에는 `batchImageStatus` 각 컴퓨팅 환경의 전체를 보여주는 배치 이미지 상태 열이 표시됩니다. 컴퓨팅 환경에 여러 AMIs가 있고 하나의 AMI에 사용 가능한 업데이트가 있는 경우 콘솔에 전체 컴퓨팅 환경에 사용 가능한 업데이트가 표시됩니다.

> **Note**
>
> 컴퓨팅 환경이 모든 이미지 유형에 대한 크기 조정을 시작한 후 상태가 나타납니다.

컴퓨팅 환경 세부 정보 페이지

컴퓨팅 환경 세부 정보 페이지의 컴퓨팅 리소스 탭의 Ec2 구성 섹션에는 컴퓨팅 환경의 각 이미지 유형에 대한 배치 이미지 상태가 표시됩니다. 이미지 유형에 여러 AMIs가 있고 하나의 AMI에 사용 가능한 업데이트가 있는 경우 콘솔에 해당 이미지 유형에 사용 가능한 업데이트가 표시됩니다.

> **Note**
>
> 컴퓨팅 환경이 해당 특정 이미지 유형에 대한 인스턴스 조정을 시작한 후에만 각 이미지 유형에 대한 상태가 나타납니다.

CLI

describe-compute-environments를 호출하면 응답에 다음 값과 함께 AMI 가시성을 제공하는 `batchImageStatus` 필드가 포함됩니다.

- `LATEST` -에서 지원하는 최신 AMI를 사용합니다 AWS Batch.
- `UPDATE_AVAILABLE` - 업데이트를 사용할 수 있습니다.

> **Note**
>
> `batchImageStatus` 필드는 AWS Batch관리형 AMIs에만 표시됩니다. 사용자 지정 AMIs 기본 시작 템플릿에서 `imageId` (사용되지 않음), `imageIdOverride`또는에 지정된 경우에는 표시되지 않습니다. 컴퓨팅 환경에 시작 템플릿 재정의가 있는 경우 상태가 표시되지 않습니다. 가 AMI를 AWS Batch 선택하는 방법에 대한 자세한 내용은 섹션을 참조하세요AMI 선택 순서. AMIs
> 필드는 컴퓨팅 환경이 해당를 사용하여 인스턴스를 조정하기 시작한 후 에`Ec2Configuration`만 각에 대해 독립적으로 표시됩니다`imageType`.

```
{
    "computeEnvironments": [
        {
            "computeEnvironmentName": "my-compute-environment",
            "computeResources": {
                "ec2Configuration": [
```

```
                    {
                        "imageType": "ECS_AL2023"
                    },
                    {
                        "imageType": "ECS_AL2023_NVIDIA",
                        "batchImageStatus": "UPDATE_AVAILABLE"
                    }
                ]
            }
        }
    ]
}
```

AMI 버전 업데이트

가 AMI 업데이트를 사용할 수 있음을 AWS Batch 나타내는 경우 AMIs를 최신 버전으로 업데이트가 true로 설정된 컴퓨팅 환경을 업데이트하여 최신 AMI를 사용하도록 컴퓨팅 환경을 업데이트할 수 있습니다.

새 AMI IDs를 지정할 필요가 없습니다.는 AMIs 업데이트를 최신 버전으로 설정할 때 지원되는 최신 AMI를 AWS Batch 자동으로 선택합니다.

Important

AMIs 업데이트하면 조정 업데이트가 아닌 인프라 업데이트가 트리거됩니다. 즉,는 기존 인스턴스를 업데이트된 AMI를 사용하는 새 인스턴스로 AWS Batch 대체합니다. 업데이트 프로세스는 조정 업데이트보다 오래 걸리며 업데이트 정책 구성에 따라 실행 중인 작업을 중단할 수 있습니다.

Important

할당 전략이 BEST_FIT인 경우 블루/그린 업데이트를 수행해야 합니다.

Console

AWS Batch 콘솔AMIs를 업데이트하려면:

1. https://console.aws.amazon.com/batch/ AWS Batch 콘솔을 엽니다.
2. 탐색 창에서 환경을 선택합니다.
3. 업데이트와 함께 AMI 상태를 표시하는 컴퓨팅 환경을 선택합니다.
4. 지금 업데이트(AMI 상태 기준) 또는 작업 > 편집을 선택하여 업데이트 모달을 엽니다.
5. AMI 업데이트 모달에서 현재 AMI 버전과 상태를 검토합니다.
6. 확인 또는 저장을 선택하여 인프라 업데이트를 시작합니다.

인프라 업데이트 UPDATING 중에 컴퓨팅 환경 상태가 로 변경됩니다. 콘솔에서 진행 과정을 모니터링할 수 있습니다.

CLI

AWS CLIAMIs를 업데이트하려면 `update-compute-environment` 명령을 사용합니다.

```
aws batch update-compute-environment \
    --compute-environment my-compute-environment \
    --compute-resources updateToLatestImageVersion=true
```

이 명령은 최신 AWS Batch지원 AMIs.

사용자 지정 AMI 고려 사항

컴퓨팅 환경에서 사용자 지정 AMIs, 즉 `ComputeResources.imageId` (사용되지 않음)에 지정된 AMIs를 사용하는 경우 `Ec2Configuration.imageIdOverride`, 기본 시작 템플릿 또는 시작 템플릿 재정의는 이러한 AMIs에 대한 상태 정보를 제공할 AWS Batch 수 없습니다.

- 상태 가시성 - 사용자 지정 AMIs 콘솔의 배치 이미지 상태에 대해 "-"를 표시하고 API 응답에 `batchImageStatus` 필드를 포함하지 않습니다.
- 수동 관리 - 사용자 지정 AMIs. AMI 공급자의 보안 및 소프트웨어 패치에 대한 최신 정보를 확인하고 그에 따라 사용자 지정 AMIs 업데이트합니다.
- EC2 관리 - Amazon EC2 콘솔 또는 APIs 사용하여 새 버전 생성 및 이전 버전 사용 중단 등 사용자 지정 AMI 수명 주기를 관리합니다.

사용자 지정 AMIsthe section called “컴퓨팅 리소스 AMI”.

AMI 업데이트 모범 사례

이 섹션은 사용자 지정 AMIs 모두에 적용됩니다.

- 정기 모니터링 - 컴퓨팅 환경의 AMI 상태를 정기적으로 확인하여 업데이트를 사용할 수 있는 시기를 식별합니다. 기본 AMIs 경우 업데이트를 사용할 수 있는 시기가 `batchImageStatus`에 표시됩니다. 사용자 지정 AMIs 경우 AWS 보안 게시판과 같은 다른 리소스를 사용해야 합니다.
- 유지 관리 기간 - 인프라 업데이트가 기존 인스턴스를 대체하므로 작업 중단이 허용되는 유지 관리 기간 동안 AMI 업데이트를 예약합니다.
- 작업 재시도 전략 - 인프라 업데이트 중에 중단될 수 있는 작업을 처리하도록 작업 재시도 전략을 구성합니다. 자세한 내용은 the section called “작업 자동 재시도” 단원을 참조하십시오.
- 업데이트 정책 구성 - 인프라 업데이트 중에 실행 중인 작업이 처리되는 방식을 제어하도록 적절한 업데이트 정책을 구성합니다. 자세한 내용은 the section called “인프라 업데이트 수행” 단원을 참조하십시오.
- 테스트 - 프로덕션 컴퓨팅 환경에 적용하기 전에 개발 환경에서 AMI 업데이트를 테스트합니다.

자습서: 컴퓨팅 리소스 AMI 생성

사용자 지정 컴퓨팅 리소스를 직접 생성하여 관리형 및 비관리형 컴퓨팅 환경에 사용할 수 있습니다. 지침은 다음(컴퓨팅 리소스 AMI 사양)을 참조하세요. 사용자 지정 AMI를 생성한 후에는 이 AMI를 사용하는 컴퓨팅 환경을 생성하여 작업 대기열에 연결할 수 있습니다. 마지막으로 해당 대기열에 작업 제출을 시작합니다.

사용자 지정 컴퓨팅 리소스 AMI 생성하기

1. 시작할 기본 AMI를 선택합니다 기본 AMI는 HVM 가상화를 사용해야 합니다. 기본 AMI는 Windows AMI가 될 수 없습니다.

> **Note**
>
> 컴퓨팅 환경에 대해 선택한 AMI는 해당 컴퓨팅 환경에 사용할 인스턴스 유형의 아키텍처와 일치해야 합니다. 예를 들어, 컴퓨팅 환경에서 A1 인스턴스 유형을 사용하는 경우 선택한 컴퓨팅 리소스 AMI는 ARM 인스턴스를 지원해야 합니다. Amazon ECS는 Amazon ECS에 최적화된 Amazon Linux 2 AMI의 x86 버전과 ARM 버전을 모두 제공합니다. 자세한 내용을 알아보려면 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS 최적화 Amazon Linux 2 AMI를 참조하세요.

Amazon ECS 최적화 Amazon Linux 2 AMI는 관리형 컴퓨팅 환경의 컴퓨팅 리소스에 대한 기본 AMI입니다. Amazon ECS 최적화 Amazon Linux 2 AMI는 AWS 엔지니어가에서 사전 구성하고 테스트 AWS Batch 합니다. 를 시작하고에서 AWS 빠르게 실행되는 컴퓨팅 리소스를 가져올 수 있는 최소한의 AMI입니다. 자세한 내용은 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS 최적화 AMI를 참조하세요.

아니면 다른 Amazon Linux 2을 선택하고 다음 명령을 사용하여 `ecs-init` 패키지를 설치할 수 있습니다. 자세한 내용은 Amazon Elastic Container Service 개발자 안내서의 Amazon Linux 2 ECS 컨테이너 인스턴스에 Amazon ECS 컨테이너 에이전트 설치를 참조하세요.

```
$ sudo amazon-linux-extras disable docker
$ sudo amazon-linux-extras install ecs-init
```

예를 들어 AWS Batch 컴퓨팅 리소스에서 GPU 워크로드를 실행하려는 경우 Amazon Linux 딥 러닝 AMI로 시작할 수 있습니다. 그런 다음 AWS Batch 작업을 실행하도록 AMI를 구성합니다. 자세한 내용은 GPU 워크로드 AMI 사용 단원을 참조하십시오.

> **Important**
>
> 사용자는 `ecs-init` 패키지를 지원하지 않는 기본 AMI를 선택할 수 있습니다. 하지만 그렇게 할 경우 부팅 시 Amazon ECS 에이전트를 시작하고 계속 실행되도록 하는 방법을 구성해야 합니다. 또한 Amazon ECS 컨테이너 에이전트를 시작하고 모니터링하는 데 `systemd`를 사용하는 몇 가지 예제 사용자 데이터 구성 스크립트도 볼 수 있습니다. 자세한 내용은 Amazon Elastic Container Service 개발자 안내서의 예제 컨테이너 인스턴스 사용자 데이터 구성 스크립트를 참조하세요.

2. 선택한 기본 AMI에서 해당 AMI에 적절한 스토리지 옵션을 사용하여 인스턴스를 시작합니다. 연결된 Amazon EBS 볼륨의 크기와 수를 구성하거나 선택한 인스턴스 유형이 이 인스턴스를 지원하는 경우에는 인스턴스 스토리지 볼륨의 크기와 수를 구성할 수 있습니다. 자세한 내용은 Amazon EC2 사용 설명서에서 인스턴스 시작 및 Amazon EC2 인스턴스 저장소를 참조하세요.
3. 인스턴스를 SSH에 연결하고 필요한 구성 작업을 수행합니다. 다음 단계 중 하나 또는 전체가 포함될 수 있습니다.

   - Amazon ECS 컨테이너 에이전트 설치 자세한 내용은 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS 컨테이너 에이전트 설치를 참조하세요.
   - 인스턴스 스토어 볼륨을 포맷하기 위한 스크립트 구성

- 부팅 시 인스턴스 스토어 볼륨 또는 Amazon EFS 파일 시스템이 탑재될 수 있도록 `/etc/fstab` 파일에 추가
- 디버깅 설정, 기본 이미지 크기 조정 등 도커 옵션 구성.
- 패키지 설치 또는 파일 복사

자세한 내용은 Amazon EC2 사용 설명서의 SSH를 사용하여 Linux 인스턴스에 연결을 참조하세요.

4. 인스턴스에서 Amazon ECS 컨테이너 에이전트를 시작한 경우 AMI를 생성하기 전에 중지하고 영구 데이터 체크 포인트 파일을 모두 제거해야 합니다. 이렇게 하지 않으면 AMI에서 시작된 인스턴스에서 에이전트가 시작되지 않습니다.

   a. Amazon ECS 컨테이너 에이전트를 중지합니다.

      - Amazon ECS 최적화 Amazon Linux 2 AMI:

        ```
        sudo systemctl stop ecs
        ```

      - Amazon ECS 최적화 Amazon Linux AMI:

        ```
        sudo stop ecs
        ```

   b. 영구적인 데이터 체크포인트 파일을 제거합니다. 기본적으로 `/var/lib/ecs/data/` 디렉터리에 파일이 위치합니다. 다음 명령을 사용하여 해당 파일을 제거합니다.

      ```
      sudo rm -rf /var/lib/ecs/data/*
      ```

5. 실행 중인 인스턴스에서 AMI를 새로 만듭니다. 자세한 내용은 Amazon EC2 사용 설명서의 Amazon EBS 지원 Linux AMI 생성을 참조하세요.

에서 새 AMI를 사용하려면 AWS Batch

1. 새 AMI가 생성되면 새로운 AMI 컴퓨팅 환경을 생성합니다. 이렇게 하려면 AWS Batch 컴퓨팅 환경을 생성할 때 이미지 유형을 선택하고 이미지 ID 재정의 상자에 사용자 지정 AMI ID를 입력합니다. 자세한 내용은 the section called “자습서: Amazon EC2 리소스를 사용하여 관리형 컴퓨팅 환경 생성” 단원을 참조하십시오.

> **Note**
>
> 컴퓨팅 환경에 대해 선택한 AMI는 해당 컴퓨팅 환경에 사용할 인스턴스 유형의 아키텍처와 일치해야 합니다. 예를 들어, 컴퓨팅 환경에서 A1 인스턴스 유형을 사용하는 경우 선택한 컴퓨팅 리소스 AMI는 ARM 인스턴스를 지원해야 합니다. Amazon ECS는 Amazon ECS에 최적화된 Amazon Linux 2 AMI의 x86 버전과 ARM 버전을 모두 제공합니다. 자세한 내용을 알아보려면 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS 최적화 Amazon Linux 2 AMI를 참조하세요.

2. 작업 대기열을 만들고 새 컴퓨팅 환경에 연결합니다. 자세한 내용은 작업 대기열 생성 단원을 참조하십시오.

> **Note**
>
> 작업 대기열과 연관된 모든 컴퓨팅 환경은 동일한 아키텍처를 공유해야 합니다. AWS Batch (은)는 단일 작업 대기열에서의 혼합 컴퓨팅 환경 아키텍처 유형을 지원하지 않습니다.

3. (선택 사항) 새 작업 대기열에 샘플 작업을 제출합니다. 자세한 내용은 작업 정의 예, 단일 노드 작업 정의 생성, 자습서: 작업 제출 섹션을 참조하세요.

GPU 워크로드 AMI 사용

AWS Batch 컴퓨팅 리소스에서 GPU 워크로드를 실행하려면 GPU 지원이 포함된 AMI를 사용해야 합니다. 자세한 내용을 알아보려면 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS의 GPU 작업과 Amazon ECS 최적화 AMI를 참조하세요.

관리형 컴퓨팅 환경에서 컴퓨팅 환경이 p3, p4, p5, p6, g3, g3s, g4, g5 또는 g6 인스턴스 유형이나 인스턴스 패밀리를 지정하는 경우, AWS Batch는 Amazon ECS GPU 최적화 AMI를 사용합니다.

비관리형 컴퓨팅 환경에서는 Amazon ECS GPU 최적화 AMI가 권장됩니다. AWS Command Line Interface 또는 AWS Systems Manager 파라미터 스토어 GetParameter, GetParameters, GetParametersByPath 작업을 사용하여 권장되는 Amazon ECS GPU 최적화 AMI에 대한 메타데이터를 검색할 수 있습니다.

> **Note**
>
> p5 인스턴스 패밀리는 Amazon ECS GPU 최적화 AM의 20230912(와)과 같거나 이후 버전에서만 지원되며, p2 및 g2 인스턴스 유형과는 호환되지 않습니다. p5 인스턴스를 사용해야 하는 경우 컴퓨팅 환경에 p2 또는 g2 인스턴스가 포함되어 있지 않고 최신 기본 Batch AMI를 사용하는지 확인하세요. 새 컴퓨팅 환경을 생성하면 최신 AMI가 사용되지만, p5(을)를 포함하도록 컴퓨팅 환경을 업데이트하는 경우 `ComputeResource` 속성에서 `updateToLatestImageVersion`(을)를 `true`(으)로 설정하여 최신 AMI를 사용하고 있는지 확인할 수 있습니다. 자세한 내용을 알아보려면 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS의 GPU 작업을 참조하세요.

다음 예제에서는 GetParameter 명령을 사용하는 방법을 보여줍니다.

AWS CLI

```
$ aws ssm get-parameter --name /aws/service/ecs/optimized-ami/amazon-linux-2/gpu/
recommended \
                        --region us-east-2 --output json
```

출력에는 `Value` 파라미터의 AMI 정보가 포함됩니다.

```
{
    "Parameter": {
        "Name": "/aws/service/ecs/optimized-ami/amazon-linux-2/gpu/recommended",
        "LastModifiedDate": 1555434128.664,
        "Value": "{\"schema_version\":1,\"image_name\":\"amzn2-ami-ecs-gpu-
hvm-2.0.20190402-x86_64-ebs\",\"image_id\":\"ami-083c800fe4211192f\",\"os\":\"Amazon
 Linux 2\",\"ecs_runtime_version\":\"Docker version 18.06.1-ce\",\"ecs_agent_version
\":\"1.27.0\"}",
        "Version": 9,
        "Type": "String",
        "ARN": "arn:aws:ssm:us-east-2::parameter/aws/service/ecs/optimized-ami/
amazon-linux-2/gpu/recommended"
    }
}
```

Python

```
from __future__ import print_function
```

```
import json
import boto3

ssm = boto3.client('ssm', 'us-east-2')

response = ssm.get_parameter(Name='/aws/service/ecs/optimized-ami/amazon-linux-2/
gpu/recommended')
jsonVal = json.loads(response['Parameter']['Value'])
print("image_id   = " + jsonVal['image_id'])
print("image_name = " + jsonVal['image_name'])
```

출력에는 AMI ID와 AMI 이름만 포함됩니다.

```
image_id   = ami-083c800fe4211192f
image_name = amzn2-ami-ecs-gpu-hvm-2.0.20190402-x86_64-ebs
```

다음 예제에서는 GetParameters의 사용에 대해 설명합니다.

AWS CLI

```
$ aws ssm get-parameters --names  /aws/service/ecs/optimized-ami/amazon-linux-2/gpu/
recommended/image_name \
                                  /aws/service/ecs/optimized-ami/amazon-linux-2/gpu/
recommended/image_id \
                         --region us-east-2 --output json
```

출력에는 각 파라미터에 대한 전체 메타데이터가 포함됩니다.

```
{
    "InvalidParameters": [],
    "Parameters": [
        {
            "Name": "/aws/service/ecs/optimized-ami/amazon-linux-2/gpu/recommended/
image_id",
            "LastModifiedDate": 1555434128.749,
            "Value": "ami-083c800fe4211192f",
            "Version": 9,
            "Type": "String",
            "ARN": "arn:aws:ssm:us-east-2::parameter/aws/service/ecs/optimized-ami/
amazon-linux-2/gpu/recommended/image_id"
```

```
        },
        {
            "Name": "/aws/service/ecs/optimized-ami/amazon-linux-2/gpu/recommended/
image_name",
            "LastModifiedDate": 1555434128.712,
            "Value": "amzn2-ami-ecs-gpu-hvm-2.0.20190402-x86_64-ebs",
            "Version": 9,
            "Type": "String",
            "ARN": "arn:aws:ssm:us-east-2::parameter/aws/service/ecs/optimized-ami/
amazon-linux-2/gpu/recommended/image_name"
        }
    ]
}
```

Python

```
from __future__ import print_function

import boto3

ssm = boto3.client('ssm', 'us-east-2')

response = ssm.get_parameters(
            Names=['/aws/service/ecs/optimized-ami/amazon-linux-2/gpu/recommended/
image_name',
                   '/aws/service/ecs/optimized-ami/amazon-linux-2/gpu/recommended/
image_id'])
for parameter in response['Parameters']:
    print(parameter['Name'] + " = " + parameter['Value'])
```

출력에는 이름의 전체 경로를 사용하여 AMI ID와 AMI 이름이 포함됩니다.

```
/aws/service/ecs/optimized-ami/amazon-linux-2/gpu/recommended/image_id =
 ami-083c800fe4211192f
/aws/service/ecs/optimized-ami/amazon-linux-2/gpu/recommended/image_name = amzn2-
ami-ecs-gpu-hvm-2.0.20190402-x86_64-ebs
```

다음 예제에서는 GetParametersByPath 명령을 사용하는 방법을 보여 줍니다.

AWS CLI

```
$ aws ssm get-parameters-by-path --path /aws/service/ecs/optimized-ami/amazon-
linux-2/gpu/recommended \
                                --region us-east-2 --output json
```

출력에는 지정된 경로에 있는 모든 파라미터에 대한 전체 메타데이터가 포함됩니다.

```
{
    "Parameters": [
        {
            "Name": "/aws/service/ecs/optimized-ami/amazon-linux-2/gpu/recommended/
ecs_agent_version",
            "LastModifiedDate": 1555434128.801,
            "Value": "1.27.0",
            "Version": 8,
            "Type": "String",
            "ARN": "arn:aws:ssm:us-east-2::parameter/aws/service/ecs/optimized-ami/
amazon-linux-2/gpu/recommended/ecs_agent_version"
        },
        {
            "Name": "/aws/service/ecs/optimized-ami/amazon-linux-2/gpu/recommended/
ecs_runtime_version",
            "LastModifiedDate": 1548368308.213,
            "Value": "Docker version 18.06.1-ce",
            "Version": 1,
            "Type": "String",
            "ARN": "arn:aws:ssm:us-east-2::parameter/aws/service/ecs/optimized-ami/
amazon-linux-2/gpu/recommended/ecs_runtime_version"
        },
        {
            "Name": "/aws/service/ecs/optimized-ami/amazon-linux-2/gpu/recommended/
image_id",
            "LastModifiedDate": 1555434128.749,
            "Value": "ami-083c800fe4211192f",
            "Version": 9,
            "Type": "String",
            "ARN": "arn:aws:ssm:us-east-2::parameter/aws/service/ecs/optimized-ami/
amazon-linux-2/gpu/recommended/image_id"
        },
        {
            "Name": "/aws/service/ecs/optimized-ami/amazon-linux-2/gpu/recommended/
image_name",
```

```
            "LastModifiedDate": 1555434128.712,
            "Value": "amzn2-ami-ecs-gpu-hvm-2.0.20190402-x86_64-ebs",
            "Version": 9,
            "Type": "String",
            "ARN": "arn:aws:ssm:us-east-2::parameter/aws/service/ecs/optimized-ami/
amazon-linux-2/gpu/recommended/image_name"
        },
        {
            "Name": "/aws/service/ecs/optimized-ami/amazon-linux-2/gpu/recommended/
os",
            "LastModifiedDate": 1548368308.143,
            "Value": "Amazon Linux 2",
            "Version": 1,
            "Type": "String",
            "ARN": "arn:aws:ssm:us-east-2::parameter/aws/service/ecs/optimized-ami/
amazon-linux-2/gpu/recommended/os"
        },
        {
            "Name": "/aws/service/ecs/optimized-ami/amazon-linux-2/gpu/recommended/
schema_version",
            "LastModifiedDate": 1548368307.914,
            "Value": "1",
            "Version": 1,
            "Type": "String",
            "ARN": "arn:aws:ssm:us-east-2::parameter/aws/service/ecs/optimized-ami/
amazon-linux-2/gpu/recommended/schema_version"
        }
    ]
}
```

Python

```
from __future__ import print_function

import boto3

ssm = boto3.client('ssm', 'us-east-2')

response = ssm.get_parameters_by_path(Path='/aws/service/ecs/optimized-ami/amazon-
linux-2/gpu/recommended')
for parameter in response['Parameters']:
    print(parameter['Name'] + " = " + parameter['Value'])
```

출력에는 이름의 전체 경로를 사용하여 지정된 경로에 있는 모든 파라미터 이름의 값이 포함됩니다.

```
/aws/service/ecs/optimized-ami/amazon-linux-2/gpu/recommended/ecs_agent_version =
 1.27.0
/aws/service/ecs/optimized-ami/amazon-linux-2/gpu/recommended/ecs_runtime_version =
 Docker version 18.06.1-ce
/aws/service/ecs/optimized-ami/amazon-linux-2/gpu/recommended/image_id =
 ami-083c800fe4211192f
/aws/service/ecs/optimized-ami/amazon-linux-2/gpu/recommended/image_name = amzn2-
ami-ecs-gpu-hvm-2.0.20190402-x86_64-ebs
/aws/service/ecs/optimized-ami/amazon-linux-2/gpu/recommended/os = Amazon Linux 2
/aws/service/ecs/optimized-ami/amazon-linux-2/gpu/recommended/schema_version = 1
```

자세한 내용을 알아보려면 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS 최적화 AMI 메타데이터 검색을 참조하세요.

Amazon Linux 사용 중단

Amazon Linux AMI(Amazon Linux 1이라고도 함)는 2023년 12월 31일에 수명이 종료되었습니다. AWS Batch 는 2024년 1월 1일부터 보안 업데이트 또는 버그 수정을 받지 못하므로 Amazon Linux AMI에 대한 지원을 종료했습니다. Amazon Linux 수명 주기 종료에 대한 자세한 내용은 AL FAQ를 참조하세요.

예기치 않은 워크로드 중단을 방지하고 보안 및 기타 업데이트를 계속 받을 수 있도록 기존 Amazon Linux 기반 컴퓨팅 환경을 Amazon Linux 2023으로 업데이트하는 것이 좋습니다.

Amazon Linux AMI를 사용하는 컴퓨팅 환경은 수명 주기 종료 날짜인 2023년 12월 31일 이후에도 계속 작동할 수 있습니다. 그러나 이러한 컴퓨팅 환경은 더 이상 새로운 소프트웨어 업데이트, 보안 패치 또는 버그 수정을 받지 않습니다 AWS. 수명 주기 종료 후 Amazon Linux AMI에서 이러한 컴퓨팅 환경을 유지하는 것은 사용자의 책임입니다. 최적의 성능과 보안을 유지하려면 AWS Batch 컴퓨팅 환경을 Amazon Linux 2023 또는 Amazon Linux 2로 마이그레이션하는 것이 좋습니다.

Amazon Linux AMI AWS Batch 에서 Amazon Linux 2023 또는 Amazon Linux 2로 마이그레이션하는 데 도움이 필요하면 컴퓨팅 환경 업데이트 -를 참조하세요 AWS Batch.

Amazon EKS Amazon Linux 2 AMI 사용 중단

AWS 는 2025년 11월 26일에 Amazon EKS 최적화 Amazon Linux 2 AMIs에 대한 지원을 종료했습니다. 2025년 10월 27일에는 새 Amazon EKS 컴퓨팅 환경의 기본 AMI를 Amazon Linux 2023으로 AWS

Batch 변경했습니다. Amazon EKS 최적화 Amazon Linux 2 AMIs 더 이상 소프트웨어 업데이트, 보안 패치 또는 버그 수정을 수신하지 않습니다 AWS.

> **Important**
>
> 아직 AWS Batch Amazon Linux 2를 사용하는 Amazon EKS 컴퓨팅 환경이 있는 경우 Amazon Linux 2023으로 마이그레이션하는 것이 좋습니다. end-of-life 후 Amazon EKS 최적화 Amazon Linux 2 컴퓨팅 환경을 유지하는 것은 사용자의 책임입니다.

Amazon EKS AL2 수명 주기 종료에 대한 자세한 내용은 Amazon EKS 사용 설명서의 Amazon EKS AMI 사용 중단 FAQ를 참조하세요.

AWS Batch Amazon Linux 2에서 Amazon Linux 2023으로 Amazon EKS 컴퓨팅 환경을 마이그레이션하는 데 도움이 필요하면 섹션을 참조하세요EKS AL2에서 EKS AL2023으로 업그레이드하는 방법.

Amazon ECS Amazon Linux 2 AMI 사용 중단

AWS 는 2026년 6월 30일에 Amazon ECS Amazon Linux 2 최적화 및 가속화된 AMIs에 대한 지원을 종료합니다. 2026년 1월 12일에 새 Amazon ECS 컴퓨팅 환경의 기본 AMI를 Amazon Linux 2에서 Amazon Linux 2023으로 AWS Batch 변경했습니다. 2026년 6월 30일부터 배치 제공 Amazon Linux 2 AMI를 사용하여 새 Amazon ECS 컴퓨팅 환경 생성을 AWS Batch 차단합니다. AMIs 이 날짜 이후에는 Amazon Linux 2023 또는 고객 제공 AMIs.

> **Important**
>
> 2026년 6월 30일 이전에 기존 AWS Batch Amazon ECS 컴퓨팅 환경을 Amazon Linux 2023으로 마이그레이션하는 것이 좋습니다. 기존 컴퓨팅 환경은이 날짜 이후에도 계속 작동하지만에서 소프트웨어 업데이트, 보안 패치 또는 버그 수정을 더 이상 받지 않습니다 AWS. end-of-life 후 Amazon Linux 2 컴퓨팅 환경을 유지하는 것은 사용자의 책임입니다.

상태 계획 수명 주기 이벤트를 사용하여 AWS 영향을 받는 Amazon ECS 컴퓨팅 환경의 마이그레이션 상태를 추적할 수 있습니다. 자세한 내용은 AWS 계획된 상태 수명 주기 이벤트 단원을 참조하십시오.

Amazon Linux 2 수명 주기 종료에 대한 자세한 내용은 Amazon Linux 2 FAQ를 참조하세요.

Amazon Linux 2 및 Amazon Linux 2023 운영 체제의 차이점에 대한 자세한 내용은 Amazon Linux 2023 사용 설명서의 Amazon Linux 2023 및 Amazon Linux 2 비교를 참조하세요.

Amazon ECS 최적화 AMI를 위한 Amazon Linux 2023의 변경 사항에 대한 자세한 내용은 Amazon ECS 사용 설명서의 Amazon Linux 2에서 Amazon Linux 2023 Amazon ECS 최적화 AMI로 마이그레이션을 참조하세요.

AWS Batch Amazon Linux 2에서 Amazon Linux 2023으로 Amazon ECS 컴퓨팅 환경을 마이그레이션하는 데 도움이 필요하면 섹션을 참조하세요ECS AL2에서 ECS AL2023으로 마이그레이션하는 방법.

에서 Amazon EC2 시작 템플릿 사용 AWS Batch

AWS Batch 는 Amazon EC2 EC2 시작 템플릿을 사용할 수 있도록 지원합니다. 시작 템플릿을 사용하면 사용자 지정 AMI를 생성할 필요 없이 AWS Batch 컴퓨팅 리소스의 기본 구성을 수정할 수 있습니다. AMIs

시작 템플릿은 AMIs 선택 순서에 우선하는 AMI를 지정할 수 있습니다. 자세한 내용은 AMI 선택 순서 단원을 참조하십시오.

> **Note**
>
> AWS Batch 컴퓨팅 환경에 대한 사용자 지정 시작 템플릿을 지정할 때 AWS Batch 는 기본 Auto Scaling 그룹에 시작 템플릿을 직접 연결하지 않습니다. 대신는 Auto Scaling 그룹에 대한 별도의 배치 관리형 시작 템플릿을 AWS Batch 생성하고 사용자 지정 시작 템플릿의 관련 설정을 여기에 통합합니다. 따라서 Auto Scaling 그룹과 연결된 시작 템플릿이 원래 지정한 시작 템플릿과 다릅니다. 이는 예상된 동작입니다.

> **Note**
>
> 시작 템플릿은 AWS Fargate 리소스에서 지원되지 않습니다.

시작 템플릿을 컴퓨팅 환경과 연결하려면 먼저 시작 템플릿을 생성해야 합니다. 사용자는 Amazon EC2 콘솔에서 시작 템플릿을 생성할 수 있습니다. 또는 AWS CLI 또는 AWS SDK를 사용할 수 있습니다. 예를 들어 다음 JSON 파일은 기본 AWS Batch 컴퓨팅 리소스 AMI의 Docker 데이터 볼륨 크기를 조정하고 암호화되도록 설정하는 시작 템플릿을 나타냅니다.

```
{
    "LaunchTemplateName": "increase-container-volume-encrypt",
    "LaunchTemplateData": {
        "BlockDeviceMappings": [
```

```
            {
                "DeviceName": "/dev/xvda",
                "Ebs": {
                    "Encrypted": true,
                    "VolumeSize": 100,
                    "VolumeType": "gp2"
                }
            }
        ]
    }
}
```

라는 파일에 JSON을 저장`lt-data.json`하고 다음 AWS CLI 명령을 실행하여 이전 시작 템플릿을 생성할 수 있습니다.

```
aws ec2 --region <region> create-launch-template --cli-input-json file://lt-data.json
```

시작 템플릿에 대한 자세한 내용은 Amazon EC2 사용 설명서의 시작 템플릿에서 인스턴스 시작을 참조하세요.

시작 템플릿을 사용하여 컴퓨팅 환경을 생성하는 경우 다음 기존 컴퓨팅 환경 파라미터를 시작 템플릿으로 이동할 수 있습니다.

> **Note**
>
> 이러한 파라미터 중 어느 하나가(Amazon EC2 태깅 제외)가 시작 템플릿과 컴퓨팅 환경 구성 둘 다에서 지정되었다고 가정해 보겠습니다. 이 경우, 컴퓨팅 환경 파라미터가 우선합니다. Amazon EC2 태그는 시작 템플릿과 컴퓨팅 환경 구성 사이에 병합됩니다. 태그의 키 값이 충돌하는 경우, 컴퓨팅 환경 구성의 값이 우선적으로 적용됩니다.

- Amazon EC2 키 페어
- Amazon EC2 AMI ID
- 보안 그룹 ID
- Amazon EC2 태그

AWS Batch다음 시작 템플릿 파라미터는 무시됩니다.

- 인스턴스 유형(컴퓨팅 환경을 생성할 때 원하는 인스턴스 유형을 지정)

- 인스턴스 역할(컴퓨팅 환경을 생성할 때 원하는 인스턴스 역할을 지정)
- 네트워크 인터페이스 서브넷(컴퓨팅 환경을 생성할 때 원하는 서브넷을 지정)
- 인스턴스 시장 옵션(스팟 인스턴스 구성을 제어AWS Batch 해야 함)
- API 종료 비활성화(인스턴스 수명 주기를 제어해야AWS Batch 함)

AWS Batch 는 인프라 업데이트 중에 새 시작 템플릿 버전으로만 시작 템플릿을 업데이트합니다. 자세한 내용은 에서 컴퓨팅 환경 업데이트 AWS Batch 단원을 참조하십시오.

기본 및 재정의 시작 템플릿

컴퓨팅 환경의 기본 시작 템플릿과 특정 인스턴스 유형 및 패밀리를 위한 재정의 시작 템플릿을 정의할 수 있습니다. 이는 컴퓨팅 환경의 대부분의 인스턴스 유형에 기본 템플릿이 사용되도록 하여 유용할 수 있습니다.

특정 버전의 이름을 지정하는 대신 대체 변수 `$Default` 및 `$Latest`를 사용할 수 있습니다. 재정의 시작 템플릿을 제공하지 않는 경우, 기본 시작 템플릿이 자동으로 적용됩니다.

`$Default` 또는 `$Latest` 변수를 사용하는 경우는 컴퓨팅 환경이 생성될 때 현재 정보를 AWS Batch 적용합니다. 향후 기본 또는 최신 버전이 변경되는 경우 UpdateComputeEnvironment 또는 AWS Management Console -를 통해 정보를 업데이트해야 합니다 AWS Batch.

추가적인 유연성을 위해 특정 컴퓨팅 인스턴스 유형 또는 패밀리에 적용되는 재정의 시작 템플릿을 정의할 수 있습니다.

> **Note**
>
> 각 컴퓨팅 환경에 최대 10개의 재정의 시작 템플릿을 지정할 수 있습니다.

`targetInstanceTypes` 파라미터를 사용하여 이 재정의 시작 템플릿을 사용해야 하는 인스턴스 유형 또는 패밀리를 선택합니다. 인스턴스 유형 또는 패밀리는 먼저 `instanceTypes` 파라미터에 의해 식별되어야 합니다.

시작 템플릿 재정의를 정의하고 나중에 제거하기로 결정하는 경우, 빈 배열을 전달하여 `UpdateComputeEnvironment` API 작업에서 `overrides` 파라미터를 설정 해제할 수 있습니다. `UpdateComputeEnvironment` API 작업을 제출할 때 `overrides` 파라미터를 포함하지 않도록 선택할 수도 있습니다. 자세한 내용은 단원을 참조하십시오. `LaunchTemplateSpecification.overrides`

자세한 내용은 AWS Batch API 참조 가이드LaunchTemplateSpecificationOverride.targetInstanceTypes의 섹션을 참조하세요.

시작 템플릿의 Amazon EC2 사용자 데이터

사용자는 인스턴스를 시작할 때 cloud-init에 의해 실행되는 시작 템플릿에 Amazon EC2 사용자 데이터를 제공할 수 있습니다. 사용자 데이터는 다음 사항을 포함하여 제한 없이 일반적인 구성 시나리오를 수행할 수 있습니다.

- 사용자 또는 그룹 포함
- 패키지 설치
- 파티션 및 파일 시스템 생성

시작 템플릿의 Amazon EC2 사용자 데이터는 MIME 멀티파트 아카이브 형식이어야 합니다. 이는 사용자 데이터가 컴퓨팅 리소스를 구성하는 데 필요한 다른 AWS Batch 사용자 데이터와 병합되기 때문입니다. 여러 사용자 데이터 블록을 단일 MIME 멀티파트 파일로 결합할 수 있습니다. 예를 들어, 도커 대몬(daemon)을 구성하는 클라우드 boothook를 Amazon ECS 컨테이너 에이전트에 대한 구성 정보를 작성하는 사용자 데이터 쉘 스크립트와 결합할 수 있습니다.

AWS CloudFormation를 사용하는 경우 AWS::CloudFormation::Init 유형을 cfn-init 헬퍼 스크립트와 함께 사용하여 일반적인 구성 시나리오를 수행할 수 있습니다.

MIME 멀티파트 파일은 다음과 같은 구성 요소로 이루어집니다.

- 콘텐츠 유형 및 부분 경계 선언: `Content-Type: multipart/mixed; boundary="==BOUNDARY=="`
- MIME 버전 선언: `MIME-Version: 1.0`
- 다음 구성 요소를 포함하는 하나 이상의 사용자 데이터 블록:
  - 사용자 데이터 블록의 시작을 나타내는 시작 경계: `--==BOUNDARY==` 이 경계 앞의 라인은 비워 두어야 합니다.
  - 블록에 대한 콘텐츠 유형 선언: `Content-Type: text/cloud-config; charset="us-ascii"`. 콘텐츠 유형에 대한 자세한 내용은 Cloud-Init 설명서를 참조하세요. 콘텐츠 유형 선언 뒤의 라인은 비워 두어야 합니다.
  - 쉘 명령 또는 `cloud-init` 지시어 목록 등의 사용자 데이터 콘텐츠
- MIME 멀티파트 파일의 끝을 나타내는 종료 경계: `--==BOUNDARY==--` 종료 경계 앞의 라인은 비워 두어야 합니다.

Note

Amazon EC2 콘솔에서 시작 템플릿에 사용자 데이터를 추가하는 경우 사용자 데이터를 일반 텍스트처럼 붙여 넣을 수 있습니다. 또는 파일에서 업로드할 수도 있습니다. AWS CLI 또는 AWS SDK를 사용하는 경우 먼저 사용자 데이터를 base64 인코딩하고이 JSON 파일에 표시된 대로 CreateLaunchTemplate을 호출할 때 해당 문자열을 UserData 파라미터 값으로 제출해야 합니다.

```
{
    "LaunchTemplateName": "base64-user-data",
    "LaunchTemplateData": {
        "UserData":
 "ewogICAgIkxhdW5jaFRlbXBsYXRlTmFtZSI6ICJpbmNyZWFzZS1jb250YWluZXItdm9sdW..."
    }
}
```

예약 Amazon ECS 에이전트 구성 값

AWS Batch 는 Amazon ECS를 사용하는 EC2-based 관리형 컴퓨팅 환경에서 인스턴스가 시작될 / etc/ecs/ecs.config 때 특정 Amazon ECS 에이전트 구성 값을에 기록합니다. AWS Batch 는 컴퓨팅 환경 업데이트 중에 이러한 값을 덮어쓸 수도 있으므로 변경 사항이 지속되지 않을 수 있습니다.

Important

사용자 데이터와 AWS Batch관리형 값이 충돌하면 작업이 예약되지 않고 예기치 않은 조정 동작이 발생하며 예기치 않은 비용이 발생할 수 있습니다.

Important

a AWS Batch관리형 시작 템플릿의 사용자 데이터를 자체 시작 템플릿으로 복사하지 마십시오. AWS Batch관리형 사용자 데이터에는 컴퓨팅 환경 간에 다르고 컴퓨팅 환경이 업데이트될 때 변경되는 값이 포함됩니다. 시작 템플릿에 복사된 값이 포함된 경우 현재 컴퓨팅 환경의 예상 값과 충돌할 수 있습니다.

시작 템플릿 사용자 데이터에서 다음 값을 설정하지 마십시오. 이러한 값은에서 관리 AWS Batch 하며 충돌로 인해 예약 및 조정 문제가 발생할 수 있습니다.

- `ECS_CLUSTER`
- `ECS_INSTANCE_ATTRIBUTES`

위에 나열된 파라미터에 대한 자세한 내용은 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS 컨테이너 에이전트 구성을 참조하세요.

주제

- 참조: Amazon EC2 시작 템플릿 예제

참조: Amazon EC2 시작 템플릿 예제

다음은 고유한 템플릿을 생성하는 데 사용할 수 있는 MIME 멀티파트 파일의 예입니다.

예제

- 예: 기존 Amazon EFS 파일 시스템 탑재
- 예: 기본 Amazon ECS 컨테이너 에이전트 구성 기본값 재정의
- 예: 기존 Amazon FSx for Lustre 파일 시스템 탑재

예: 기존 Amazon EFS 파일 시스템 탑재

Example

이 예제 MIME 멀티파트 파일은 `amazon-efs-utils` 패키지를 설치하고 `/mnt/efs`에 기존 Amazon EFS 파일 시스템을 탑재하도록 컴퓨팅 리소스를 구성합니다.

```
MIME-Version: 1.0
Content-Type: multipart/mixed; boundary="==MYBOUNDARY=="

--==MYBOUNDARY==
Content-Type: text/cloud-config; charset="us-ascii"

packages:
- amazon-efs-utils
```

```
runcmd:
- file_system_id_01=fs-abcdef123
- efs_directory=/mnt/efs

- mkdir -p ${efs_directory}
- echo "${file_system_id_01}:/ ${efs_directory} efs tls,_netdev" >> /etc/fstab
- mount -a -t efs defaults

--==MYBOUNDARY==--
```

예: 기본 Amazon ECS 컨테이너 에이전트 구성 기본값 재정의

Example

이 예제 MIME 멀티파트 파일은 컴퓨팅 리소스에 대한 기본 도커 이미지 정리 설정을 재정의합니다.

```
MIME-Version: 1.0
Content-Type: multipart/mixed; boundary="==MYBOUNDARY=="

--==MYBOUNDARY==
Content-Type: text/x-shellscript; charset="us-ascii"

#!/bin/bash
echo ECS_IMAGE_CLEANUP_INTERVAL=60m >> /etc/ecs/ecs.config
echo ECS_IMAGE_MINIMUM_CLEANUP_AGE=60m >> /etc/ecs/ecs.config

--==MYBOUNDARY==--
```

예: 기존 Amazon FSx for Lustre 파일 시스템 탑재

Example

이 예제 MIME 멀티파트 파일은 Extras Library에서 `lustre2.10` 패키지를 설치하고 `/scratch`에 기존 FSx for Lustre 파일 시스템과 `fsx`의 이름을 탑재하도록 컴퓨팅 리소스를 구성합니다. 이 예제는 Amazon Linux 2의 예제입니다. 다른 Linux 배포판의 설치 가이드는 Amazon FSx for Lustre의 Lustre 사용 설명서의 Lustre 클라이언트 설치를 참조하세요. 자세한 내용은 Amazon FSx for Lustre 사용 설명서의 Amazon FSx 파일 시스템 자동 탑재를 참조하세요.

```
MIME-Version: 1.0
Content-Type: multipart/mixed; boundary="==MYBOUNDARY=="
```

```
--===MYBOUNDARY==
Content-Type: text/cloud-config; charset="us-ascii"

runcmd:
- file_system_id_01=fs-0abcdef1234567890
- region=us-east-2
- fsx_directory=/scratch
- amazon-linux-extras install -y lustre2.10
- mkdir -p ${fsx_directory}
- mount -t lustre ${file_system_id_01}.fsx.${region}.amazonaws.com@tcp:fsx
 ${fsx_directory}

--===MYBOUNDARY==--
```

컨테이너 속성의 volumes 및 mountPoints 멤버에서 탑재 지점을 컨테이너에 매핑해야 합니다.

```
{
    "volumes": [
        {
            "host": {
                "sourcePath": "/scratch"
            },
            "name": "Scratch"
        }
    ],
    "mountPoints": [
        {
            "containerPath": "/scratch",
            "sourceVolume": "Scratch"
        }
    ],
}
```

인스턴스 메타데이터 서비스(IMDS) 구성

인스턴스 메타데이터 서비스(IMDS)는 EC2 인스턴스에 대한 메타데이터를 해당 인스턴스에서 실행되는 애플리케이션에 제공합니다. 향상된 보안을 위해 모든 새 워크로드에 IMDSv2를 사용하고 기존 워크로드를 IMDSv1에서 IMDSv2로 마이그레이션합니다. IMDS 및 IMDS 구성에 대한 자세한 내용은 Amazon EC2 사용 설명서의 인스턴스 메타데이터를 사용하여 EC2 인스턴스 관리 및 새 인스턴스에 대한 인스턴스 메타데이터 옵션 구성을 참조하세요.

구성 시나리오

컴퓨팅 환경 설정에 따라 적절한 구성 방법을 선택합니다.

시작 템플릿이 없는 기본 AMI

기본 AWS Batch AMI를 사용하고 시작 템플릿을 지정하지 않는 경우 다음 옵션 중 하나를 선택합니다.

1. Amazon Linux 2023 기본 AMI 사용 - Amazon Linux 2023에는 기본적으로 IMDSv2가 필요합니다. 컴퓨팅 환경을 생성할 때 Amazon Linux 2023을 이미지 유형으로 선택합니다.
2. 계정 수준 IMDSv2 구성 설정 - 모든 새 인스턴스에 IMDSv2를 요구하도록 AWS 계정을 구성합니다. 이 설정은 계정에서 시작하는 모든 새 인스턴스에 영향을 미칩니다. 지침은 Amazon EC 2 사용 설명서의 IMDSv2를 계정의 기본값으로 설정을 참조하세요.

> **Note**
>
> 계정 수준 IMDS 구성은 시작 템플릿 또는 AMI 구성에 의해 재정의될 수 있습니다. 시작 템플릿 설정은 계정 수준 설정보다 우선합니다.

시작 템플릿이 없는 사용자 지정 AMI

시작 템플릿 없이 사용자 지정 AMI를 사용하는 경우 다음 옵션 중 하나를 선택합니다.

1. Amazon Linux 2023을 기본으로 사용 - Amazon Linux 2023을 기본 이미지로 사용하여 사용자 지정 AMI를 구축합니다. Batch를 위한 사용자 지정 AMI를 생성하는 방법에 대한 자세한 내용은 자습서: 컴퓨팅 리소스 AMI 생성 섹션을 참조하세요.
2. 사용자 지정 AMI에서 IMDSv2 구성 - 사용자 지정 AMI를 생성할 때 IMDSv2가 요구되도록 구성합니다. 자세한 내용은 Amazon EC2 사용 설명서의 사용자 지정 AMI에 대한 인스턴스 메타데이터 옵션 구성을 참조하세요.
3. 계정 수준 IMDSv2 구성 설정 - 모든 새 인스턴스에 IMDSv2를 요구하도록 AWS 계정을 구성합니다. 이 설정은 계정에서 시작하는 모든 새 인스턴스에 영향을 미칩니다. 지침은 Amazon EC 2 사용 설명서의 IMDSv2를 계정의 기본값으로 설정을 참조하세요.

> **Note**
>
> 계정 수준 IMDS 구성은 시작 템플릿 또는 AMI 구성에 의해 재정의될 수 있습니다. 시작 템플릿 설정은 계정 수준 설정보다 우선합니다.

시작 템플릿 사용

컴퓨팅 환경에서 시작 템플릿을 사용할 때, IMDSv2를 요구하도록 시작 템플릿에 메타데이터 옵션을 추가합니다. Batch에 시작 템플릿을 사용하는 방법에 대한 자세한 내용은 에서 Amazon EC2 시작 템플릿 사용 AWS Batch 섹션을 참조하세요.

```
{
    "LaunchTemplateName": "batch-imdsv2-template",
    "VersionDescription": "IMDSv2 only template for Batch",
    "LaunchTemplateData": {
        "MetadataOptions": {
            "HttpTokens": "required"
        }
    }
}
```

AWS CLI를 사용하여 시작 템플릿을 생성합니다.

```
aws ec2 create-launch-template --cli-input-json file://imds-template.json
```

EC2 구성

AWS Batch는 EC2 및 EC2 스팟 컴퓨팅 환경에 Amazon ECS에 최적화된 AMI를 사용합니다. 기본값은 Amazon Linux 2(ECS_AL2)입니다. 2026년 1월부터 기본값이 AL2023(ECS_AL2023)으로 변경됩니다.

AWS는 Amazon Linux 2에 대한 지원을 종료합니다. 최적의 성능과 보안을 유지하려면 AWS Batch Amazon ECS 컴퓨팅 환경을 Amazon Linux 2023으로 마이그레이션하는 것이 좋습니다. 자세한 내용은 Amazon ECS Amazon Linux 2 AMI 사용 중단 섹션을 참조하세요.

예기치 않은 워크로드 중단을 방지하고 보안 및 기타 업데이트를 계속 받을 수 있도록 기존 Amazon Linux 기반 컴퓨팅 환경을 Amazon Linux 2023으로 업데이트하는 것이 좋습니다.

AWS Batch를 Amazon Linux AMI에서 Amazon Linux 2023으로 마이그레이션하는 데 도움이 필요하면 ECS AL2에서 ECS AL2023으로 마이그레이션하는 방법 섹션을 참조하세요.

주제

- ECS AL2에서 ECS AL2023으로 마이그레이션하는 방법

ECS AL2에서 ECS AL2023으로 마이그레이션하는 방법

AL2023은 클라우드 애플리케이션에 안전하고 안정적이면서 고성능의 환경을 제공하도록 설계된 Linux 기반 운영 체제입니다. AL2와 AL2023의 차이점에 대한 자세한 내용은 Amazon Linux 2023 사용 설명서에서 Amazon Linux 2023과 Amazon Linux 2 비교를 참조하세요.

> **Important**
>
> 2026년 6월 30일부터 배치 제공 Amazon Linux 2 AMI를 사용하여 새 Amazon ECS 컴퓨팅 환경 생성을 AWS Batch 차단합니다. AMIs 이 날짜 이전에는 기존 AWS Batch Amazon ECS 컴퓨팅 환경을 Amazon Linux 2023으로 마이그레이션하는 것이 좋습니다. 자세한 내용은 Amazon ECS Amazon Linux 2 AMI 사용 중단 단원을 참조하십시오.

는 Amazon Linux 2에 대한 지원을 종료하므로 2026년 1월 12일에 새 Amazon ECS 컴퓨팅 환경의 기본 AMI를 Amazon Linux 2에서 Amazon Linux 2023 AWS 으로 AWS Batch 변경했습니다. 새 컴퓨팅 환경을 생성할 때 imageType.Ec2Configuration 필드의 값을 지정하지 않으면 기본 AMI가 사용됩니다. 최적의 성능과 보안을 유지하려면 AWS Batch Amazon ECS 컴퓨팅 환경을 Amazon Linux 2023으로 마이그레이션하는 것이 좋습니다.

상태 계획 수명 주기 이벤트를 사용하여 AWS 영향을 받는 Amazon ECS 컴퓨팅 환경의 마이그레이션 상태를 추적할 수 있습니다. 자세한 내용은 AWS 계획된 상태 수명 주기 이벤트 단원을 참조하십시오.

컴퓨팅 환경의 구성 방식에 따라 AL2에서 AL2023으로의 다음 업그레이드 경로 중 하나를 사용할 수 있습니다.

Ec2Configuration.ImageType을 사용한 업그레이드

- 시작 템플릿 또는 시작 템플릿 재정의를 사용하지 않는 경우에는 Ec2Configuration.ImageType을 `ECS_AL2023`(또는 GPU 인스턴스를 사용하는 경우 `ECS_AL2023_NVIDIA`)으로 변경한 다음 UpdateComputeEnvironment를 실행합니다.
- Ec2Configuration.ImageIdOverride를 지정하는 경우, Ec2Configuration.ImageType이 Ec2Configuration.ImageIdOverride에 지정된 AMI 유형과 일치해야 합니다.

  `ImageIdOverride`와 `ImageType`이 일치하지 않는 경우 컴퓨팅 환경이 제대로 작동하지 않을 수 있습니다.

시작 템플릿을 사용하여 업그레이드

- ECS_AL2023을 기반으로 하는 AMI를 지정하는 시작 템플릿을 사용하는 경우 시작 템플릿이 Amazon Linux 2023과 호환되는지 확인합니다. Amazon ECS 최적화 AMI를 위한 Amazon Linux 2023의 변경 사항에 대한 자세한 내용은 Amazon ECS 사용 설명서의 Amazon Linux 2에서 Amazon Linux 2023 Amazon ECS 최적화 AMI로 마이그레이션을 참조하세요.
- AL2023 AMI의 경우 사용자 지정 사용자 데이터 또는 초기화 스크립트가 AL2023 환경 및 패키지 관리 시스템과 호환되는지 확인하세요.

를 사용하여 업그레이드 CloudFormation

- CloudFormation 를 사용하여 컴퓨팅 환경을 관리하는 경우 템플릿을 업데이트하여의 `ImageType` 속성을 `Ec2Configuration`에서 `ECS_AL2`로`ECS_AL2023`(또는 GPU 인스턴스를 사용할 `ECS_AL2023_NVIDIA` 때) 변경합니다.

```
ComputeEnvironment:
  Type: AWS::Batch::ComputeEnvironment
  Properties:
    ComputeResources:
      Ec2Configuration:
        - ImageType: ECS_AL2023
```

  그런 다음 CloudFormation 스택을 업데이트하여 변경 사항을 적용합니다.
- CloudFormation 템플릿이를 사용하여 사용자 지정 AMI를 지정하는 경우 AMI ID가 AL2023-based AMI에 해당하고 `ImageType` 설정과 일치하는지 `ImageId0verride`확인합니다.

마이그레이션 고려 사항

Amazon Linux 2에서 Amazon Linux 2023으로 마이그레이션할 때는 다음 사항을 고려하세요.

- 패키지 관리 - Amazon Linux 2023은 패키지 관리를 위해 `yum` 대신 `dnf`을 사용합니다.
- 시스템 서비스 - 일부 시스템 서비스 및 해당 구성은 AL2와 AL2023 간에 다를 수 있습니다.
- 컨테이너 런타임 - AL2와 AL2023 모두 Docker를 지원하지만 AL2023의 기본 구성은 다를 수 있습니다.
- 보안 - AL2023에는 향상된 보안 기능이 포함되어 있으며 보안 관련 구성을 업데이트해야 할 수 있습니다.

- 인스턴스 메타데이터 서비스 버전 2(IMDSv2) - IMDSv2는 EC2 인스턴스 메타데이터 액세스에 토큰 기반 인증을 요구하여 향상된 보안을 제공하는 세션 지향 서비스입니다. IMDS에 대한 자세한 내용은 Amazon EC22 사용 설명서의 인스턴스 메타데이터 서비스 버전 2 작동 방식을 참조하세요.

변경 사항 및 마이그레이션 고려 사항의 전체 목록은 Amazon ECS 사용 설명서의 Amazon Linux 2에서 Amazon Linux 2023 Amazon ECS 최적화 AMI로 마이그레이션을 참조하세요.

에 대한 인스턴스 유형 할당 전략 AWS Batch

관리형 컴퓨팅 환경이 생성되면는 작업의 요구 사항에 가장 적합한 `instanceTypes` 지정된에서 인스턴스 유형을 AWS Batch 선택합니다. 할당 전략은가 추가 용량을 AWS Batch 필요로 하는 경우의 동작을 정의합니다. 이 파라미터는 Fargate 리소스에서 실행되는 작업에는 적용되지 않습니다. 이 파라미터는 지정하지 마세요.

`BEST_FIT`(기본값)

AWS Batch 는 가장 저렴한 인스턴스 유형을 선호하는 작업의 요구 사항에 가장 적합한 인스턴스 유형을 선택합니다. 선택한 인스턴스 유형의 추가 인스턴스를 사용할 수 없는 경우 추가 인스턴스를 사용할 수 있을 때까지 AWS Batch 기다립니다. 사용 가능한 인스턴스가 충분하지 않거나 사용자가 Amazon EC2 service quotas에 도달한 경우에는 현재 실행 중인 작업이 완료될 때까지 추가 작업은 실행되지 않습니다. 이 할당 전략은 비용은 낮게 유지하지만 확장을 제한할 수 있습니다. `BEST_FIT`과 스팟 집합을 함께 사용하는 경우 스팟 플릿 IAM 역할이 반드시 지정되어야 합니다. 컴퓨팅 환경을 업데이트할 때 `BEST_FIT`는 지원되지 않습니다. 자세한 내용은 에서 컴퓨팅 환경 업데이트 AWS Batch 단원을 참조하십시오.

> **Note**
>
> AWS Batch 는 계정의 AWS 리소스를 관리합니다. BEST_FIT 할당 전략을 사용하는 컴퓨팅 환경은 원래 기본적으로 시작 구성을 사용했습니다. 그러나 새 AWS 계정에서 시작 구성을 사용하는 것은 시간이 지남에 따라 제한됩니다. 따라서 2024년 4월 말부터 새로 생성된 BEST_FIT 컴퓨팅 환경은 기본적으로 시작 템플릿으로 설정됩니다. 서비스 역할에 시작 템플릿을 관리할 권한이 없는 경우 시작 구성을 계속 활용할 AWS Batch 수 있습니다. 기존 컴퓨팅 환경은 시작 구성을 계속 사용합니다.

`BEST_FIT_PROGRESSIVE`

AWS Batch 는 대기열에 있는 작업의 요구 사항을 충족할 만큼 충분히 큰 추가 인스턴스 유형을 선택합니다. 단위 vCPU 비용이 저렴한 인스턴스 유형이 선호됩니다. 이전에 선택한 인스턴스 유형의 추가 인스턴스를 사용할 수 없는 경우 AWS Batch 는 새 인스턴스 유형을 선택합니다.

> **Note**
>
> 다중 노드 병렬 작업의 경우 AWS Batch 는 사용 가능한 최적의 인스턴스 유형을 선택합니다. 용량 부족으로 인해 인스턴스 유형을 사용할 수 없게 되는 경우, 패밀리 내의 다른 인스턴스 유형이 시작되지 않습니다.

`SPOT_CAPACITY_OPTIMIZED`

AWS Batch 는 대기열에 있는 작업의 요구 사항을 충족할 만큼 충분히 큰 인스턴스 유형을 하나 이상 선택합니다. 중단될 가능성이 적은 인스턴스 유형이 선호됩니다. 이 할당 전략은 스팟 인스턴스 컴퓨팅 리소스에만 사용할 수 있습니다.

`SPOT_PRICE_CAPACITY_OPTIMIZED`

가격 및 용량 최적화 할당 전략은 가격과 용량을 모두 고려하여 중단될 가능성이 가장 낮으면서 가장 저렴한 스팟 인스턴스 풀을 선택합니다. 이 할당 전략은 스팟 인스턴스 컴퓨팅 리소스에만 사용할 수 있습니다.

> **Note**
>
> 대부분의 인스턴스에서 `SPOT_CAPACITY_OPTIMIZED`보다 `SPOT_PRICE_CAPACITY_OPTIMIZED`을 사용하는 것을 권장합니다.

`BEST_FIT_PROGRESSIVE` 및 `BEST_FIT` 전략은 온디맨드 또는 스팟 인스턴스를 사용하고, `SPOT_CAPACITY_OPTIMIZED` 및 `SPOT_PRICE_CAPACITY_OPTIMIZED` 전략은 스팟 인스턴스를 사용합니다. 그러나 용량 요구 사항을 충족하려면를 초과 AWS Batch 해야 `maxvCpus` 할 수 있습니다. 이 경우가 단일 인스턴스를 초과`maxvCpus`해서는 AWS Batch 안 됩니다.

컴퓨팅 리소스 메모리 관리

Amazon ECS 컨테이너 에이전트가 컨테이너 컴퓨팅 리소스를 컴퓨팅 환경에 등록할 때 에이전트는 컨테이너 컴퓨팅 리소스가 작업용으로 예약할 수 있는 메모리 양을 결정해야 합니다. 플랫폼 메모리 오

버헤드와 운영 체제 커널이 차지하는 메모리로 인해 이 수치가 Amazon EC2 인스턴스에 대해 설치된 메모리 용량과 다르기 때문입니다 예를 들어, `m4.large` 인스턴스의 설치된 메모리는 8GiB입니다. 하지만 이는 컴퓨팅 리소스 등록 시 작업에 대해 정확히 8,192MiB의 메모리를 사용할 수 있음을 의미하지 않습니다.

작업에 8,192MiB를 지정하는 것을 가정하고, 컴퓨팅 리소스 중 사용할 수 있는 메모리가 8,192MiB 이상인 컴퓨팅 리소스가 없다고 가정해 보겠습니다. 그러면 해당 작업을 컴퓨팅 환경에 배치할 수 없습니다. 관리형 컴퓨팅 환경을 사용 중인 경우 AWS Batch은(는) 요청을 수용하기 위해 더 큰 인스턴스 유형을 시작해야 합니다.

기본 AWS Batch 컴퓨팅 리소스 AMI는 또한 Amazon ECS 컨테이너 에이전트 및 기타 중요 시스템 프로세스에 대해 32MiB의 메모리를 예약합니다. 이 메모리는 작업 할당에 사용할 수 없습니다. 자세한 내용은 시스템 메모리 예약 섹션을 참조하세요.

Amazon ECS 컨테이너 에이전트는 Docker `ReadMemInfo()` 함수를 사용하여 운영 체제가 사용할 수 있는 전체 메모리를 쿼리합니다. Linux와 Windows 모두 명령줄 유틸리티를 제공하여 총 메모리를 결정합니다.

Example- Linux 총 메모리 결정

free 명령은 운영 체제가 인식한 총 메모리를 반환합니다.

```
$ free -b
```

다음은 Amazon ECS 최적화된 Amazon Linux AMI를 실행하는 `m4.large` 인스턴스의 출력의 예입니다.

```
             total       used       free     shared    buffers     cached
Mem:    8373026816  348180480 8024846336      90112   25534464  205418496
-/+ buffers/cache:  117227520 8255799296
```

이 인스턴스의 총 메모리는 8,373,026,816바이트입니다. 즉, 태스크에 7,985MiB을 사용할 수 있습니다.

주제

- 시스템 메모리 예약
- 자습서: 컴퓨팅 리소스 메모리 보기
- Amazon EKS의 AWS Batch에 대한 메모리 및 vCPU 고려 사항

시스템 메모리 예약

컴퓨팅 리소스의 모든 메모리를 작업에 사용하는 경우 작업 및 메모리의 중요 시스템 프로세스 경합으로 인해 시스템 오류가 발생할 수 있습니다. Amazon ECS 컨테이너 에이전트는 `ECS_RESERVED_MEMORY`라는 구성 변수를 제공합니다. 이 구성 변수를 사용하여 작업에 할당된 풀의 지정된 메모리 용량(MiB)을 제거할 수 있습니다. 이를 통해 중요 시스템 프로세스에 대한 메모리를 효율적으로 예약합니다.

기본 AWS Batch 컴퓨팅 리소스 AMI는 Amazon ECS 컨테이너 에이전트 및 기타 중요 시스템 프로세스에 대해 32MiB의 메모리를 예약합니다. Amazon ECS 컨테이너 에이전트 및 기타 중요한 시스템 프로세스에 대해 5% 메모리 버퍼를 예약하는 것이 좋습니다.

자습서: 컴퓨팅 리소스 메모리 보기

컨테이너 컴퓨팅 리소스가 Amazon ECS 콘솔 또는 DescribeContainerInstances API 작업에 등록하는 메모리 양을 확인할 수 있습니다. 특정 인스턴스 유형에 대해 가능한 많은 메모리를 제공하여 리소스 사용률을 최대화하고자 하는 경우 컴퓨팅 리소스에 대해 사용 가능한 메모리를 관찰한 다음 작업에 메모리를 할당할 수 있습니다.

컴퓨팅 리소스 메모리를 보려면

1. https://console.aws.amazon.com/ecs/v2에서 콘솔을 엽니다.
2. 클러스터를 선택한 다음 보려는 컴퓨팅 리소스를 호스팅하는 클러스터를 선택합니다.

   컴퓨팅 환경에 대한 클러스터 이름은 컴퓨팅 환경 이름으로 시작됩니다.
3. 인프라를 선택합니다.
4. 컨테이너 인스턴스에서 연결할 컨테이너 인스턴스를 선택합니다.
5. 리소스 및 네트워킹(Resources and networking) 섹션에서는 컴퓨팅 리소스에 대해 등록된 메모리와 사용 가능한 메모리를 보여줍니다.

   총 용량 메모리 값은 처음 시작할 때 Amazon ECS에 등록된 컴퓨팅 리소스의 값이며, 사용 가능 메모리 값은 작업에 아직 할당되지 않은 값입니다.

Amazon EKS의 AWS Batch에 대한 메모리 및 vCPU 고려 사항

Amazon EKS의 AWS Batch에서는 컨테이너에 사용할 수 있는 리소스를 지정할 수 있습니다. 예를 들어 vCPU 및 메모리 리소스에 대해 `requests` 또는 `limits` 값을 지정할 수 있습니다.

vCPU 리소스를 지정하기 위한 제약 조건은 다음과 같습니다.

- vCPU `requests` 또는 `limits` 중 하나는 지정해야 합니다.
- vCPU 유닛 1개는 물리적 코어 또는 가상 코어 1개와 동일합니다.
- vCPU 값은 정수로 입력하거나 0.25씩 증분하여 입력해야 합니다.
- 유효한 vCPU 값 중 가장 작은 값은 0.25입니다.
- 두 값이 지정되면, `requests` 값은 `limits` 값에 대해 지정된 값과 같거나 이보다 작아야 합니다. 이렇게 하면 소프트 vCPU 구성과 하드 vCPU 구성을 모두 구성할 수 있습니다.
- vCPU 값은 milliCPU 형식으로 지정할 수 없습니다. 예를 들어, `100m`(은)는 유효한 형식이 아닙니다.
- AWS Batch(은)는 `requests` 값을 조정 결정에 사용합니다. `requests` 값이 지정되지 않은 경우 `limits` 값이 `requests` 값에 복사됩니다.

다음은 메모리 리소스를 지정하기 위한 제약 조건입니다.

- 메모리 `requests` 또는 `limits` 중 하나는 지정해야 합니다.
- 메모리 값은 mebibytes(MiBs)이어야 합니다.
- 둘 다 지정된 경우 `requests` 값은 `limits` 값과 같아야 합니다.
- AWS Batch(은)는 `requests` 값을 조정 결정에 사용합니다. `requests` 값이 지정되지 않은 경우 `limits` 값이 `requests` 값에 복사됩니다.

다음은 GPU 리소스를 지정하기 위한 제약 조건입니다.

- 둘 다 지정된 경우 `requests` 값은 `limits` 값과 같아야 합니다.
- AWS Batch(은)는 `requests` 값을 조정 결정에 사용합니다. `requests` 값이 지정되지 않은 경우 `limits` 값이 `requests` 값에 복사됩니다.

예제: 작업 정의

Amazon EKS 작업 정의의 AWS Batch(은)는 소프트 vCPU 공유를 구성합니다. 이렇게 하면 Amazon EKS의 AWS Batch(이)가 해당 인스턴스 유형의 vCPU 용량을 모두 사용할 수 있습니다. 하지만 실행 중인 다른 작업이 있는 경우 작업에는 최대 2개의 vCPU가 할당됩니다. 메모리는 2GB로 제한됩니다.

```
{
   "jobDefinitionName": "MyJobOnEks_Sleep",
   "type": "container",
```

```
    "eksProperties": {
        "podProperties": {
            "containers": [
                {
                    "image": "public.ecr.aws/amazonlinux/amazonlinux:2",
                    "command": ["sleep", "60"],
                    "resources": {
                        "requests": {
                            "cpu": "2",
                            "memory": "2048Mi"
                        }
                    }
                }
            ]
        }
    }
}
```

다음 Amazon EKS의 AWS Batch 작업 정의는 `request` 값이 1이고 최대 4개의 vCPU를 작업에 할당합니다.

```
{
    "jobDefinitionName": "MyJobOnEks_Sleep",
    "type": "container",
    "eksProperties": {
        "podProperties": {
            "containers": [
                {
                    "image": "public.ecr.aws/amazonlinux/amazonlinux:2",
                    "command": ["sleep", "60"],
                    "resources": {
                        "requests": {
                            "cpu": "1"
                        },
                        "limits": {
                            "cpu": "4",
                            "memory": "2048Mi"
                        }
                    }
                }
            ]
        }
    }
```

```
}
```

다음 Amazon EKS의 AWS Batch 작업 정의는 vCPU `limits` 값을 1로, 메모리 `limits` 값을 1GB로 설정합니다.

```
{
   "jobDefinitionName": "MyJobOnEks_Sleep",
   "type": "container",
   "eksProperties": {
       "podProperties": {
           "containers": [
               {
                   "image": "public.ecr.aws/amazonlinux/amazonlinux:2",
                   "command": ["sleep", "60"],
                   "resources": {
                       "limits": {
                           "cpu": "1",
                           "memory": "1024Mi"
                       }
                   }
               }
           ]
       }
   }
}
```

AWS Batch(이)가 Amazon EKS의 AWS Batch 작업을 Amazon EKS 포드로 변환하면 AWS Batch가 `limits` 값을 `requests` 값에 복사합니다. 이는 `requests` 값이 지정되지 않은 경우입니다. 위의 예제 작업 정의를 제출하면 포드 `spec`(은)는 다음과 같습니다.

```
apiVersion: v1
kind: Pod
...
spec:
  ...
  containers:
    - command:
        - sleep
        - 60
      image: public.ecr.aws/amazonlinux/amazonlinux:2
      resources:
        limits:
```

```
          cpu: 1
          memory: 1024Mi
        requests:
          cpu: 1
          memory: 1024Mi
      ...
```

노드 CPU 및 메모리 예약

AWS Batch(은)는 vCPU 및 메모리 예약에 `bootstrap.sh` 파일의 기본 로직을 사용합니다. `bootstrap.sh` 파일에 대한 자세한 내용은 bootstrap.sh을 참조하세요. vCPU와 메모리 리소스의 크기를 조정할 때는 다음 예제를 고려하세요.

> **Note**
>
> 실행 중인 인스턴스가 없는 경우 vCPU 및 메모리 예약은 초기에 AWS Batch 조정 로직과 의사 결정에 영향을 미칠 수 있습니다. 인스턴스가 실행된 후 AWS Batch(은)는 초기 할당량을 조정합니다.

예제: 노드 CPU 예약

CPU 예약 값은 인스턴스에 사용할 수 있는 총 vCPU 수를 사용하여 밀리코어 단위로 계산됩니다.

vCPU 번호	예약 비율(%)
1	6%
2	1%
3~4	0.5%
4 이상	0.25%

위 값을 사용하면 다음과 같습니다.

- vCPU가 2개 있는 `c5.large` 인스턴스의 CPU 예약 값은 70m입니다. 이 값은 다음과 같은 방식으로 계산됩니다. (1*60)+(1*10)=70m.

- vCPU가 96개 있는 c5.24xlarge 인스턴스의 CPU 예약 값은 310m입니다. 이는 다음과 같은 방식으로 계산됩니다. (1*60)+(1*10)+(2*5)+(92*2.5)=310m.

이 예제에서는 c5.large 인스턴스에서 작업을 실행하는 데 사용할 수 있는 밀리코어 vCPU 유닛이 1,930개(2,000~70개로 계산)입니다. 작업에 2 (2*1000 m) vCPU 유닛이 필요한데, 작업이 단일 c5.large 인스턴스에 맞지 않다고 가정하겠습니다. 하지만 1.75 vCPU 유닛이 필요한 작업에는 적합합니다.

예제: 노드 메모리 예약

메모리 예약 값은 다음을 사용하여 메비바이트 단위로 계산됩니다.

- 인스턴스 용량(MiB) 예를 들어 8GB 인스턴스는 7,748MiB입니다.
- kubeReserved 값 kubeReserved 값은 시스템 대몬(daemon)용으로 예약할 메모리 양입니다. kubeReserved 값은 다음과 같은 방식으로 계산됩니다. ((11*인스턴스 유형에서 지원하는 최대 포드 수)+255). 각 인스턴스 유형별로 지원되는 최대 포드 수 목록은 GitHub에서 eni-max-pods.txt를 참조하세요.
- HardEvictionLimit 값 사용 가능한 메모리가 HardEvictionLimit 값 아래로 떨어지면 인스턴스는 포드를 제거하려고 시도합니다.

할당 가능한 메모리를 계산하는 공식은 다음과 같습니다. (*instance_capacity_in_MiB*)-(11*(*maximum_number_of_pods*))-255-(*HardEvictionLimit value.*)).

c5.large 인스턴스는 최대 29개의 포드를 지원합니다. c5.large 값이 100MiB인 8GB HardEvictionLimit 인스턴스의 경우 할당 가능한 메모리는 7,074MiB입니다. 이 값은 다음과 같은 방식으로 계산됩니다. (7748-(11* 29)-255-100)=7074MiB. 이 예제에서 8,192MiB 작업은 8gibibyte(GiB) 인스턴스임에도 불구하고 이 인스턴스에 맞지 않습니다.

DaemonSets

DaemonSets(을)를 사용할 때 다음 사항을 고려하세요.

- 실행 중인 Amazon EKS의 AWS Batch 인스턴스가 없는 경우, DaemonSets가 초기에 AWS Batch 조정 로직 및 의사 결정에 영향을 미칠 수 있습니다. AWS Batch는 초기에 예상된 DaemonSets에 0.5 vCPU 유닛과 500MiB를 할당합니다. 인스턴스가 실행된 후 AWS Batch(은)는 초기 할당량을 조정합니다.

- DaemonSet(은)는 vCPU 또는 메모리 제한을 정의하면 Amazon EKS의 AWS Batch 작업은 더 적은 리소스를 가집니다. AWS Batch 작업에 할당되는 DaemonSets의 수를 가능한 한 적게 유지하는 것이 좋습니다.

Fargate 컴퓨팅 환경

Fargate는 Amazon EC2 인스턴스의 서버 또는 클러스터를 관리할 필요 없이 컨테이너를 실행 AWS Batch 하기 위해와 함께 사용할 수 있는 기술입니다. Fargate를 사용하면 더 이상 컨테이너를 실행하기 위해 가상 머신의 클러스터를 프로비저닝, 구성 또는 조정할 필요가 없습니다. 따라서 서버 유형을 선택하거나, 클러스터를 조정할 시점을 결정하거나, 클러스터 패킹을 최적화할 필요가 없습니다.

Fargate 리소스를 사용하여 작업을 실행할 때는 애플리케이션 패키지를 컨테이너에 작성하고, CPU 및 메모리 요구 사항을 지정한 다음, 네트워킹 및 IAM 정책을 정의하고, 애플리케이션을 시작합니다. 각 Fargate 작업에는 자체 격리 경계가 있으며 다른 작업과 기본 커널, CPU 리소스, 메모리 리소스 또는 탄력적 네트워크 인터페이스를 공유하지 않습니다.

Fargate는 Amazon ECS를 오케스트레이터로 사용하는 AWS Batch 컴퓨팅 환경에서만 사용할 수 있습니다. Fargate는 Amazon EKS 컴퓨팅 환경의 AWS Batch 에 대해 지원되지 않습니다. 자세한 내용은 Amazon EKS 컴퓨팅 환경 단원을 참조하십시오.

주제

- Fargate를 사용하는 경우
- Fargate의 작업 정의
- Fargate의 작업 대기열
- Fargate의 컴퓨팅 환경

Fargate를 사용하는 경우

대부분의 시나리오에서 Fargate를 사용하는 것이 좋습니다. Fargate는 컨테이너에 지정하는 리소스 요구 사항에 근접하게 일치하도록 컴퓨팅을 시작하고 규모를 조정합니다. Fargate를 사용하면 과도하게 프로비저닝하거나 추가 서버를 위해 비용을 지불할 필요가 없습니다. 또한 인스턴스 유형과 같은 인프라 관련 파라미터의 세부 사항에 대해서도 걱정할 필요가 없습니다. 컴퓨팅 환경을 스케일 업해야 하는 경우 Fargate 리소스에서 실행되는 작업을 더 빠르게 시작할 수 있습니다. 일반적으로 새 Amazon EC2 인스턴스를 가동하는 데 몇 분 정도 걸립니다. 하지만 Fargate에서 실행되는 작업은 약 30초 내에 프로비저닝할 수 있습니다. 정확한 소요 시간은 컨테이너 이미지 크기, 작업 수 등 여러 요인에 따라 달라집니다.

그러나 작업에 다음이 필요한 경우에는 Amazon EC2를 사용하는 것이 좋습니다.

- 16개 이상의 vCPU
- 120기가바이트(GiB) 이상의 메모리
- GPU
- 사용자 지정 Amazon Machine Image(AMI)
- 모든 linuxParameters 파라미터

작업 수가 많은 경우 Amazon EC2 인프라를 사용하는 것이 좋습니다. 동시 실행 작업 수가 Fargate의 조절 제한을 초과하는 경우를 예로 들 수 있습니다. 이는 EC2를 사용하면 Fargate 리소스보다 EC2 리소스로 작업을 더 빠르게 전송할 수 있기 때문입니다. 또한 EC2를 사용하면 더 많은 작업을 동시에 실행할 수 있습니다. 자세한 내용은 Amazon Elastic Container Service 개발자 안내서의 Fargate Service Quotas를 참조하세요.

Fargate의 작업 정의

AWS Batch 의 작업은 사용 가능한 모든 작업 정의 파라미터를 지원하지 AWS Fargate 않습니다. 전혀 지원되지 않는 파라미터도 있고 Fargate 작업에 다르게 작동하는 파라미터도 있습니다.

다음 목록은 Fargate 작업에서 유효하지 않거나 제한되는 작업 정의 파라미터를 설명합니다.

`platformCapabilities`

`FARGATE`로 지정되어야 합니다.

```
"platformCapabilities": [ "FARGATE" ]
```

`type`

`container`로 지정되어야 합니다.

```
"type": "container"
```

`containerProperties`의 파라미터

`executionRoleArn`

Fargate 리소스에서 실행되는 작업에 지정해야 합니다. 자세한 내용은 Amazon Elastic Container Service 개발자 안내서의 태스크에 대한 IAM 역할을 참조하세요.

```
"executionRoleArn": "arn:aws:iam::123456789012:role/ecsTaskExecutionRole"
```

`fargatePlatformConfiguration`

(선택 사항, Fargate 작업 정의에만 해당). Fargate 플랫폼 버전을 지정하거나 최신 플랫폼 버전에 대한 `LATEST`를 지정합니다. `platformVersion`에 가능한 값은 `1.3.0`, `1.4.0` 및 `LATEST`입니다(기본값).

```
"fargatePlatformConfiguration": { "platformVersion": "1.4.0" }
```

`instanceType`, `ulimits`

Fargate 리소스에서 실행되는 작업에는 적용되지 않습니다.

`memory`, `vcpus`

이러한 설정은 `resourceRequirements`에서 지정해야 합니다.

`privileged`

이 파라미터를 지정하지 않거나 `false`를 지정합니다.

```
"privileged": false
```

`resourceRequirements`

메모리 및 vCPU 요구 사항 모두 지원되는 값을 사용하여 지정해야 합니다. Fargate 리소스에서 실행되는 작업에는 GPU 리소스가 지원되지 않습니다.

GuardDuty Runtime Monitoring을 사용하는 경우 GuardDuty 보안 에이전트에 약간의 메모리 오버헤드가 있습니다. 따라서 메모리 제한에 GuardDuty 보안 에이전트의 크기가 포함되어야 합니다. GuardDuty 보안 에이전트 메모리 제한에 대한 자세한 내용은 GuardDuty 사용 설명서의 CPU and memory limits를 참조하세요. 모범 사례에 대한 자세한 내용은 Amazon ECS 개발자 안내서의 런타임 모니터링을 활성화한 후 Fargate 작업에서 메모리 오류를 해결하는 방법을 참조하세요.

```
"resourceRequirements": [
  {"type": "MEMORY", "value": "512"},
  {"type": "VCPU",   "value": "0.25"}
]
```

`linuxParameters`의 파라미터

`devices`, `maxSwap`, `sharedMemorySize`, `swappiness`, `tmpfs`

Fargate 리소스에서 실행되는 작업에는 적용되지 않습니다.

`logConfiguration`의 파라미터

`logDriver`

`awslogs` 및 `splunk`만 지원됩니다. 자세한 내용은 awslogs 로그 드라이버 사용 단원을 참조하십시오.

`networkConfiguration`의 멤버

`assignPublicIp`

프라이빗 서브넷에 인터넷으로 트래픽을 전송하기 위한 NAT 게이트웨이가 연결되어 있지 않은 경우 `assignPublicIp`는 "`ENABLED`"여야 합니다. 자세한 내용은 AWS Batch IAM 실행 역할 단원을 참조하십시오.

Fargate의 작업 대기열

AWS Batch 의 작업 대기열 AWS Fargate 은 기본적으로 변경되지 않습니다. 유일한 제한 사항은 `computeEnvironmentOrder`에 나열된 컴퓨팅 환경이 모두 Fargate 컴퓨팅 환경(`FARGATE` 또는 `FARGATE_SPOT`)이어야 한다는 것입니다. EC2와 Fargate 컴퓨팅 환경은 함께 사용할 수 없습니다.

Fargate의 컴퓨팅 환경

AWS Batch 의 컴퓨팅 환경은 사용 가능한 모든 컴퓨팅 환경 파라미터를 지원하지 AWS Fargate 않습니다. 전혀 지원되지 않는 파라미터도 있습니다. 다른 파라미터는 Fargate에 대한 특정 요구 사항이 있습니다.

다음 목록은 Fargate 작업에서 유효하지 않거나 제한되는 컴퓨팅 환경 파라미터를 설명합니다.

`type`

이 파라미터는 `MANAGED`여야 합니다.

```
"type": "MANAGED"
```

`computeResources` 객체의 파라미터

`allocationStrategy`, `bidPercentage`, `desiredvCpus`, `imageId`, `instanceTypes`, `ec2Configuration`, `ec2KeyPair`, `instanceRole`, `launchTemplate`, `minvCpus`, `placementGroup`, `spotIamFleetRole`

이는 Fargate 컴퓨팅 환경에는 적용할 수 없으며 제공될 수 없습니다.

`subnets`

이 파라미터에 나열된 서브넷에 NAT 게이트웨이가 연결되어 있지 않은 경우 작업 정의의 `assignPublicIp` 파라미터를 `ENABLED`로 설정해야 합니다.

`tags`

이는 Fargate 컴퓨팅 환경에는 적용할 수 없으며 제공될 수 없습니다. Fargate 컴퓨팅 환경에 태그를 지정하려면 `computeResources` 객체에 없는 `tags` 파라미터를 사용합니다.

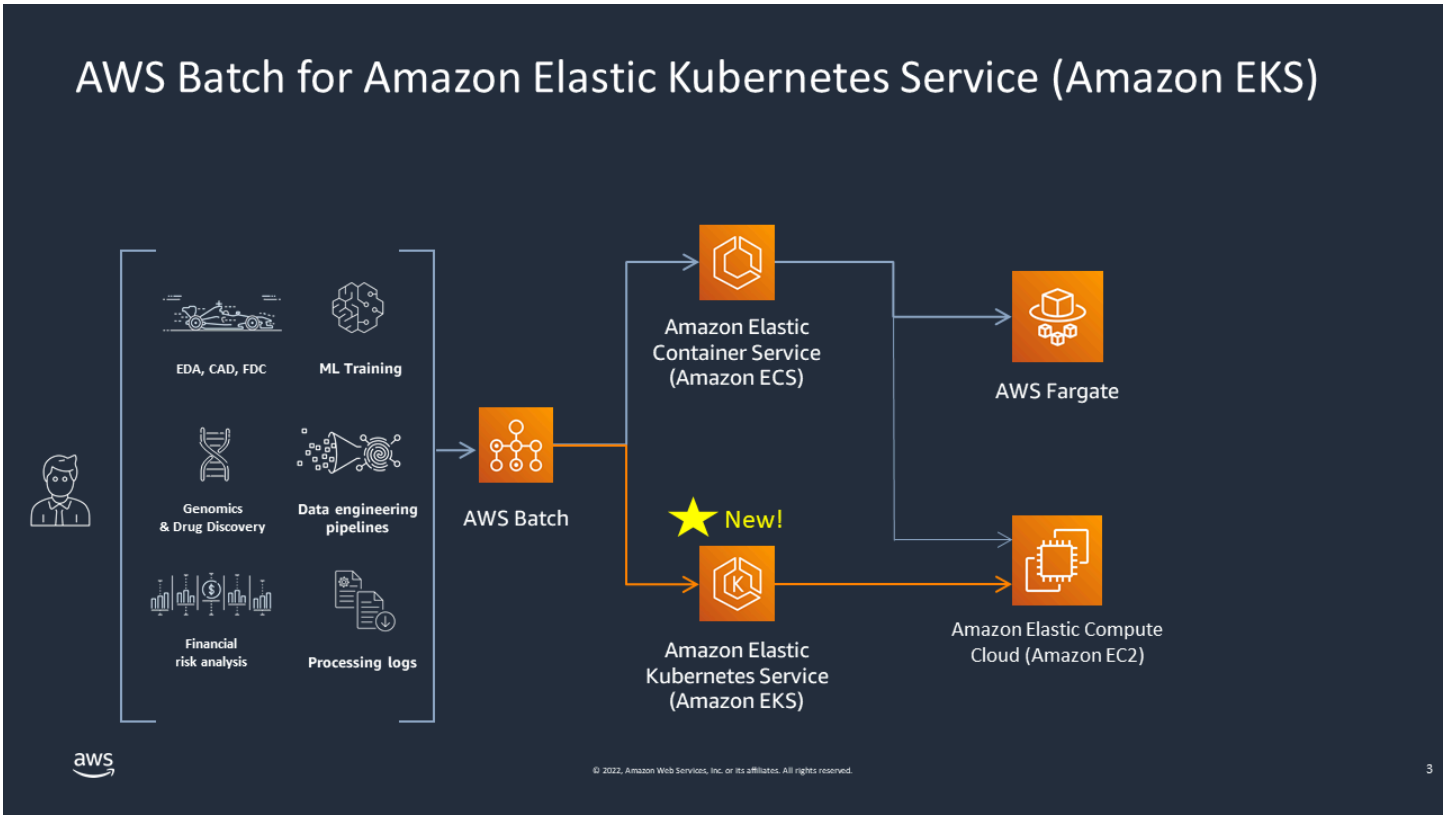
`type`

`FARGATE` 또는 `FARGATE_SPOT`여야 합니다.

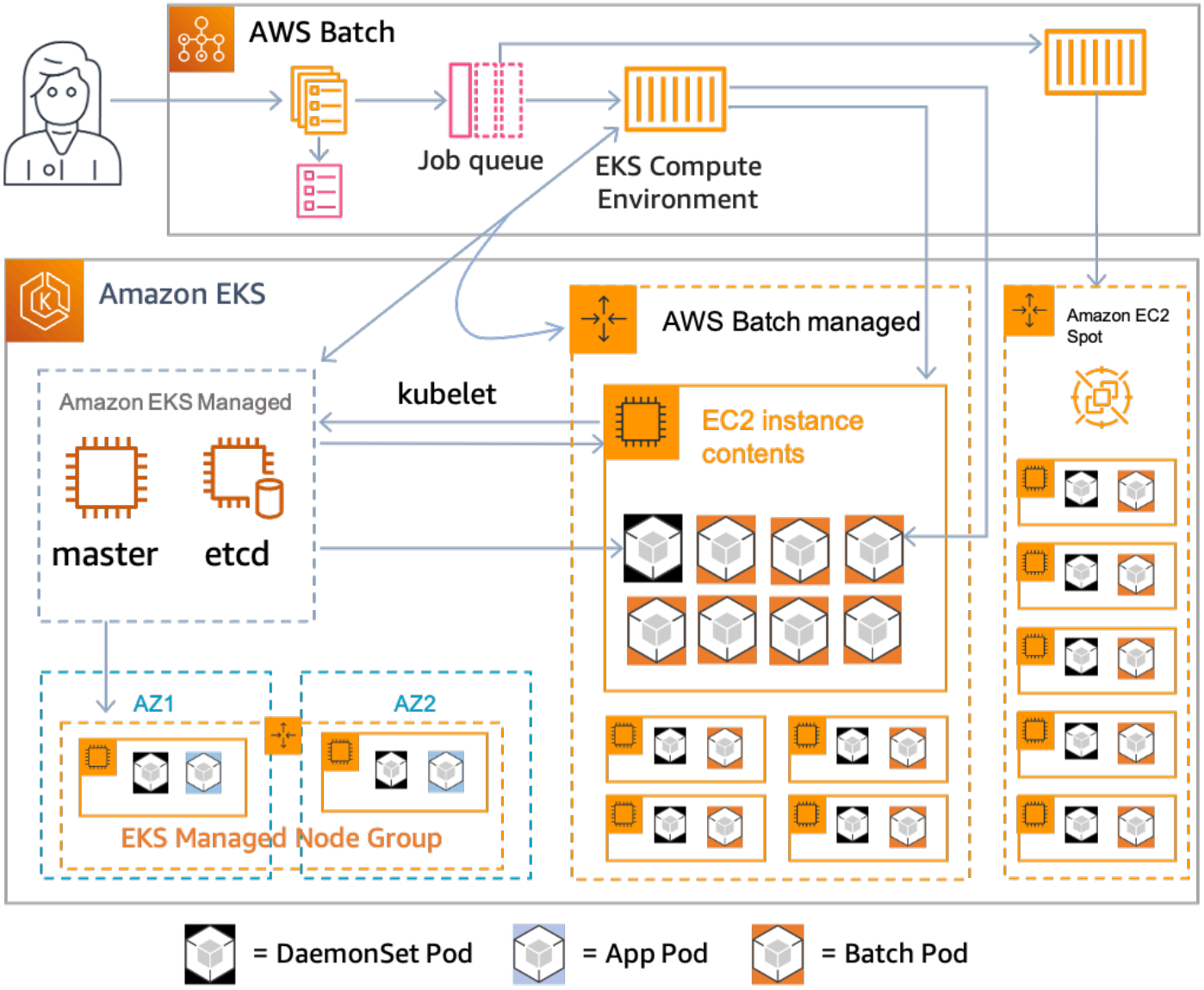
```
"type": "FARGATE_SPOT"
```

Amazon EKS 컴퓨팅 환경

Amazon EKS AWS Batch 에서 시작하기는 EKS 컴퓨팅 환경 생성에 대한 간략한 안내서를 제공합니다. 이 섹션에서는 Amazon EKS 컴퓨팅 환경에 대한 자세한 내용을 제공합니다.



AWS Batch 는 관리형 배치 기능을 제공하여 Amazon EKS 클러스터의 배치 워크로드를 간소화합니다. 여기에는 대기열, 종속성 추적, 관리형 작업 재시도 및 우선 순위, 포드 관리, 노드 조정이 포함됩니다.는 여러 가용 영역과 여러 Amazon EC2 인스턴스 유형 및 크기를 처리할 AWS Batch 수 있습니다. 는 여러 Amazon EC2 스팟 모범 사례를 AWS Batch 통합하여 내결함성 방식으로 워크로드를 실행하므로 중단을 줄일 수 있습니다. AWS Batch (을)를 사용하면 안심하고 소수의 야간 작업이나 수백만 개의 미션 크리티컬한 작업을 실행할 수 있습니다.



AWS Batch 는 Amazon Elastic Kubernetes Service(Amazon EKS)에서 관리하는 Kubernetes 클러스터의 배치 워크로드를 오케스트레이션하는 관리형 서비스입니다.는 "오버레이" 모델을 사용하여 클러스터 외부에서이 오케스트레이션을 AWS Batch 수행합니다. AWS Batch 는 관리형 서비스이므로 클러스터에 설치하거나 관리할 Kubernetes 구성 요소(예: 운영자 또는 사용자 지정 리소스)가 없습니다.는가 Kubernetes API 서버와 AWS Batch 통신할 수 있도록 하는 역할 기반 액세스 제어(RBAC)로 클러스터를 구성하기 AWS Batch 만 하면 됩니다. Kubernetes는 포드 및 노드를 생성, 모니터링 및 삭제하기 위해 Kubernetes APIs를 AWS Batch 호출합니다.

AWS Batch 에는 작업 용량 할당 측면에서 최적화를 통해 작업 대기열 로드를 기반으로 Kubernetes 노드를 확장하는 조정 로직이 내장되어 있습니다. 작업 대기열이 비어 있으면 AWS Batch 는 기본적으로 0인 설정한 최소 용량으로 노드를 축소합니다.는 이러한 노드의 전체 수명 주기를 AWS Batch 관리하고 레이블과 테인트로 노드를 장식합니다. 이렇게 하면에서 관리하는 노드에 다른 Kubernetes 워크로

드가 배치되지 않습니다 AWS Batch. 이에 대한 예외는 노드를 대상으로 AWS Batch 하여 작업의 적절한 실행에 필요한 모니터링 및 기타 기능을 제공할 수 DaemonSets있는 입니다. 또한 AWS Batch 는 관리하지 않는 클러스터의 노드에서 작업, 특히 포드를 실행하지 않습니다. 이렇게 하면 클러스터의 다른 애플리케이션에 대해 별도의 확장 로직과 서비스를 사용할 수 있습니다.

작업을 제출하려면 AWS Batch API. AWS Batch translates 작업과 직접 AWS Batch상호 작용podspecs한 다음 Amazon EKS 클러스터의 AWS Batch 에서 관리하는 노드에 포드를 배치하라는 요청을 생성합니다. kubectl(와)과 같은 도구를 사용하여 실행 중인 포드 및 노드를 볼 수 있습니다. 포드 실행이 완료되면 Kubernetes 시스템에서 더 낮은 부하를 유지하기 위해 생성한 포드를 AWS Batch 삭제합니다.

유효한 Amazon EKS 클러스터를에 연결하여 시작할 수 있습니다 AWS Batch. 그런 다음 AWS Batch 작업 대기열을 연결하고 podspec 동등한 속성을 사용하여 Amazon EKS 작업 정의를 등록합니다. 마지막으로, 작업 정의를 참조하는 SubmitJob API 작업을 사용하여 작업을 제출합니다. 자세한 내용은 Amazon EKS AWS Batch 에서 시작하기 단원을 참조하십시오.

AWS Batch Amazon EKS의는 Amazon EC2 인스턴스(온디맨드 및 스팟)를 컴퓨팅 리소스로 지원합니다. 에서 Fargate를 사용하려면 대신 Amazon ECS 컴퓨팅 환경을 AWS Batch사용합니다. 자세한 내용은 Fargate 컴퓨팅 환경 단원을 참조하십시오.

Amazon EKS

주제

- Amazon EKS 기본 AMI
- 혼합 AMI 환경
- 지원되는 Kubernetes 버전
- 컴퓨팅 환경의 Kubernetes 버전 업데이트
- Kubernetes 노드에 대한 공동 책임
- AWS Batch 관리형 노드DaemonSet에서 실행
- Amazon EKS 시작 템플릿 사용자 지정
- EKS AL2에서 EKS AL2023으로 업그레이드하는 방법

Amazon EKS 기본 AMI

Amazon EKS 컴퓨팅 환경을 생성할 때 Amazon Machine Image(AMI)를 지정할 필요가 없습니다.는 CreateComputeEnvironment 요청에 지정된 Kubernetes 버전 및 인스턴스 유형을 기반으로 Amazon

EKS 최적화 AMI를 AWS Batch 선택합니다. 일반적으로 기본 AMI를 선택하는 것이 좋습니다. AMI 선택 우선 순위에 대한 자세한 내용은 섹션을 참조하세요AMI 선택 순서. Amazon EKS에 최적화된 AMI에 대한 자세한 내용은 Amazon EKS 사용 설명서의 Amazon EKS에 최적화된 Amazon Linux AMI를 참조하세요.

> **Important**
>
> Amazon Linux 2023 AMIs는 AWS Batch Amazon EKS의 기본 입니다.
> AWS 는 11/26/25부터 Amazon EKS AL2-optimized 및 AL2-accelerated AMIs에 대한 지원을 종료합니다. 11/26/25 end-of-support 이후에도 Amazon EKS 컴퓨팅 환경에서 AWS Batch제공 Amazon EKS 최적화 Amazon Linux 2 AMIs를 계속 사용할 수 있지만 이러한 컴퓨팅 환경은 더 이상 새로운 소프트웨어 업데이트, 보안 패치 또는 버그 수정을 받지 않습니다 AWS. AL2에서 AL2023으로 업그레이드하는 방법에 대한 자세한 내용은 AWS Batch 사용 설명서의 EKS AL2에서 EKS AL2023으로 업그레이드하는 방법 섹션을 참조하세요.

다음 명령을 실행하여 Amazon EKS 컴퓨팅 환경에 대해 AWS Batch 선택한 AMI 유형을 확인합니다. 다음 예제는 GPU가 아닌 인스턴스 유형입니다.

```
# compute CE example: indicates Batch has chosen the AL2023 x86 or ARM EKS 1.35 AMI,
 depending on instance types
    $ aws batch describe-compute-environments --compute-environments My-Eks-CE1 \
        | jq '.computeEnvironments[].computeResources.ec2Configuration'
    [
      {
        "imageType": "EKS_AL2023",
        "imageKubernetesVersion": "1.35"
      }
    ]
```

다음 예제는 GPU 인스턴스 유형입니다.

```
# GPU CE example: indicates Batch has chosen the AL2023 x86 EKS Accelerated 1.35 AMI
    $ aws batch describe-compute-environments --compute-environments My-Eks-GPU-CE \
        | jq '.computeEnvironments[].computeResources.ec2Configuration'
    [
      {
        "imageType": "EKS_AL2023_NVIDIA",
        "imageKubernetesVersion": "1.35"
      }
```

```
    ]
```

혼합 AMI 환경

시작 템플릿 재정의를 사용하여 Amazon Linux 2(AL2) 및 Amazon Linux 2023(AL2023) AMI를 모두 포함하는 컴퓨팅 환경을 생성할 수 있습니다. 이는 서로 다른 아키텍처에 서로 다른 AMI를 사용하거나 AL2에서 AL2023으로 전환하는 마이그레이션 기간 동안 사용하는 데 유용합니다.

> **Note**
>
> AWS 는 11/26/25부터 Amazon EKS AL2-optimized 및 AL2-accelerated AMIs에 대한 지원을 종료합니다. 11/26/25 end-of-support 이후에도 Amazon EKS 컴퓨팅 환경에서 AWS Batch제공 Amazon EKS 최적화 Amazon Linux 2 AMIs를 계속 사용할 수 있지만 이러한 컴퓨팅 환경은 더 이상 새로운 소프트웨어 업데이트, 보안 패치 또는 버그 수정을 받지 않습니다 AWS. 혼합 AMI 환경은 기존 AL2 기반 워크로드와의 호환성을 유지하면서 워크로드를 AL2023으로 점진적으로 마이그레이션할 수 있게 해 주므로 전환 기간 동안 유용합니다.

두 AMI 유형을 모두 사용하는 예제 구성:

```
{
  "computeResources": {
    "launchTemplate": {
      "launchTemplateId": "TemplateId",
      "version": "1",
      "userdataType": "EKS_BOOTSTRAP_SH",
      "overrides": [
        {
          "instanceType": "c5.large",
          "imageId": "ami-al2-custom",
          "userdataType": "EKS_BOOTSTRAP_SH"
        },
        {
          "instanceType": "c6a.large",
          "imageId": "ami-al2023-custom",
          "userdataType": "EKS_NODEADM"
        }
      ]
    },
    "instanceTypes": ["c5.large", "c6a.large"]
  }
```

```
}
```

지원되는 Kubernetes 버전

AWS Batch Amazon EKS의는 현재 다음 Kubernetes 버전을 지원합니다.

- 1.35
- 1.34
- 1.33
- 1.32
- 1.31
- 1.30
- 1.29

CreateComputeEnvironment API 작업 또는 UpdateComputeEnvironment API 작업을 사용하여 컴퓨팅 환경을 생성하거나 업데이트할 때 다음과 유사한 오류 메시지가 표시될 수 있습니다. EC2Configuration에서 지원되지 않는 Kubernetes 버전을 지정하는 경우 이 문제가 발생합니다.

```
At least one imageKubernetesVersion in EC2Configuration is not supported.
```

이 문제를 해결하려면 컴퓨팅 환경을 삭제하고 지원되는 Kubernetes 버전으로 다시 생성하세요.

Amazon EKS 클러스터에서 마이너 버전 업그레이드를 수행할 수 있습니다. 예를 들어 마이너 버전이 지원되지 않는 경우에도 클러스터를 1.xx에서 1.yy로 업그레이드할 수 있습니다.

하지만 메이저 버전 업데이트 후에는 컴퓨팅 환경 상태가 INVALID로 변경될 수 있습니다. 메이저 버전을 1.xx에서 2.yy로 업그레이드하는 경우를 예로 들 수 있습니다. 메이저 버전이에서 지원되지 않는 경우 다음과 유사한 오류 메시지가 AWS Batch표시됩니다.

```
reason=CLIENT_ERROR - ... EKS Cluster version [2.yy] is unsupported
```

컴퓨팅 환경의 Kubernetes 버전 업데이트

를 사용하면 Amazon EKS 클러스터 업그레이드를 지원하도록 컴퓨팅 환경의 Kubernetes 버전을 업데이트할 AWS Batch수 있습니다. 컴퓨팅 환경의 Kubernetes 버전은가 작업을 실행하기 위해 AWS

Batch 시작하는 Kubernetes 노드의 Amazon EKS AMI 버전입니다. Amazon EKS 클러스터의 컨트롤 플레인 버전을 업데이트하기 전 또는 후에 Amazon EKS 노드에서 Kubernetes 버전 업그레이드를 수행할 수 있습니다. 컨트롤 플레인을 업그레이드한 후에는 노드를 업데이트하는 것이 좋습니다. 자세한 내용은 Amazon EKS 사용 설명서의 Amazon EKS 클러스터 Kubernetes 버전 업데이트를 참조하세요.

컴퓨팅 환경의 Kubernetes 버전을 업그레이드하려면 UpdateComputeEnvironment API 작업을 사용합니다.

```
$ aws batch update-compute-environment \
    --compute-environment <compute-environment-name> \
    --compute-resources \
      'ec2Configuration=[{imageType=EKS_AL2023,imageKubernetesVersion=1.35}]'
```

Kubernetes 노드에 대한 공동 책임

컴퓨팅 환경의 유지 관리는 공동의 책임입니다.

- AWS Batch 노드, 레이블, 테인트, 네임스페이스, 시작 템플릿 또는 오토 스케일링 그룹을 변경하거나 제거하지 마십시오. AWS Batch 관리형 노드에 테인트를 추가하지 마세요. 이러한 변경을 수행하면 컴퓨팅 환경이 지원되지 없으며 유휴 인스턴스를 비롯한 장애가 발생합니다.
- 포드를 AWS Batch 관리형 노드로 대상으로 지정하지 마세요. 포드의 대상을 관리형 노드로 설정하면 규모 조정이 중단되고 작업 대기열이 멈추는 현상이 발생합니다. 자체 관리형 노드 또는 관리형 노드 그룹에서를 사용하지 않는 워크로드 AWS Batch 를 실행합니다. 자세한 내용은 Amazon EKS 사용 설명서의 관리형 노드 그룹을 참조하세요.
- AWS Batch 관리형 노드에서 실행DaemonSet되도록를 대상으로 지정할 수 있습니다. 자세한 내용은 AWS Batch 관리형 노드DaemonSet에서 실행 단원을 참조하십시오.

AWS Batch 는 컴퓨팅 환경 AMIs 자동으로 업데이트하지 않습니다. 업데이트하는 것은 사용자의 책임입니다. AMI를 최신 AMI 버전으로 업그레이드하려면 다음 명령을 실행합니다.

```
$ aws batch update-compute-environment \
    --compute-environment <compute-environment-name> \
    --compute-resources 'updateToLatestImageVersion=true'
```

AWS Batch 는 Kubernetes 버전을 자동으로 업그레이드하지 않습니다. 다음 명령을 실행하여 컴퓨팅 환경의 Kubernetes 버전을 *1.35*로 업데이트합니다.

```
$ aws batch update-compute-environment \
```

```
    --compute-environment <compute-environment-name> \
    --compute-resources \
      'ec2Configuration=[{imageType=EKS_AL2023,imageKubernetesVersion=1.35}]'
```

최신 AMI 또는 Kubernetes 버전으로 업데이트하는 경우 작업이 업데이트될 때 작업을 종료할지 여부(`terminateJobsOnUpdate`)와 실행 중인 작업 실행이 완료되지 않은 경우 인스턴스가 교체될 때까지 기다릴 시간을 지정할 수 있습니다(`jobExecutionTimeoutMinutes`). 자세한 내용은 에서 컴퓨팅 환경 업데이트 AWS Batch 섹션 및 UpdateComputeEnvironment API 작업에 설정된 인프라 업데이트 정책(UpdatePolicy)을 참조하세요.

AWS Batch 관리형 노드DaemonSet에서 실행

AWS Batch 는 AWS Batch 관리형 Kubernetes 노드에 테인트를 설정합니다. 다음를 사용하여 AWS Batch 관리형 노드에서 실행DaemonSet되도록를 대상으로 지정할 수 있습니다`tolerations`.

```
tolerations:
  - key: "batch.amazonaws.com/batch-node"
    operator: "Exists"
```

이 작업을 수행하는 또 다른 방법은 다음 `tolerations`를 사용하는 것입니다.

```
tolerations:
  - key: "batch.amazonaws.com/batch-node"
    operator: "Exists"
    effect: "NoSchedule"
  - key: "batch.amazonaws.com/batch-node"
    operator: "Exists"
    effect: "NoExecute"
```

Amazon EKS 시작 템플릿 사용자 지정

AWS Batch Amazon EKS의는 시작 템플릿을 지원합니다. 시작 템플릿으로 수행할 수 있는 작업에는 제약이 있습니다.

Important

- EKS AL2 AMIs 경우를 AWS Batch 실행합니다`/etc/eks/bootstrap.sh`. 시작 템플릿이나 cloud-init user-data 스크립트에서 `/etc/eks/bootstrap.sh`를 실행하지

않습니다. `--kubelet-extra-args` 파라미터 외에 다른 파라미터를 bootstrap.sh로 추가할 수 있습니다. 이렇게 하려면 `/etc/aws-batch/batch.config` 파일에 `AWS_BATCH_KUBELET_EXTRA_ARGS` 변수를 설정합니다. 자세한 내용은 다음 예제를 참조하세요.

- EKS AL2023의 경우는 EKS의 NodeConfigSpec을 AWS Batch 사용하여 인스턴스를 EKS 클러스터에 조인합니다. AWS Batch 는 EKS 클러스터에 대해 NodeConfigSpec의 ClusterDetails를 채우므로 지정할 필요가 없습니다.

Note

가 값을 AWS Batch 재정의하므로 시작 템플릿에서 다음 NodeConfigSpec 설정을 설정하지 않는 것이 좋습니다. 자세한 내용은 Kubernetes 노드에 대한 공동 책임 단원을 참조하십시오.

- `Taints`
- `Cluster Name`
- `apiServerEndpoint`
- `certificatAuthority`
- `CIDR`
- 접두사 `batch.amazonaws.com/`가 있는 레이블을 생성하지 마세요.

Note

CreateComputeEnvironment를 호출한 후 시작 템플릿이 변경된 경우 UpdateComputeEnvironment를 호출하여 교체할 시작 템플릿의 버전을 평가해야 합니다.

주제

- kubelet 추가 인수 추가
- 컨테이너 런타임 구성
- Amazon EFS 볼륨 마운트
- IPv6 지원

`kubelet` 추가 인수 추가

AWS Batch 는 `kubelet` 명령에 추가 인수 추가를 지원합니다. 지원되는 파라미터 목록은 Kubernetes 설명서의 kubelet 섹션을 참조하세요. EKS AL2 AMI에 대한 다음 예제에서는 *`--node-labels mylabel=helloworld`*가 `kubelet` 명령줄에 추가됩니다.

```
MIME-Version: 1.0
      Content-Type: multipart/mixed; boundary="==MYBOUNDARY=="

      --==MYBOUNDARY==
      Content-Type: text/x-shellscript; charset="us-ascii"

      #!/bin/bash
      mkdir -p /etc/aws-batch

      echo AWS_BATCH_KUBELET_EXTRA_ARGS=\"--node-labels mylabel=helloworld\" >> /etc/
aws-batch/batch.config

      --==MYBOUNDARY==--
```

EKS AL2023 AMI의 경우, 파일 형식은 YAML입니다. 지원되는 파라미터 목록은 Kubernetes 설명서의 NodeConfigSpec 섹션을 참조하세요. EKS AL2023 AMI에 대한 다음 예제에서는 *`--node-labels mylabel=helloworld`*가 `kubelet` 명령줄에 추가됩니다.

```
MIME-Version: 1.0
Content-Type: multipart/mixed; boundary="==MYBOUNDARY=="

--==MYBOUNDARY==
Content-Type: application/node.eks.aws

apiVersion: node.eks.aws/v1alpha1
kind: NodeConfig
spec:
  kubelet:
    flags:
    - --node-labels=mylabel=helloworld

--==MYBOUNDARY==--
```

컨테이너 런타임 구성

환경 변수를 사용하여 AWS Batch CONTAINER_RUNTIME 관리형 노드에서 컨테이너 런타임을 구성할 수 있습니다. 다음 예제에서는 컨테이너 런타임을 containerd 실행 시점의 bootstrap.sh로 설정합니다. 자세한 내용은 Kubernetes 설명서의 containerd 섹션을 참조하세요.

최적화 EKS_AL2023 또는 EKS_AL2023_NVIDIA AMI를 사용하는 경우, containerd만 지원되므로 컨테이너 런타임을 지정할 필요가 없습니다.

> **Note**
>
> CONTAINER_RUNTIME 환경 변수는 bootstrap.sh의 --container-runtime 옵션과 동일합니다. 자세한 내용은 Kubernetes 설명서의 Options 섹션을 참조하세요.

```
MIME-Version: 1.0
      Content-Type: multipart/mixed; boundary="==MYBOUNDARY=="

      --==MYBOUNDARY==
      Content-Type: text/x-shellscript; charset="us-ascii"

      #!/bin/bash
      mkdir -p /etc/aws-batch

      echo CONTAINER_RUNTIME=containerd >> /etc/aws-batch/batch.config

      --==MYBOUNDARY==--
```

Amazon EFS 볼륨 마운트

시작 템플릿을 사용하여 볼륨을 노드에 마운트할 수 있습니다. 다음 예제에서는 cloud-config packages 및 runcmd 설정이 사용됩니다. 자세한 내용은 cloud-init 설명서의 Cloud 구성 예제를 참조하세요.

```
MIME-Version: 1.0
Content-Type: multipart/mixed; boundary="==MYBOUNDARY=="

--==MYBOUNDARY==
Content-Type: text/cloud-config; charset="us-ascii"
```

```
packages:
- amazon-efs-utils

runcmd:
- file_system_id_01=fs-abcdef123
- efs_directory=/mnt/efs

- mkdir -p ${efs_directory}
- echo "${file_system_id_01}:/ ${efs_directory} efs _netdev,noresvport,tls,iam 0 0"
 >> /etc/fstab
- mount -t efs -o tls ${file_system_id_01}:/ ${efs_directory}

--==MYBOUNDARY==--
```

작업에서 이 볼륨을 사용하려면 RegisterJobDefinition에 대한 eksProperties 파라미터에 이 볼륨을 추가해야 합니다. 다음 예제는 작업 정의의 많은 부분입니다.

```
{
    "jobDefinitionName": "MyJobOnEks_EFS",
    "type": "container",
    "eksProperties": {
        "podProperties": {
            "containers": [
                {
                    "image": "public.ecr.aws/amazonlinux/amazonlinux:2",
                    "command": ["ls", "-la", "/efs"],
                    "resources": {
                        "limits": {
                            "cpu": "1",
                            "memory": "1024Mi"
                        }
                    },
                    "volumeMounts": [
                        {
                            "name": "efs-volume",
                            "mountPath": "/efs"
                        }
                    ]
                }
            ],
            "volumes": [
                {
                    "name": "efs-volume",
```

```
                    "hostPath": {
                        "path": "/mnt/efs"
                    }
                }
            ]
        }
    }
}
```

노드에서 Amazon EFS 볼륨은 `/mnt/efs` 디렉터리에 마운트됩니다. Amazon EKS 작업을 위한 컨테이너에서 볼륨은 `/efs` 디렉터리에 마운트됩니다.

IPv6 지원

AWS Batch 는 IPv6 주소가 있는 Amazon EKS 클러스터를 지원합니다. AWS Batch 지원에는 사용자 지정이 필요하지 않습니다. 하지만 시작하기 전에 Amazon EKS 사용 설명서의 포드 및 서비스에 IPv6 주소 할당에 설명된 고려 사항 및 조건을 검토하는 것이 좋습니다.

EKS AL2에서 EKS AL2023으로 업그레이드하는 방법

Amazon EKS 최적화 AMI는 Amazon Linux 2(AL2) 및 Amazon Linux 2023(AL2023) 기반의 두 패밀리에서 사용할 수 있습니다. AL2023은 클라우드 애플리케이션에 안전하고 안정적이면서 고성능의 환경을 제공하도록 설계된 Linux 기반 운영 체제입니다. AL2와 AL2023의 주요 차이점에 대한 자세한 내용은 Amazon EKS 사용 설명서의 Amazon Linux 2와 Amazon Linux 2023으로의 업그레이드를 참조하세요.

> **Important**
>
> AWS 는 AL2-optimized AL2-accelerated AMIs에 대한 지원을 종료했습니다. AWS Batch Amazon Linux 2를 사용하는 Amazon EKS 컴퓨팅 환경은 더 이상 소프트웨어 업데이트, 보안 패치 또는 버그 수정을 수신하지 않습니다 AWS. 최적의 성능과 보안을 유지하려면 AWS Batch Amazon EKS 컴퓨팅 환경을 Amazon Linux 2023으로 마이그레이션하는 것이 좋습니다. 수명 주기 종료 후 Amazon EKS 최적화 Amazon Linux 2 AMI에서 이러한 컴퓨팅 환경을 유지하는 것은 사용자의 책임입니다.

컴퓨팅 환경의 구성 방식에 따라 AL2에서 AL2023으로의 다음 업그레이드 경로 중 하나를 사용할 수 있습니다.

Ec2Configuration.ImageType을 사용한 업그레이드

- 시작 템플릿 또는 시작 템플릿 재정의를 사용하지 않는 경우 Ec2Configuration.ImageType을 EKS_AL2023 또는 로 변경EKS_AL2023_NVIDIA한 다음 UpdateComputeEnvironment를 실행합니다.
- Ec2Configuration.ImageIdOverride를 지정하는 경우, Ec2Configuration.ImageType이 Ec2Configuration.ImageIdOverride에 지정된 AMI 유형과 일치해야 합니다.

  ImageIdOverride와 ImageType을 불일치시키는 경우에는 노드가 클러스터에 조인하지 않습니다.

시작 템플릿을 사용하여 업그레이드

- 시작 템플릿 또는 시작 템플릿 재정의에 정의된 kubelet 추가 인수가 있는 경우 새 kubelet 추가 인수 형식으로 업데이트해야 합니다.

  kubelet 추가 인수 형식을 불일치시키는 경우에는 추가 인수가 적용되지 않습니다.
- AL2023 AMI의 경우, containerd가 유일하게 지원되는 컨테이너 런타임입니다. 시작 템플릿에 EKS_AL2023에 대한 이름을 지정할 필요가 없습니다.

  를 사용하여 사용자 지정 컨테이너 런타임을 지정할 수 없습니다EKS_AL2023.
- EKS_AL2023 기반 AMI를 지정하는 시작 템플릿 또는 시작 템플릿 재정의를 사용하는 경우에는 userdataType을 EKS_NODEADM으로 설정해야 합니다.

  userdataType과 AMI를 불일치시키는 경우에는 노드가 EKS 클러스터에 조인하지 않습니다.

용 서비스 환경 AWS Batch

서비스 환경을 사용하면 AWS Batch 를 SageMaker AI와 통합할 수 있습니다. 서비스 환경에는의 대기열, 예약 및 우선 순위 관리 기능을 제공하면서 SageMaker 훈련 작업을 제출하고 관리하는 AWS Batch 데 필요한 SageMaker AI별 구성 파라미터 AWS Batch가 포함되어 있습니다.

서비스 환경에서는 데이터 과학자와 ML 엔지니어가 우선 순위가 있는 SageMaker 훈련 작업을 서비스 작업 대기열에 제출할 수 있습니다. 이 통합은 ML 워크로드를 수동으로 조정할 필요를 제거하고, 우발적인 과다 지출을 방지하고, 조직의 기계 학습 워크플로 전반에서 리소스 사용률을 개선합니다.

주제

의 서비스 환경이란? AWS Batch

서비스 환경은 SageMaker AI AWS Batch 와 통합하는 데 필요한 구성 파라미터를 포함하는 AWS Batch 리소스입니다. 서비스 환경을 사용하면 AWS Batch 가 SageMaker 훈련 작업을 제출하고 관리하는 동시에 AWS Batch의 대기열, 일정 및 우선 순위 관리 기능을 제공할 수 있습니다.

서비스 환경은 데이터 과학 팀이 기계 학습 워크로드를 관리할 때 직면하는 일반적인 문제를 해결합니다. 조직은 우발적인 과다 지출을 방지하거나, 예산 제약을 충족하거나, 예약 인스턴스 비용을 절감하거나, 워크로드에 특정 인스턴스 유형을 사용하기 위해 모델 훈련에 사용할 수 있는 인스턴스 수를 종종 제한합니다. 그러나 데이터 과학자는 할당된 인스턴스에서 가능한 것보다 더 많은 워크로드를 동시에 실행하기를 원할 수 있으며, 이 경우 어느 워크로드가 언제 실행될지를 결정하기 위해 수동 조정이 필요합니다.

이 조정 문제는 데이터 과학자가 몇 명뿐인 팀부터 대규모 운영에 이르기까지 모든 규모의 조직에 영향을 미칩니다. 조직이 성장함에 따라 복잡성이 증가하여 워크로드 조정을 관리하는 데 더 많은 시간이 필요하고 종종 인프라 관리자의 개입이 필요해집니다. 이러한 수동 작업은 시간을 낭비하고 인스턴스 효율성을 줄여 고객에게 실제 비용을 초래합니다.

서비스 환경에서는 데이터 과학자와 ML 엔지니어가 우선순위가 있는 SageMaker 훈련 작업을 구성 가능한 대기열에 제출하여 리소스를 사용할 수 있게 되는 즉시 개입 없이 워크로드가 자동으로 실행되도

록 할 수 있습니다. 이 통합은 AWS Batch의 광범위한 대기열 및 예약 기능을 활용하여 고객이 조직의 목표에 맞게 대기열 및 예약 정책을 사용자 지정할 수 있도록 합니다.

서비스 환경이 다른 AWS Batch 구성 요소와 작동하는 방식

서비스 환경은 다른 AWS Batch 구성 요소와 통합되어 SageMaker 훈련 작업 대기열을 활성화합니다.

- 작업 대기열 - 서비스 환경은 작업 대기열과 연결되어 대기열이 SageMaker 훈련 작업에 대한 서비스 작업을 처리할 수 있도록 해 줍니다.
- 서비스 작업 - 서비스 작업을 서비스 환경과 연결된 대기열에 제출하면는 환경의 구성을 AWS Batch 사용하여 해당 SageMaker 훈련 작업을 제출합니다.
- 예약 정책 - 서비스 환경은 AWS Batch 예약 정책과 함께 작동하여 SageMaker 훈련 작업의 실행 순서의 우선 순위를 지정하고 관리합니다.

이 통합을 통해 SageMaker 훈련 작업 AWS Batch의 전체 기능과 유연성을 유지하면서의 성숙한 대기열 및 예약 기능을 활용할 수 있습니다.

서비스 환경 모범 사례

서비스 환경은 대규모로 SageMaker 훈련 작업을 관리할 수 있는 기능을 제공합니다. 이러한 모범 사례를 따르면 기계 학습 워크플로에 영향을 미칠 수 있는 일반적인 구성 문제를 방지하면서 비용, 성능 및 운영 효율성을 최적화하는 데 도움이 됩니다.

서비스 환경 용량을 계획할 때는 SageMaker 훈련 작업 대기열에 적용되는 특정 할당량 및 제한을 고려하세요. 각 서비스 환경에는 동시에 실행할 수 있는 SageMaker 훈련 작업 수를 직접 제어하는, 인스턴스 수로 표현된 최대 용량 제한이 있습니다. 이러한 제한을 이해하면 리소스 경합을 방지하는 데 도움이 되고 예측 가능한 작업 실행 시간을 보장합니다.

최적의 서비스 환경 성능은 SageMaker 훈련 작업 예약의 고유한 특성을 이해하는 데 달려 있습니다. 기존의 컨테이너화된 작업과 달리 서비스 작업은 SageMaker AI가 필요한 훈련 인스턴스를 획득하고 프로비저닝하는 동안 `SCHEDULED` 상태를 거치면서 전환됩니다. 이는 작업 시작 시간이 인스턴스 가용성 및 리전 용량에 따라 크게 달라질 수 있음을 의미합니다.

> **Important**
>
> 서비스 환경에는 SageMaker 훈련 워크로드의 규모를 조정하는 능력에 영향을 미칠 수 있는 특정 할당량이 있습니다. 계정당 최대 50개의 서비스 환경을 생성할 수 있으며, 각 작업 대기열은 한 개의 연결된 서비스 환경만 지원합니다. 또한 개별 작업에 대한 서비스 요청 페이로드는

10KiB로 제한되며 SubmitServiceJob API는 계정당 초당 5개의 트랜잭션으로 제한됩니다. 용량 계획 중에 이러한 제한을 이해하면 예상치 못한 규모 조정 제약을 방지할 수 있습니다.

서비스 환경을 효과적으로 모니터링하려면 AWS Batch 및 SageMaker AI 서비스 지표 모두에 주의를 기울여야 합니다. 작업 상태 전환은 시스템 성능에 대한 소중한 정보를 제공합니다. 특히 SCHEDULED 상태에서 소요된 시간은 용량 가용성 패턴을 나타냅니다. 컴퓨팅 환경과 유사한 자체 수명 주기 상태를 유지하는 서비스 환경은 CREATING, VALID, INVALID 및 DELETING 상태를 거치며, 이러한 상태는 운영 상태 확인을 위해 모니터링되어야 합니다. 성숙한 모니터링 방식을 가진 조직은 일반적으로 대기열 깊이, 작업 완료율 및 인스턴스 사용률 패턴을 추적하여 시간 경과에 따라 서비스 환경 구성을 최적화합니다.

의 서비스 환경 상태 및 수명 주기 AWS Batch

서비스 환경은 현재 운영 상태와 SageMaker 훈련 작업을 처리할 준비가 되었음을 나타내는 수명 주기 상태를 유지합니다. 이러한 상태를 이해하면 서비스 환경 상태를 모니터링하고 구성 문제를 해결하며 신뢰적인 작업 처리를 보장하는 데 도움이 됩니다. 상태 관리 시스템은 컴퓨팅 환경의 설정된 패턴을 따르면서 SageMaker 훈련 작업 통합의 고유한 요구 사항을 수용합니다.

서비스 환경 상태는 구성 검증, 리소스 가용성 및 운영 상태 확인을 AWS Batch 기반으로에서 자동으로 관리됩니다. 물리적 인프라를 관리하는 컴퓨팅 환경과 달리 서비스 환경은 구성 검증 및 SageMaker AI 서비스와의 통합 준비 태세에 중점을 둡니다. 상태 전환은 서비스 환경이 SageMaker 훈련 작업을 성공적으로 제출하고 관리할 수 있는지 여부에 대한 가시성을 제공합니다.

서비스 환경 상태 정의

서비스 환경은 현재 운영 상태와 SageMaker 훈련 작업을 처리할 준비가 되었음을 나타내는 네 가지 상태 중 하나일 수 있습니다. 각 상태는 최초 생성부터 운영 준비, 최종 삭제에 이르기까지 서비스 환경 수명 주기의 특정 단계를 나타냅니다. 다음 표는 각 상태와 그 의미를 설명합니다.

State	설명
CREATING	서비스 환경을 생성할 때의 최초 상태입니다. 이 상태에서는 구성 파라미터를 AWS Batch 검증하고 SageMaker AI 서비스와의 통합을 설정합니다. 서비스 환경은 작업을 처리할 수 없으며, 이와 연결된 모든 작업 대기열은 서비스 작업 제

State	설명
	출을 수락하지 않습니다. 일반적으로 생성 프로세스는 올바르게 구성된 서비스 환경에서 몇 초 내에 완료됩니다.
VALID	서비스 환경이 모든 구성 검증 검사를 통과했으며 SageMaker 훈련 작업을 처리할 준비가 되었음을 나타내는 운영 상태. 이 상태는 서비스 환경 구성이 올바르고, 필요한 모든 권한이 있으며, 사용자를 대신하여 SageMaker AI에 작업을 성공적으로 제출할 AWS Batch 수 있음을 나타냅니다. 서비스 환경은 이 상태에서 대부분의 운영 수명 주기를 소비합니다.
INVALID	서비스 환경에서 SageMaker 훈련 작업을 처리할 수 없는 구성 또는 권한 문제가 발생했음을 나타내는 상태. 유효하지 않은 서비스 환경과 연결된 작업 대기열은 해당 문제가 해결될 때까지 새 서비스 작업 제출을 처리할 수 없습니다.
DELETING	서비스 환경 삭제를 요청했을 때 발생하는 상태. 이 상태에서는 활성 SageMaker 훈련 작업이 환경과 연결되어 있지 않고 필요한 정리 작업을 수행하는지 AWS Batch 확인합니다. 이 상태의 서비스 환경은 새 작업 제출을 처리할 수 없으며, 연결된 모든 리소스가 제대로 정리되면 삭제 프로세스가 완료됩니다.

서비스 환경 상태 전환

서비스 환경 상태 전환은 구성 변경, 검증 결과 및 운영 상태 모니터링을 기반으로 자동 수행됩니다. AWS Batch 서비스는 서비스 환경 상태를 지속적으로 모니터링하고 그에 따라 상태를 업데이트합니다. 이러한 전환을 이해하면 구성 변경이 적용되는 시기를 예측하고 잘못된 상태를 유발하는 문제를 해결하는 방법을 찾는 데 도움이 됩니다.

성공적으로 생성 및 검증이 완료되면 서비스 환경이 CREATING에서 VALID로 전환됩니다. 이 전환은 모든 구성 파라미터가 올바르고, 필수 IAM 권한이 올바르게 구성되었으며, 서비스 환경이 SageMaker

AI 서비스와 성공적으로 통합될 수 있음을 확인합니다. 일단 VALID 상태가 되면 해당 작업 대기열이 서비스 작업 제출을 처리하기 시작할 수 있습니다.

구성 검증에 실패하거나 종속성을 사용할 수 없게 되면 서비스 환경이 VALID에서 INVALID로 전환됩니다. 이는 IAM 역할 수정, 할당량을 위반하는 용량 제한 변경 또는 서비스 환경의 올바른 작동에 영향을 미치는 외부 리소스 수정으로 인해 발생할 수 있습니다. 상태 사유 필드는 유효하지 않은 상태의 원인에 대한 구체적인 세부 정보를 제공합니다.

관련 문제가 해결되면 서비스 환경이 INVALID에서 VALID로 다시 전환될 수 있습니다. 여기에는 IAM 권한 업데이트, 용량 구성 수정 또는 필요한 AWS 리소스에 대한 액세스 복원이 포함될 수 있습니다. 전환은 일반적으로 AWS Batch 가 구성 문제가 해결되었음을 감지하면 자동으로 이루어집니다.

에서 서비스 환경 생성 AWS Batch

에서 SageMaker 훈련 작업을 실행하려면 먼저 서비스 환경을 생성 AWS Batch해야 합니다. 가 SageMaker AI 서비스와 통합하고 사용자를 대신하여 SageMaker 훈련 작업을 제출하는 AWS Batch 데 필요한 구성 파라미터가 포함된 서비스 환경을 생성할 수 있습니다.

사전 조건

서비스 환경을 생성하려면 먼저 다음을 갖추어야 합니다.

- IAM 권한 - 서비스 환경을 생성하고 관리할 수 있는 권한. 자세한 내용은 AWS Batch IAM 정책, 역할 및 권한 단원을 참조하십시오.

Create a service environment (AWS Console)

AWS Batch 콘솔을 사용하여 웹 인터페이스를 통해 서비스 환경을 생성합니다.

서비스 환경을 생성하는 방법

1. https://console.aws.amazon.com/batch/ AWS Batch 콘솔을 엽니다.
2. 탐색 창에서 환경을 선택합니다.
3. 환경 생성을 선택하고 서비스 환경을 선택합니다.
4. 서비스 환경 구성에서 SageMaker AI를 선택합니다.
5. 이름에 서비스 환경의 고유한 이름을 입력합니다. 유효한 문자는 a~z, A~Z, 0~9, 하이픈(-) 및 밑줄(_)입니다.

6. 최대 인스턴스 수에 동시 훈련 인스턴스의 최대 수를 입력합니다.
7. (선택 사항) 태그 추가를 선택하고 키-값 페어를 입력하여 태그를 추가합니다.
8. 다음을 선택합니다.
9. 새 서비스 환경의 세부 정보를 검토하고 서비스 환경 생성을 선택합니다.

Create a service environment (AWS CLI)

`create-service-environment` 명령을 사용하여 AWS CLI를 사용하여 서비스 환경을 생성합니다.

서비스 환경을 생성하는 방법

1. 기본 필수 파라미터를 사용하여 서비스 환경을 생성합니다.

```
aws batch create-service-environment \
    --service-environment-name my-sagemaker-service-env \
    --service-environment-type SAGEMAKER_TRAINING \
    --capacity-limits capacityUnit=NUM_INSTANCES,maxCapacity=10
```

2. (선택 사항) 태그와 함께 서비스 환경을 선택합니다.

```
aws batch create-service-environment \
    --service-environment-name my-sagemaker-service-env \
    --service-environment-type SAGEMAKER_TRAINING \
    --capacity-limits capacityUnit=NUM_INSTANCES,maxCapacity=10 \
    --tags team=data-science,project=ml-training
```

3. 서비스 환경이 성공적으로 생성되었는지 확인합니다.

```
aws batch describe-service-environments \
    --service-environment my-sagemaker-service-env
```

서비스 환경이 환경 목록에 `CREATING` 상태로 나타납니다. 생성이 성공적으로 완료되면 상태가 `VALID`로 변경되고 서비스 환경에서 작업 처리를 시작할 수 있도록 서비스 작업 대기열을 추가할 준비가 됩니다.

에서 서비스 환경 업데이트 AWS Batch

서비스 환경을 업데이트하여 용량 제한을 수정하거나, 운영 상태를 변경하거나, 리소스 태그를 업데이트할 수 있습니다. 서비스 환경 업데이트는 환경을 다시 생성하지 않고도 변경된 SageMaker 훈련 워크로드 요구 사항에 맞게 용량을 조정하거나 운영 설정을 수정할 수 있게 해 줍니다. 서비스 환경을 업데이트하기 전에, 어느 파라미터를 수정할 수 있으며 변경 사항이 실행 중인 작업에 어떤 영향을 미치는지 이해해야 합니다.

서비스 환경의 용량 제한, 상태 또는 태그를 변경할 수 있습니다.

> **Note**
>
> 용량 제한이 낮아지면 이미 SCHEDULED, STARTING 또는 RUNNING 상태인 작업은 종료되지 않습니다. 이미 예약된 작업으로 인해 현재 사용률이 일시적으로 새 용량 제한을 초과할 수 있지만 새 작업은 예약 중에 업데이트된 제한을 준수합니다.

Update a service environment (AWS Console)

AWS Batch 콘솔을 사용하여 웹 인터페이스를 통해 서비스 환경을 업데이트합니다.

서비스 환경을 업데이트하려면

1. https://console.aws.amazon.com/batch/ AWS Batch 콘솔을 엽니다.
2. 탐색 창에서 환경을 선택합니다.
3. 서비스 환경 탭을 선택합니다.
4. 업데이트할 서비스 환경을 선택합니다.
5. 작업을 선택한 후 다음 중 하나를 선택합닏.
   - 상태 - 활성화 또는 비활성화를 선택하여 상태를 변경합니다.
   - 용량 제한 - 최대 인스턴스 수를 수정합니다.
6. 변경 사항 저장을 선택하여 변경 사항을 적용합니다.

서비스 환경은 즉시 업데이트됩니다. 환경 세부 정보를 확인하여 변경 사항이 성공적으로 적용되었는지 확인합니다. 서비스 환경을 비활성화한 경우, 다시 활성화할 때까지 연결된 작업 대기열이 새 서비스 작업 제출의 처리를 중지합니다.

Update a service environment (AWS CLI)

`update-service-environment` 명령을 사용하여 AWS CLI를 사용하여 서비스 환경을 수정합니다.

서비스 환경 용량 제한을 업데이트하려면

1. 서비스 환경의 용량 제한을 업데이트합니다.

```
aws batch update-service-environment \
    --service-environment my-sagemaker-service-env \
    --capacity-limits capacityUnit=NUM_INSTANCES,maxCapacity=20
```

2. 업데이트가 성공적으로 적용되었는지 확인합니다.

```
aws batch describe-service-environments \
    --service-environments my-sagemaker-service-env
```

서비스 환경 상태를 업데이트하려면

1. 새 작업 처리를 중지하려면 서비스 환경을 비활성화합니다.

```
aws batch update-service-environment \
    --service-environment my-sagemaker-service-env \
    --state DISABLED
```

2. 서비스 환경을 다시 활성화하여 처리를 재개합니다.

```
aws batch update-service-environment \
    --service-environment my-sagemaker-service-env \
    --state ENABLED
```

서비스 환경 업데이트는 즉시 적용됩니다. 새 작업을 제출하기 전에 서비스 환경 상태를 모니터링하여 업데이트가 성공적으로 완료되었는지 확인합니다.

에서 서비스 환경 삭제 AWS Batch

SageMaker 훈련 작업에 더 이상 필요하지 않은 서비스 환경을 삭제할 수 있습니다. 서비스 환경을 삭제하면 구성이 제거되고 추가 작업 제출이 방지됩니다. 서비스 환경을 삭제하기 전에 해당 환경에 의존하는 활성 SageMaker 훈련 작업이 없으며 서비스 환경과 연결된 작업 대기열이 없는지 확인합니다.

Important

서비스 환경 삭제는 되돌릴 수 없습니다. 삭제한 후에는 서비스 환경 또는 해당 구성을 복구할 수 없습니다. 향후 유사한 기능이 필요한 경우 필요한 설정을 사용하여 새 서비스 환경을 생성해야 합니다. 나중에 다시 활성화해야 할 수 있는 경우, 서비스 환경을 삭제 대신 비활성화하는 것을 고려하세요.

Note

계정의 모든 서비스 환경을 삭제해도 AWS Batch 및 SageMaker AI 통합에 대해 생성된 서비스 연결 역할이 자동으로 제거되지는 않습니다. 서비스 연결 역할은 향후 서비스 환경을 생성에 계속 사용할 수 있습니다. 서비스 연결 역할을 제거하려는 경우, 계정에 서비스 환경이 없는지 확인한 후 IAM을 사용하여 별도로 삭제해야 합니다.

삭제 사전 조건

서비스 환경을 삭제하려면 먼저 모든 서비스 작업 대기열의 연결을 해제한 다음 서비스 환경을 비활성화해야 합니다.

서비스 환경을 삭제하기 전에:

- 활성 작업 확인 - 서비스 환경을 통해 현재 실행 중인 SageMaker 훈련 작업이 없는지 확인합니다.
- 작업 대기열 검토 - 서비스 환경과 연결된 작업 대기열을 식별하고 작업 대기열을 다른 서비스 환경에 연결하거나 작업 대기열을 비활성화하고 삭제합니다.

작업 대기열 관리: 삭제된 서비스 환경과 연결된 작업 대기열은 여전히 존재할 수 있지만 서비스 작업은 처리할 수 없습니다. 원래 서비스 환경을 삭제하기 전에 사용하지 않는 작업 대기열을 삭제하거나 다른 서비스 환경에 연결해야 합니다.

Delete a service environment (AWS Console)

AWS Batch 콘솔을 사용하여 웹 인터페이스를 통해 서비스 환경을 삭제합니다.

서비스 환경을 삭제하는 방법

1. https://console.aws.amazon.com/batch/ AWS Batch 콘솔을 엽니다.
2. 탐색 창에서 환경을 선택합니다.
3. 서비스 환경 탭을 선택한 다음 서비스 환경을 선택합니다.
4. 서비스 환경이 활성화되어 있는 경우, 작업을 선택한 다음 비활성화를 선택합니다.
5. 서비스 환경이 비활성화되어 있는 경우, 작업을 선택한 다음 삭제를 선택합니다.
6. 확인 대화 상자에서 확인을 선택합니다.

삭제가 진행되는 동안 서비스 환경에 `DELETING` 상태가 표시됩니다. 삭제가 완료되면 서비스 환경이 환경 목록에서 사라집니다.

Delete a service environment (AWS CLI)

`delete-service-environment` 명령을 사용하여 AWS CLI를 사용하여 서비스 환경을 제거합니다.

서비스 환경을 삭제하는 방법

1. 서비스 환경에 연결된 작업 대기열을 확인합니다.

```
aws batch describe-job-queues
```

서비스 환경과 연결된 작업 대기열이 있는 경우, 서비스 환경에서 작업 대기열의 연결을 해제하고 다른 서비스 환경에 연결하거나 작업 대기열을 삭제할 수 있습니다.

2. 서비스 환경을 비활성화합니다.

```
aws batch update-service-environment \
    --service-environment my-sagemaker-service-env \
    --state DISABLED
```

3. 서비스 환경을 삭제합니다.

```
aws batch delete-service-environment \
    --service-environment my-sagemaker-service-env
```

4. 삭제 프로세스를 모니터링합니다.

```
aws batch describe-service-environments \
    --service-environment my-sagemaker-service-env
```

삭제 프로세스 동안 서비스 환경은 DELETING 상태로 전환됩니다. 삭제가 완료되면 서비스 환경이 더 이상 설명 작업에 나열되지 않습니다. 연결된 작업 대기열은 남아 있지만 다른 서비스 환경과 연결될 때까지 서비스 작업을 처리할 수 없습니다.

작업 대기열

작업은 컴퓨팅 환경에서 실행되도록 예약할 수 있을 때까지 상주하는 작업 대기열에 제출됩니다. AWS 계정에는 여러 작업 대기열이 있을 수 있습니다. 예를 들어, 우선순위가 높은 작업에는 Amazon EC2 온디맨드 인스턴스를 사용하는 대기열을 생성하고, 우선순위가 낮은 작업에는 Amazon EC2 스팟 인스턴스를 사용하는 대기열을 생성할 수 있습니다. 작업 대기열에는 스케줄러에서 사용하는 우선순위가 지정되어 어느 대기열의 어느 작업이 먼저 실행될지가 결정됩니다.

주제

- 작업 대기열 생성
- 에서 작업 대기열 보기 AWS Batch
- AWS Batch의 작업 대기열 삭제
- 공정 공유 예약 정책
- 리소스 인식 일정 예약
- 할당량 관리
- 서비스 작업 용량 사용률 추적
- 컴퓨팅 작업 용량 사용률 추적

작업 대기열 생성

AWS Batch에서 작업을 제출하려면 먼저 작업 대기열을 생성해야 합니다. 작업 대기열을 생성할 때, 대기열에 컴퓨팅 환경을 한 개 이상 연결하고 우선권 순서를 지정합니다.

또한 작업 대기열에 AWS 배치 스케줄러가 작업을 배치할 순서를 결정하는 우선 순위도 설정합니다. 이는 컴퓨팅 환경이 여러 작업 대기열에 연결된 경우, 우선 순위가 높은 작업 대기열에 우선권이 부여된다는 의미입니다.

주제

- Amazon EC2 작업 대기열 생성
- Fargate 작업 대기열 생성
- Amazon EKS 작업 대기열 생성
- AWS Batch에서 SageMaker 훈련 작업 대기열 생성
- 작업 대기열 템플릿

Amazon EC2 작업 대기열 생성

Amazon Elastic Compute Cloud(Amazon EC2)에 대한 작업 대기열을 생성하려면 다음 단계를 완료합니다.

Amazon EC2 작업 대기열을 생성하기 위해

1. https://console.aws.amazon.com/batch/에서 AWS Batch 콘솔을 엽니다.
2. 탐색 모음에서 사용할 AWS 리전(을)를 선택합니다.
3. 탐색 창에서 작업 대기열을 선택합니다.
4. 생성을 선택합니다.
5. 오케스트레이션 유형으로 Amazon Elastic Compute Cloud(Amazon EC2)를 선택합니다.
6. 이름에 작업 대기열의 고유 이름을 입력합니다. 이름은 최대 128자까지 포함할 수 있으며, 대문자와 소문자, 숫자, 밑줄(_)을 포함할 수 있습니다.
7. 우선 순위에 작업 대기열의 우선 순위 값을 정수로 입력합니다. 우선 순위가 높은 작업 대기열은 동일한 컴퓨팅 환경과 연결된 우선 순위가 낮은 작업 대기열보다 우선합니다. 우선 순위는 내림차순으로 결정됩니다. 예를 들어, 우선 순위 값이 1인 작업 대기열은 우선 순위 값이 10인 작업 대기열보다 먼저 일정이 예약됩니다.
8. (선택 사항)예약 정책 Amazon 리소스 이름(ARN)에서 기존 예약 정책을 선택합니다.
9. 연결된 컴퓨팅 환경 섹션의 목록에서 작업 대기열과 연결할 컴퓨팅 환경을 한 개 이상 선택합니다. 대기열에서 작업 대기열 배치를 하려는 순서대로 컴퓨팅 환경을 선택합니다. 작업 스케줄러는 컴퓨팅 환경 순서를 사용하여 주어진 작업을 실행할 컴퓨팅 환경을 결정합니다. 작업 대기열과 연결하려면 컴퓨터 환경이 `VALID` 상태여야 합니다. 한 개의 작업 대기열에 최대 3개의 컴퓨팅 환경을 연결할 수 있습니다. 기존 컴퓨팅 환경이 없는 경우 컴퓨팅 환경 생성을 선택합니다.

   > **Note**
   >
   > 작업 대기열에 연결된 모든 컴퓨팅 환경은 동일한 프로비저닝 모델을 공유해야 합니다. AWS Batch(은)는 단일 작업 대기열에서 프로비저닝 모델을 혼합하여 사용하는 것을 지원하지 않습니다.

10. 컴퓨팅 환경 순서에서 위/아래 화살표를 선택하여 원하는 순서를 구성합니다.
11. 생성을 선택하여 완료하고 작업 대기열을 생성합니다.

Fargate 작업 대기열 생성

AWS Fargate에 대한 작업 대기열을 생성하려면 다음 단계를 완료합니다.

Fargate 작업 대기열을 생성하려면

1. https://console.aws.amazon.com/batch/에서 AWS Batch 콘솔을 엽니다.
2. 탐색 모음에서 사용할 AWS 리전(을)를 선택합니다.
3. 탐색 창에서 작업 대기열을 선택합니다.
4. 생성을 선택합니다.
5. 오케스트레이션 유형에서 Fargate를 선택합니다.
6. 이름에 작업 대기열의 고유 이름을 입력합니다. 이름은 최대 128자까지 포함할 수 있으며, 대문자와 소문자, 숫자, 밑줄(_)을 포함할 수 있습니다.
7. 우선 순위에 작업 대기열의 우선 순위 값을 정수로 입력합니다. 우선 순위가 높은 작업 대기열은 동일한 컴퓨팅 환경과 연결된 우선 순위가 낮은 작업 대기열보다 우선합니다. 우선 순위는 내림차순으로 결정됩니다. 예를 들어, 우선 순위 값이 1인 작업 대기열은 우선 순위 값이 10인 작업 대기열보다 먼저 일정이 예약됩니다.
8. (선택 사항)예약 정책 Amazon 리소스 이름(ARN)에서 기존 예약 정책을 선택합니다.
9. 연결된 컴퓨팅 환경 섹션의 목록에서 작업 대기열과 연결할 컴퓨팅 환경을 한 개 이상 선택합니다. 대기열에서 작업 대기열 배치를 하려는 순서대로 컴퓨팅 환경을 선택합니다. 작업 스케줄러는 컴퓨팅 환경 순서를 사용하여 주어진 작업을 실행할 컴퓨팅 환경을 결정합니다. 작업 대기열과 연결하려면 컴퓨터 환경이 `VALID` 상태여야 합니다. 한 개의 작업 대기열에 최대 3개의 컴퓨팅 환경을 연결할 수 있습니다.

   > **Note**
   >
   > 작업 대기열에 연결된 모든 컴퓨팅 환경은 동일한 프로비저닝 모델을 공유해야 합니다. AWS Batch(은)는 단일 작업 대기열에서 프로비저닝 모델을 혼합하여 사용하는 것을 지원하지 않습니다.

10. 컴퓨팅 환경 순서에서 위/아래 화살표를 선택하여 원하는 순서를 구성합니다.
11. 생성을 선택하여 완료하고 작업 대기열을 생성합니다.

Amazon EKS 작업 대기열 생성

Amazon Elastic Kubernetes Service(Amazon EKS)에 대한 작업 대기열을 생성하려면 다음 단계를 완료합니다.

Amazon EKS 작업 대기열 생성하기 위해

1. https://console.aws.amazon.com/batch/에서 AWS Batch 콘솔을 엽니다.
2. 탐색 모음에서 사용할 AWS 리전(을)를 선택합니다.
3. 탐색 창에서 작업 대기열을 선택합니다.
4. 생성을 선택합니다.
5. 오케스트레이션 유형으로 Amazon Elastic Kubernetes Service(Amazon EKS)를 선택합니다.
6. 이름에 작업 대기열의 고유 이름을 입력합니다. 이름은 최대 128자까지 포함할 수 있으며, 대문자와 소문자, 숫자, 밑줄(_)을 포함할 수 있습니다.
7. 우선 순위에 작업 대기열의 우선 순위 값을 정수로 입력합니다. 우선 순위가 높은 작업 대기열은 동일한 컴퓨팅 환경과 연결된 우선 순위가 낮은 작업 대기열보다 우선합니다. 우선 순위는 내림차순으로 결정됩니다. 예를 들어, 우선 순위 값이 1인 작업 대기열은 우선 순위 값이 10인 작업 대기열보다 먼저 일정이 예약됩니다.
8. (선택 사항)예약 정책 Amazon 리소스 이름(ARN)에서 기존 예약 정책을 선택합니다.
9. 연결된 컴퓨팅 환경 섹션의 목록에서 작업 대기열과 연결할 컴퓨팅 환경을 한 개 이상 선택합니다. 대기열에서 작업 대기열 배치를 하려는 순서대로 컴퓨팅 환경을 선택합니다. 작업 스케줄러는 컴퓨팅 환경 순서를 사용하여 주어진 작업을 실행할 컴퓨팅 환경을 결정합니다. 작업 대기열과 연결하려면 컴퓨터 환경이 `VALID` 상태여야 합니다. 한 개의 작업 대기열에 최대 3개의 컴퓨팅 환경을 연결할 수 있습니다.

> **Note**
>
> 작업 대기열에 연결된 모든 컴퓨팅 환경은 동일한 프로비저닝 모델을 공유해야 합니다. AWS Batch(은)는 단일 작업 대기열에서 프로비저닝 모델을 혼합하여 사용하는 것을 지원하지 않습니다.

> **Note**
>
> 작업 대기열과 연관된 모든 컴퓨팅 환경은 동일한 아키텍처를 공유해야 합니다. AWS Batch(은)는 단일 작업 대기열에서의 혼합 컴퓨팅 환경 아키텍처 유형을 지원하지 않습니다.

10. 컴퓨팅 환경 순서에서 위/아래 화살표를 선택하여 원하는 순서를 구성합니다.
11. 생성을 선택하여 완료하고 작업 대기열을 생성합니다.

AWS Batch에서 SageMaker 훈련 작업 대기열 생성

SageMaker 훈련 작업 대기열은 SageMaker AI 서비스와 직접 통합되어 기본 컴퓨팅 인프라를 관리할 필요 없이 서버리스 작업 예약을 제공합니다.

사전 조건

SageMaker 훈련 작업 대기열을 생성하기 전에 다음을 갖추어야 합니다.

- 서비스 환경 - 용량 제한을 정의하는 서비스 환경. 자세한 내용은 에서 서비스 환경 생성 AWS Batch 섹션을 참조하세요.
- IAM 권한 - AWS Batch 작업 대기열 및 서비스 환경을 생성하고 관리할 수 있는 권한. 자세한 내용은 AWS Batch IAM 정책, 역할 및 권한 섹션을 참조하세요.

Create a SageMaker Training job queue (AWS Batch console)

1. https://console.aws.amazon.com/batch/에서 AWS Batch 콘솔을 엽니다.
2. 탐색 창에서 작업 대기열을 선택한 다음 생성을 선택합니다.
3. 오케스트레이션 유형에서 SageMaker 훈련을 선택합니다.
4. 작업 대기열 구성에서 다음을 수행합니다.
    a. 이름에 작업 대기열의 이름을 입력합니다.
    b. 우선 순위에 0에서 1000 사이의 값을 입력합니다. 우선 순위가 높은 작업 대기열에는 서비스 환경에 대해 더 높은 선호도가 지정됩니다.
    c. (선택 사항)예약 정책 Amazon 리소스 이름(ARN)에서 기존 예약 정책을 선택합니다.
    d. 연결된 서비스 환경의 목록에서 작업 대기열과 연결할 서비스 환경을 선택합니다.

5. (선택 사항) 작업 상태 제한의 경우:

   a. 잘못된 구성에서 SERVICE_ENVIRONMENT_MAX_RESOURCE를 선택하고 최대 실행 가능 시간(초)을 입력합니다.

   b. 용량에서 INSUFFICIENT_INSTANCE_CAPACITY를 선택하고 최대 실행 가능 시간(초)을 입력합니다.

6. 작업 대기열 생성을 선택합니다.

Create a SageMaker Training job queue (AWS CLI)

`create-job-queue` 명령을 사용하여 SageMaker 훈련 작업 대기열을 생성합니다.

다음 예제는 서비스 환경을 사용하는 기본 SageMaker 훈련 작업 대기열을 생성합니다.

```
aws batch create-job-queue \
  --job-queue-name my-sm-training-fifo-jq \
  --job-queue-type SAGEMAKER_TRAINING \
  --priority 1 \
  --service-environment-order order=1,serviceEnvironment=ExampleServiceEnvironment
```

*ExampleServiceEnvironment*를 서비스 환경의 이름으로 바꿉니다.

이 명령은 다음과 비슷한 출력을 반환합니다.

```
{
  "jobQueueName": "my-sm-training-fifo-jq",
  "jobQueueArn": "arn:aws:batch:region:account:job-queue/my-sm-training-fifo-jq"
}
```

작업 대기열을 생성한 후 성공적으로 생성되었고 유효한 상태에 있는지 확인합니다.

`describe-job-queues` 명령을 사용하여 작업 대기열에 대한 세부 정보를 봅니다.

```
aws batch describe-job-queues --job-queues my-sm-training-fifo-jq
```

이 명령은 다음과 비슷한 출력을 반환합니다.

```
{
  "jobQueues": [
    {
```

```
      "jobQueueName": "my-sm-training-fifo-jq",
      "jobQueueArn": "arn:aws:batch:region:account:job-queue/my-sm-training-fifo-
jq",
      "state": "ENABLED",
      "status": "VALID",
      "statusReason": "JobQueue Healthy",
      "priority": 1,
      "computeEnvironmentOrder": [],
      "serviceEnvironmentOrder": [
        {
          "order": 1,
          "serviceEnvironment": "arn:aws:batch:region:account:service-
environment/ExampleServiceEnvironment"
        }
      ],
      "jobQueueType": "SAGEMAKER_TRAINING",
      "tags": {},
      "jobStateTimeLimitActions": []
    }
  ]
}
```

다음을 확인하세요.

- `state`는 `ENABLED`입니다.
- `status`는 `VALID`입니다.
- `statusReason`은 `JobQueue Healthy`입니다.
- `jobQueueType`은 `SAGEMAKER_TRAINING`입니다.
- `serviceEnvironmentOrder`는 서비스 환경을 참조합니다.

작업 대기열 템플릿

다음은 빈 작업 대기열 템플릿입니다. 이 템플릿을 사용하여 작업 대기열을 생성할 수 있습니다. 그런 다음 이 작업 대기열을 파일에 저장하고 AWS CLI `--cli-input-json` 옵션과 함께 사용할 수 있습니다. 이러한 파라미터에 대한 자세한 내용은 AWS Batch API 참조에서 CreateJobQueue를 참조하세요.

> **Note**
>
> 다음 AWS CLI 명령을 사용하여 위의 작업 대기열 템플릿을 생성할 수 있습니다.

```
$ aws batch create-job-queue --generate-cli-skeleton
```

```
{
   "computeEnvironmentOrder": [
      {
         "computeEnvironment": "",
         "order": 0
      }
   ],
   "jobQueueName": "",
   "jobStateTimeLimitActions": [
      {
         "state": "RUNNABLE",
         "action": "CANCEL",
   "maxTimeSeconds": 0,
         "reason": ""

      }
   ],
   "priority": 0,
   "schedulingPolicyArn": "",
   "state": "ENABLED",
   "tags": {
      "KeyName": ""
   }
}
```

에서 작업 대기열 보기 AWS Batch

작업 대기열을 생성하고 작업을 제출한 후에는 진행 상황을 모니터링할 수 있어야 합니다. 작업 세부 정보 페이지를 사용하여 작업 대기열을 검토하고, 관리하고, 모니터링할 수 있습니다.

작업 대기열 정보 보기

AWS Batch 콘솔에서 탐색 창에서 작업 대기열을 선택하고 원하는 작업 대기열을 선택하여 세부 정보를 봅니다. 이 페이지에서 작업 대기열을 검토 및 관리하고 작업 대기열 스냅샷, 작업 상태 제한, 환경 순서, 태그, 작업 대기열 JSON 코드 등과 같은 대기열 작업에 대한 추가 정보를 볼 수 있습니다.

작업 대기열 세부 정보

이 섹션에서는 작업 대기열에 대한 개요와 유지 관리 옵션을 제공합니다. 또한 이 섹션에서 Amazon 리소스 이름(ARN)을 찾을 수 있습니다.

를 통해이 정보를 찾으려면 작업 대기열 이름 또는 해당 ARN과 함께 `DescribeJobQueues` 작업을 AWS Command Line Interface사용합니다.

활성 공유

공정 공유 일정을 사용하는 작업 대기열의 경우는 서로 다른 공유 식별자가 용량을 소비하는 방식에 대한 가시성을 AWS Batch 제공합니다. 이 정보는 리소스 배포를 이해하고 조정이 필요할 수 있는 공유를 식별하는 데 도움이 됩니다.

> **Note**
>
> 활성 공유 탭은 작업 대기열의 예약 알고리즘이 공정 공유인 경우에만 표시됩니다.

상위 20개 활성 공유 섹션에는 예약, 시작 및 실행 중인 작업이 있는 공유 식별자가 표시됩니다. 이 보기에는 다음이 포함됩니다.

- 공유 식별자 이름 - 공유의 고유 식별자입니다.

  공유 식별자는 공정 공유 예약을 위해 작업을 그룹화하는 레이블입니다. 공유 식별자가 동일한 작업을 제출하면는 리소스 할당을 위해 해당 작업을 동일한 워크로드의 일부로 AWS Batch 처리합니다. 공유 식별자는 다양한 팀, 프로젝트 또는 워크로드 유형에 컴퓨팅 용량을 공평하게 분산하는 데 도움이 됩니다. 자세한 내용은 공유 식별자를 사용하여 워크로드 식별 단원을 참조하십시오.
- 용량 사용률 - 작업이 사용하도록 구성된 리소스의 양입니다. 이는 환경에 vCPU 따라 cpu, 또는 `instances`로 측정됩니다.
- 작업 보기 작업 - 해당 공유에 대한 모든 작업을 볼 수 있는 링크입니다.

이 정보는 목록 형식과 차트 형식으로 모두 볼 수 있습니다.

- 목록 보기 - 정확한 용량 번호가 포함된 테이블 형식 표시
- 차트 보기 - 상대적 사용률을 보여주는 시각적 막대 차트

작업 대기열 스냅샷

이 섹션에서는 대기열에 있는 처음 100개 RUNNABLE 작업의 정적 목록을 제공합니다. 검색 필드를 사용하면 결과 섹션의 모든 열에서 정보를 검색하여 목록의 범위를 좁힐 수 있습니다. 스냅샷 결과 영역의 작업은 작업 대기열의 실행 전략에 따라 정렬됩니다. 선입선출(FIFO) 작업 대기열의 경우 작업 순서는 제출 시간을 기준으로 합니다. 공정 공유 예약 작업 대기열의 경우 작업 순서는 작업 우선 순위 및 공유 사용량을 기준으로 합니다. 용량 필요 필드에는 작업이 사용하도록 구성된 리소스의 양이 표시됩니다.

결과는 작업 대기열의 스냅샷이므로 결과 목록이 자동으로 업데이트되지 않습니다. 목록을 업데이트하려면 섹션 상단에 있는 새로 고침을 선택합니다. 작업의 이름 하이퍼링크를 선택하여 작업 세부 정보로 이동하고 작업의 상태 및 기타 관련 정보를 확인합니다.

를 통해이 정보를 찾으려면 작업 대기열 이름 또는 해당 ARN과 함께 `GetJobQueueSnapshot` 작업을 AWS CLI사용합니다.

```
aws batch get-job-queue-snapshot --job-queue my-sm-training-fifo-jq
```

작업 상태 제한

이 탭을 사용하여 작업이 취소되기 전에 RUNNABLE 상태를 유지할 수 있는 시간과 관련된 구성 정보를 검토합니다.

를 통해이 정보를 찾으려면 작업 대기열 이름 또는 해당 ARN과 함께 `DescribeJobQueues` 작업을 AWS CLI사용합니다.

환경 순서

작업 대기열이 여러 환경에서 실행되는 경우 이 탭은 작업 대기열 순서와 개요를 제공합니다.

를 통해이 정보를 찾으려면 작업 대기열 이름 또는 해당 ARN과 함께 `DescribeJobQueues` 작업을 AWS CLI사용합니다.

Tags

이 탭을 사용하여 이 작업 대기열과 연결된 태그를 검토하고 관리합니다.

JSON

이 탭을 사용하여 이 작업 대기열과 연결된 JSON 코드를 복사합니다. 그런 다음 AWS CloudFormation 템플릿 및 스크립트에 JSON AWS CLI 을 재사용할 수 있습니다.

AWS Batch의 작업 대기열 삭제

작업 대기열이 더 이상 필요하지 않게 되면 작업 대기열을 비활성화하고 삭제할 수 있습니다.

Delete a job queue (AWS Batch console)

1. https://console.aws.amazon.com/batch/에서 AWS Batch 콘솔을 엽니다.
2. 탐색 창에서 작업 대기열을 선택한 다음 작업 대기열을 선택합니다.
3. 작업을 선택한 다음 비활성화를 선택합니다.
4. 작업 대기열의 상태가 비활성화가 되면 작업을 선택한 다음 삭제를 선택합니다.
5. 모달 창에서 작업 대기열 삭제를 선택합니다.

Delete a job queue (AWS CLI)

1. 작업 대기열을 비활성화하여 새 작업 제출을 방지합니다.

```
aws batch update-job-queue \
  --job-queue my-sm-training-fifo-jq \
  --state DISABLED
```

2. 실행 중인 작업이 완료될 때까지 기다린 다음 작업 대기열을 삭제합니다.

```
aws batch delete-job-queue \
  --job-queue my-sm-training-fifo-jq
```

공정 공유 예약 정책

AWS Batch 스케줄러는 작업 대기열에 제출된 작업을 실행하는 시기, 위치 및 방법을 평가합니다. 작업 대기열을 생성할 때 예약 정책을 지정하지 않으면 AWS Batch 작업 스케줄러는 기본적으로 선입선출(FIFO) 전략으로 설정됩니다. FIFO 전략을 사용하면 중요한 작업이 이전에 제출된 작업 뒤에 “묶여” 있을 수 있습니다. 다른 스케줄 정책을 지정하여 특정 요구 사항에 따른 컴퓨팅 리소스를 할당할 수 있습니다.

> **Note**
>
> 작업이 실행되는 특정 순서를 예약하려면 SubmitJob의 dependsOn 파라미터를 사용하여 각 작업의 종속성을 지정합니다.

예약 정책을 생성하여 작업 대기열에 연결하면 공정 공유 예약이 활성화됩니다. 작업 대기열에 예약 정책이 있으면 예약 정책이 작업 실행 순서를 결정합니다. 자세한 내용은 공정 공유 예약 정책을 사용하여 공유 식별자 할당 단원을 참조하십시오.

주제

- 공유 식별자를 사용하여 워크로드 식별
- 공정 공유 예약 정책을 사용하여 공유 식별자 할당
- 공정 공유 일정을 사용하여 작업 예약 지원
- 자습서: 공정 공유 예약 정책 생성
- 참조: 공정 공유 예약 정책 템플릿

공유 식별자를 사용하여 워크로드 식별

공유 식별자를 사용하여 작업에 태그를 지정하고 사용자와 워크로드를 구분할 수 있습니다. AWS Batch 스케줄러는 *(T * weightFactor)* 공식을 사용하여 각 공유 식별자의 사용량을 추적합니다. 여기서 *T*는 시간 경과에 따른 vCPU 사용량입니다. 스케줄러는 공유 식별자에서 사용량이 가장 적은 작업을 선택합니다. 공유 식별자는 재정의하지 않고도 사용할 수 있습니다.

> **Note**
>
> 공유 식별자는 작업 대기열 내에서만 유효하며 작업 대기열 간에 집계되지 않습니다.

공정 공유 예약 우선 순위를 설정하여 공유 식별자에서 작업이 실행되는 순서를 구성할 수 있습니다. 예약 우선 순위가 높은 작업이 먼저 예약됩니다. 공정 공유 예약 정책을 지정하지 않으면 작업 대기열에 제출된 모든 작업이 FIFO 순서로 예약됩니다. 작업을 제출할 때 공유 식별자나 공정 공유 예약 우선 순위를 지정할 수 없습니다.

> **Note**
>
> 연결된 컴퓨팅 리소스는 명시적으로 재정의되지 않는 한 모든 공유 식별자에게 동일하게 할당됩니다.

공정 공유 예약 정책을 사용하여 공유 식별자 할당

공정 공유 예약 정책을 사용하여 작업 대기열의 컴퓨팅 리소스가 사용자 또는 워크로드 간에 할당되는 방식을 구성할 수 있습니다. 공정 공유 예약 정책을 사용하면 워크로드 또는 사용자에게 서로 다른 공유 식별자를 할당할 수 있습니다.는 일정 기간 동안 사용 가능한 총 리소스의 백분율을 각 공유 식별자에 AWS Batch 할당합니다.

공정 공유율은 `shareDecaySeconds` 및 `shareDistribution` 값을 사용하여 계산됩니다. 사용자는 정책에 공유 소멸 시간을 할당하여 공정 공유 분석에 시간을 늘릴 수 있습니다. 시간을 추가하면 시간에 더 많은 가중치가 부여되고 지정된 가중치는 줄어듭니다. 사용자는 컴퓨팅 예약을 지정하여 활성 상태가 아닌 공유 식별자를 위해 컴퓨팅 리소스를 유보해 둘 수 있습니다. 자세한 내용은 SchedulingPolicyDetail을 참조하세요.

공정 공유 일정을 사용하여 작업 예약 지원

공정 공유 예약은 작업 예약에 도움이 되는 일련의 제어 기능을 제공합니다. 예약 정책 파라미터에 대한 자세한 내용은 SchedulingPolicyDetail을 참조하세요.

- 감소 초 공유 - AWS Batch 스케줄러가 각 공유 식별자에 대한 공정 공유 백분율을 계산하는 데 사용하는 기간(초)입니다. 값이 0이면 현재 사용량만 측정됨을 나타냅니다. 성능 저하 시간이 길수록 시간에 더 많은 가중치를 부여합니다.

  > **Note**
  >
  > 성능 저하 시간은 다음과 같이 계산됩니다. *`shareDecaySeconds + OrderMinutes`* 여기서 *`OrderMinutes`* 시간은 분 단위로 정렬됩니다.

- 컴퓨팅 예약 — 단일 공유 식별자의 작업이 작업 대기열에 연결된 모든 리소스를 사용하는 것을 방지합니다. 예약 비율은 *`(computeReservation/100)^ActiveFairShares`*입니다. 여기서 `ActiveFairShares`는 활성 공유 식별자 수입니다.

> **Note**
>
> 공유 식별자에 작업이 SUBMITTED, PENDING, RUNNABLE, STARTING 또는 RUNNING 상태이면 활성 공유 식별자로 간주됩니다. 성능 저하 기간이 만료되면 공유 식별자는 비활성 상태로 간주됩니다.

- 가중치 계수 – 공유 식별자의 가중치 계수입니다. 기본값은 1입니다. 값이 낮을수록 공유 식별자의 작업이 실행되도록 허용하거나 공유 식별자에 추가 런타임이 제공됩니다. 예를 들어, 가중치 0.125(1/8)로 공유 식별자를 사용하는 작업은 가중치 1로 공유 식별자를 사용하는 작업 보다 8배의 1 컴퓨팅 리소스를 할당 받습니다.

> **Note**
>
> 기본 가중치 계수인 1을 업데이트해야 하는 경우에만 이 속성을 정의하면 됩니다.

작업 대기열이 활성 상태이고 작업을 처리 중인 경우 작업 대기열 스냅샷을 통해 처음 100개의 RUNNABLE 작업 목록을 검토할 수 있습니다. 자세한 내용은 에서 작업 대기열 보기 AWS Batch 단원을 참조하십시오.

자습서: 공정 공유 예약 정책 생성

예약 정책을 사용하여 작업 대기열을 생성하려면 먼저 예약 정책을 생성해야 합니다. 공정 공유 예약 정책을 생성할 때는 하나 이상의 공유 식별자 또는 공유 식별자 접두사를 대기열의 가중치와 연결하고 선택적으로 정책에 저하 기간 및 컴퓨팅 예약을 할당합니다.

공정 공유 예약 정책을 생성하려면

1. https://console.aws.amazon.com/batch/ AWS Batch 콘솔을 엽니다.
2. 탐색 모음에서 사용할 리전을 선택합니다.
3. 탐색 창에서 예약 정책, 생성을 선택합니다.
4. 이름에 정책의 고유 이름을 입력합니다. 최대 128개의 문자(대문자 및 소문자), 숫자, 하이픈 및 밑줄이 허용됩니다.
5. (선택 사항) 성능 저하 초 공유에는 공정 공유 예약 정책의 공유 저하 시간을 정수 값으로 입력합니다. 공유 저하 시간이 길수록 작업을 예약할 때 상당히 더 많은 컴퓨팅 리소스가 사용됩니다. 이렇게 하면 해당 공유 식별자가 최근에 컴퓨팅 리소스를 사용한 적이 없다는 가정하에 공유 식별자를

사용하는 작업이 공유 식별자에 대한 가중치가 허용하는 것보다 더 많은 컴퓨팅 리소스를 일시적으로 사용할 수 있습니다.

6. (선택 사항) 컴퓨팅 예약에는 공정 공유 예약 정책의 컴퓨팅 예약에 대한 정수 값을 입력합니다. 컴퓨팅 예약은 현재 활성화되지 않은 공유 식별자에 사용할 수 있도록 일부 vCPU를 예비 보관합니다.

   예약 비율은 `(computeReservation/100)^ActiveFairShares`이며 ActiveFairShares는 활성화된 공유 식별자 수입니다.

   예를 들어 값이 `computeReservation` 50이면가 공유 식별자가 하나뿐인 경우 사용 가능한 최대 VCPU의 50%, 공유 식별자가 두 개 있는 경우 25%, 공유 식별자가 세 개 있는 경우 12.5%를 예약 AWS Batch 해야 함을 나타냅니다. `computeReservation` 값이 25인 경우 AWS Batch 는 공정 공유 식별자가 1개뿐이면 사용 가능한 최대 VCPU의 25%, 공유 식별자가 2개이면 6.25%, 공유 식별자가 3개이면 1.56%를 예약해야 함을 나타냅니다.

7. 속성 공유 섹션에서 각 공유 식별자의 공유 식별자 및 가중치를 지정하여 공정 공유 예약 정책과 연결할 수 있습니다.

   a. 공유 식별자 추가를 선택합니다.

   b. 공유 식별자에는 공유 식별자를 지정합니다. 문자열이 '*'로 끝나는 경우 이는 작업의 공유 식별자를 일치시키는 데 사용되는 공유 식별자 접두사가 됩니다. 예약 정책의 모든 공유 식별자와 공유 식별자 접두사는 고유해야 하며 중복되지 않아야 합니다. 예를 들어 동일한 공정 공유 일정 정책에 공유 식별자 접두사 'UserA*'와 공유 식별자 'UserA1'을 사용할 수 없습니다.

   c. 가중치 계수에는 공유 식별자의 상대적 가중치를 지정합니다. 기본값은 1.0입니다. 값이 낮을수록 컴퓨팅 리소스에 대한 우선순위가 높아집니다. 공유 식별자 접두사를 사용하는 경우 해당 접두사로 시작하는 공유 식별자가 있는 작업도 가중치 계수를 공유합니다. 이렇게 하면 해당 작업에 대한 가중치 계수가 효과적으로 증가하여 개별 우선 순위는 낮아지지만 공유 식별자 접두사의 가중치 계수는 동일하게 유지됩니다.

8. (선택 사항) 태그 섹션에 각 태그에 대한 키와 값을 지정하여 예약 정책과 연결합니다. 자세한 내용은 AWS Batch 리소스 태깅 단원을 참조하십시오.

9. 제출을 선택하여 작업을 완료하고 예약 정책을 생성합니다.

참조: 공정 공유 예약 정책 템플릿

빈 공정 공유 예약 정책 템플릿은 다음과 같습니다. 이 템플릿을 사용하여 예약 정책을 생성한 다음 파일에 저장하고 옵션과 함께 사용할 수 있습니다 AWS CLI `--cli-input-json`. 이 파라미터에 대한 자세한 내용을 알아보려면 AWS Batch API 참조의 CreateSchedulingPolicy 작업을 참조하세요.

> **Note**
>
> 다음 AWS CLI 명령을 사용하여 작업 대기열 템플릿을 생성할 수 있습니다.
>
> ```
> $ aws batch create-scheduling-policy --generate-cli-skeleton
> ```

```
{
    "name": "",
    "fairsharePolicy": {
        "shareDecaySeconds": 0,
        "computeReservation": 0,
        "shareDistribution": [
            {
                "shareIdentifier": "",
                "weightFactor": 0.0
            }
        ]
    },
    "tags": {
        "KeyName": ""
    }
}
```

리소스 인식 일정 예약

AWS Batch 는 작업 대기열(JQ)과 연결된 컴퓨팅 환경(CE)에서 vCPU, GPU 및 메모리 가용성을 기반으로 작업을 예약합니다. 그러나 때로는 작업이 다른 필수 리소스에 종속될 수 있으므로 성공적으로 실행된다고 보장할 수 없고 해당 작업이 취소되거나 종료될 수 있으므로 이러한 CE 리소스의 가용성만으로는 작업이 성공적으로 실행된다고 보장할 수 없습니다. 이는 컴퓨팅 리소스의 비효율적인 사용을 유발합니다. 이 문제를 해결하기 위해 리소스 인식 예약은 CE에서 작업을 실행하도록 예약하기 전에 종속된 비CE 리소스의 가용성을 확인할 수 있습니다.

AWS Batch resource-aware scheduling을 사용하면 타사 라이선스 토큰, 데이터베이스 액세스 대역폭, 타사 API에 대한 호출 제한 필요성 등 작업을 실행하는 데 필요한 소모적인 리소스를 기반으로 작업을 예약할 수 있습니다. 작업을 실행하는 데 필요한 소모성 리소스를 지정하면 Batch는 작업을 예약할 때 이러한 리소스 종속성을 고려합니다. 소모성 리소스 부족으로 인한 작업 실패와 긴 대기 시간을 제거하기 위한 수동 개입을 방지할 수 있습니다. 필요한 모든 리소스를 사용할 수 있는 작업만 할당하여 컴퓨팅 리소스의 사용 부족을 줄일 수 있습니다.

리소스 인식 예약은 FIFO 및 공정 공유 예약 정책 모두에 사용할 수 있으며 EKS, ECS 및 Fargate를 포함하여 Batch에서 지원하는 모든 컴퓨팅 플랫폼에서 사용할 수 있습니다. 배열 작업, 다중 노드 병렬(MNP) 작업 및 일반 배치 작업과 함께 사용할 수 있습니다.

리소스 인식 예약을 구성하려면 먼저 각 리소스의 사용 가능한 총 수와 함께 작업을 실행하는 데 필요한 모든 소모성 리소스를 지정합니다. 그런 다음 소모성 리소스가 필요한 각 작업에 대해, 필요한 각 리소스의 이름과 필요한 수량을 지정합니다. Batch는 작업 대기열의 작업에 사용할 수 있는 소모성 리소스 수를 추적하고 작업이 성공적으로 실행되는 데 필요한 모든 소모성 리소스를 사용할 수 있는 경우에만 작업이 실행되도록 예약합니다.

주제

소모성 리소스 생성

먼저 작업이 실행 중일 때 사용되고 제한된 수량으로만 사용할 수 있는 비CE 리소스를 나타내는 소모성 리소스를 생성해야 합니다. 각 소모성 리소스는 다음과 같은 요소를 가집니다.

- 계정 수준에서 고유해야 하는 리소스 이름(`consumableResourceName`).
- (선택 사항) 작업이 완료된 후 리소스를 재사용할 수 있는지 여부를 나타내는 리소스 유형(`resourceType`). 이는 다음 중 하나일 수 있습니다.
  - `REPLENISHABLE`(기본값)
  - `NON_REPLENISHABLE`
- 사용 가능한 소모성 리소스의 총량을 지정하는 총 수량(`totalQuantity`).

계정당 소모성 리소스의 최대 수는 5만 개입니다.

콘솔:

1. AWS Batch 콘솔의 왼쪽 탐색 패널에서 소모성 리소스를 선택합니다.

2. 소모성 리소스 생성을 선택합니다.
3. 고유한 리소스 이름, 총 리소스 수량을 입력하고 리소스 유형이 보충 가능인지 보충 불가인지 선택합니다.
4. 소모성 리소스 생성을 선택합니다.

API:

CreateConsumableResource API를 사용하여 원하는 리소스를 정의합니다.

작업을 실행하는 데 필요한 리소스 지정

작업을 등록할 때 생성한 하나 이상의 리소스 이름(consumableResource)과 작업의 각 인스턴스에 필요한 해당 리소스의 수량(quantity)을 지정할 수 있습니다.

Batch는 주어진 시점에 각 리소스의 사용 가능한 유닛을 추적합니다. 작업 대기열의 각 작업에 대해 Batch 스케줄러는 지정된 리소스 종속성을 사용할 수 있는 경우에만 작업이 실행되도록 합니다.

작업이 대기열의 앞부분에 도달할 때 작업을 위한 소모성 리소스를 사용할 수 없는 경우, 작업은 필요한 모든 리소스를 사용할 수 있게 되거나 작업 상태 시간 제한에 도달할 때까지 RUNNABLE 상태로 대기합니다(에서 작업 대기열 보기 AWS Batch 참조). Batch가 모든 리소스를 사용할 수 있는지 확인하면 작업이 STARTING 상태로 전환되었다가 RUNNING으로 전환됩니다. 작업이 일단 STARTING으로 이동하면 리소스가 잠기고 작업이 SUCCEEDED 또는 FAILED로 이동하면 잠금 해제됩니다.

작업을 제출할 때 특정 작업에 필요한 리소스의 수량도 업데이트할 수 있습니다.

콘솔:

작업을 정의할 때 리소스와 필요한 수량을 지정하려면:

1. AWS Batch 콘솔에서 작업 정의 마법사를 사용하여 작업을 정의합니다(작업 정의 -> 생성).
2. 마법사의 4단계: 컨테이너 구성의 소모성 리소스 아래의 목록에서 필요한 리소스의 이름을 선택합니다. 요청된 값 필드에 이 작업의 인스턴스에 필요한 이 리소스의 수량을 입력한 다음 소모성 리소스 추가를 선택합니다.
3. 작업에 필요한 모든 소모성 리소스에 대해 이전 단계를 반복합니다. 정의한 각 작업에 대해 최대 5개의 리소스를 지정할 수 있습니다.
4. 작업 정의 마법사를 완료한 후 작업 정의 생성을 선택하기 전에 생성한 소모성 리소스의 목록이 표시됩니다.

작업을 제출할 때 필요한 리소스 양을 업데이트하려면:

1. AWS Batch 콘솔의 왼쪽 탐색 창에서 작업을 선택한 다음 새 작업 제출을 선택합니다.
2. 마법사의 2단계: 재정의 구성의 소모성 리소스 재정의에서 작업에 필요한 수량을 재정의할 모든 소모성 리소스에 대해 새 요청된 값을 입력합니다.
3. 이 작업에 대해 수행할 모든 재정의를 완료한 후 다음을 선택하여 검토 및 제출을 계속합니다.

API:

RegisterJobDefinition API에 작업을 등록할 때 요청의 `consumableResourceProperties` 부분에 `consumableResourceList`를 사용하여 작업의 인스턴스를 실행하는 데 필요한 소모성 리소스와 각 리소스의 수량을 지정합니다.

SubmitJob API를 사용하여 작업을 제출할 때 요청의 `consumableResourcePropertiesOverride` 부분을 사용하여 소모성 리소스 목록과 각 리소스의 수량을 재정의할 수 있습니다. 이렇게 하면 작업의 각 인스턴스에 필요한 리소스 수량만 재정의되고 사용 가능한 총 수량은 재정의되지 않습니다.

사용 중이고 사용 가능한 리소스 수 확인

Batch를 사용하면 주어진 시점에 사용 가능한 리소스 수(`availableQuantity`), 사용 중인 리소스 수(`inUseQuantity`) 및 총 리소스(`totalQuantity`)를 쿼리할 수 있습니다.

작업이 `STARTING` 상태가 되면 소모된 리소스는 해당 리소스의 사용 가능한 수량에서 차감됩니다. 리소스가 `REPLENISHABLE`인 경우, 작업이 SUCCEEDED 또는 FAILED 상태로 전환되는 즉시 소비된 리소스 수가 사용 가능한 수량에 다시 추가되고 총 수량은 동일하게 유지됩니다. 리소스가 `NON_REPLENISHABLE`인 경우, 소모된 리소스 수는 총 수량 및 사용 가능한 수량 모두에서 차감되며 작업이 `SUCCEEDED` 또는 `FAILED` 상태로 이동하는지 여부에 관계없이 다시 추가되지 않습니다.

> **Note**
>
> 이 정보는 최대 30초까지 지연될 수 있습니다.

콘솔:

1. AWS Batch 콘솔의 왼쪽 탐색 패널에서 소모성 리소스를 선택합니다.
2. 보충 가능 또는 보충 불가능 탭을 선택하여 생성한 해당 유형의 리소스를 봅니다.

3. 각 보충 가능 리소스에 대해 콘솔에는 이름, 리소스의 총 수량, 현재 사용 중인 수 및 아직 사용 가능한 수가 사용률 계산(사용 중인 리소스 수를 해당 리소스의 총 수량으로 나눈 값)과 함께 표시됩니다.

   각 보충 불가 리소스에 대해 콘솔에는 이름, 현재 사용 중인 수 및 여전히 사용 가능한 수가 표시됩니다.

콘솔의 작업 세부 정보 페이지에서 사용 가능한 리소스에 대한 현재 정보를 볼 수도 있습니다.

1. AWS Batch 콘솔의 왼쪽 탐색 패널에서 작업을 선택한 다음 작업 이름을 선택하여 해당 작업에 대한 세부 정보 페이지를 엽니다.
2. 보충 가능 리소스와 보충 불가 리소스 모두에 대한 정보는 작업에 리소스가 필요한 경우 확인할 수 있습니다. 두 유형 모두에 대해 콘솔에는 이름, 작업에 요청된 수량, 여전히 사용 가능한 수량, 현재 사용 중인 수량, 총 리소스 수량이 현재 사용률 계산(작업에서 사용 중인 리소스 수를 해당 리소스의 총 수량으로 나눈 값)과 함께 표시됩니다.

API:

다음 정보를 반환하는 DescribeConsumableResource API를 사용합니다.

```
{
   "availableQuantity": number,
   "consumableResourceArn": "string",
   "consumableResourceName": "string",
   "createdAt": number,
   "inUseQuantity": number,
   "resourceType": "string",
   "tags": {
      "string" : "string"
   },
   "totalQuantity": number
}
```

또한 ListConsumableResources API는 사용 중인 리소스 수(`inUseQuantity`)와 현재 사용 가능한 총 리소스 수(`totalQuantity`)를 계정에서 생성한 모든 소모성 리소스 목록의 일부로 보고합니다. 또한 이 API를 사용하면 소모성 리소스 이름을 기반으로 소모성 리소스 목록 쿼리를 필터링할 수 있습니다.

작업에서 사용 중인 리소스의 수량 업데이트

리소스의 총 수량을 새 값으로 재설정하거나, 총 수량에 추가하거나, 리소스에서 차감할 수 있습니다.

지정한 새 총 수량이 이전보다 큰 경우 Batch는 적절히 더 많은 작업을 예약합니다. 새 총 수량이 이전보다 작고 사용 중인 이 리소스의 유닛이 없는 경우 Batch는 총(또는 사용 가능한) 수량을 줄입니다. 사용 중인 유닛이 있는 경우, Batch는 사용 가능한 수량을 즉시 줄이고 작업이 완료되면 총(사용 가능한) 수량을 줄여 결국 새 수량에 도달합니다.

콘솔:

1. AWS Batch 콘솔의 왼쪽 탐색 패널에서 소모성 리소스를 선택합니다.
2. 보충 가능 또는 보충 불가능 탭을 선택하여 생성한 해당 유형의 리소스를 봅니다.
3. 보충 불가한 리소스의 경우:

    1. 업데이트할 리소스를 선택한 다음 작업을 선택하고 리소스 설정, 리소스 추가 또는 리소스 제거를 선택합니다.
    2. 이전 단계에서 선택한 작업에 따라 총 값을 설정하거나, 리소스를 추가하거나 리소스를 제거할 수 있는 팝업 창이 나타납니다. 새 총 값으로 설정할 수량, 총 수량에 추가할 수량 또는 총 수량에서 차감할 수량을 입력한 다음 확인을 선택합니다.

    보충 불가한 리소스의 경우:

    1. 업데이트할 리소스를 선택한 다음 작업을 선택하고 리소스 설정, 리소스 추가 또는 리소스 제거를 선택합니다.
    2. 이전 단계에서 선택한 작업에 따라 사용 가능 값을 설정하거나, 리소스를 추가하거나 리소스를 제거할 수 있는 팝업 창이 나타납니다. 새 사용 가능 값으로 설정할 수량, 사용 가능 수량에 추가할 수량 또는 사용 가능 수량에서 차감할 수량을 입력한 다음 확인을 선택합니다.

API:

`UpdateConsumableResource` API를 사용하여 리소스에 대한 새 총 수량을 설정하거나 총 수량을 늘리거나 줄입니다.

특정 소모성 리소스가 필요한 작업 찾기

Batch를 사용하여 특정 소모성 리소스가 필요한 작업 목록을 검색할 수 있습니다.

콘솔:

1. AWS Batch 콘솔의 왼쪽 탐색 패널에서 소모성 리소스를 선택합니다.
2. 목록에서 소모성 리소스의 이름을 선택합니다. 해당 리소스에 대한 세부 정보 페이지가 열립니다.
3. 작업 검색 아래에 작업 목록에 적용할 필터를 입력합니다. 작업 이름('같음', '다음으로 시작'), 날짜 범위(작업이 생성된 날짜) 및 추가 기준('작업 대기열', '작업 정의', '공유 작업 식별자')으로 필터링할 수 있습니다. 적용할 각 필터 유형에 대해 드롭다운 목록의 사용 가능한 옵션 중에서 선택하고 요청된 추가 정보를 입력합니다.

   검색을 선택합니다.
4. 작업 이름, 상태, 소모성 리소스의 요청된 유닛 수, 기타 필요한 소모성 리소스 등을 포함하여 소모성 리소스가 필요한 작업의 (필터링된) 목록이 표시됩니다. 이 목록을 사용하여 취소하거나 종료할 작업을 하나 이상 선택할 수 있습니다. 작업 이름을 선택하여 해당 작업의 세부 정보 페이지를 열 수도 있습니다.
5. 이제 작업 검색에서 결과를 새로 고침하거나 검색을 지우고 다시 시작할 수 있습니다.

API:

`ListJobsByConsumableResource` API에서 특정 소모성 리소스를 사용하는 작업의 목록을 가져올 수 있습니다. 또한 이 API를 사용하면 작업 상태 또는 작업 이름을 사용하여 작업 목록 쿼리를 필터링할 수 있습니다.

소모성 리소스 삭제

리소스가 필요한 작업이 여전히 실행 중인 경우에도 언제든지 소모성 리소스를 삭제할 수 있습니다. 소모성 리소스가 삭제되면 삭제 명령이 수신된 시간과 작업 스케줄러가 삭제를 준수하는 시간 사이에 간격이 있을 수 있으므로 리소스를 사용하는 작업이 삭제 호출 직후 예약될 수 있습니다. 삭제된 소모성 리소스에 리소스 유형(`resourceType`) `REPLENISHABLE`이 있는 경우 이는 작업이 완료되면 무시됩니다. 소모성 리소스를 삭제하고 동일한 이름으로 다시 생성하는 경우 동일한 리소스로 간주되며 `RUNNABLE` 작업에서 사용할 수 있습니다.

콘솔:

1. AWS Batch 콘솔의 왼쪽 탐색 패널에서 소모성 리소스를 선택합니다.
2. 보충 가능 또는 보충 불가능 탭을 선택하여 생성한 해당 유형의 리소스를 봅니다.
3. 삭제하려는 각 리소스를 선택한 다음 삭제를 선택합니다. 소모성 리소스 삭제 팝업 창이 나타납니다. 삭제를 확인하려면 삭제를 선택합니다.

콘솔의 세부 정보 페이지에서 소모성 리소스를 삭제할 수도 있습니다.

1. AWS Batch 콘솔의 왼쪽 탐색 패널에서 소모성 리소스를 선택합니다.
2. 보충 가능 또는 보충 불가능 탭을 선택하여 생성한 해당 유형의 리소스를 봅니다.
3. 삭제할 리소스의 이름을 선택합니다. 소모성 리소스의 세부 정보 페이지가 나타납니다. 삭제를 선택합니다. 소모성 리소스 삭제 팝업 창이 나타납니다. 삭제를 확인하려면 삭제를 선택합니다.

API:

`DeleteConsumableResource` API를 사용하여 소모성 리소스를 삭제합니다.

할당량 관리

AWS Batch 는 작업에 필요한 리소스와 연결된 서비스 환경(SE)에서 사용 가능한 용량을 기반으로 작업을 예약하여 컴퓨팅 리소스의 활용도를 높입니다. 또한 관리자는 할당량 관리를 통해 팀 또는 프로젝트가 더 세분화된 리소스 할당을 통해 사용할 수 있는 리소스 수를 제어할 수 있습니다.

할당량 관리를 통해 관리자는 컴퓨팅 할당량(용량 제한) 및 유휴 컴퓨팅을 위한 리소스 공유 전략을 포함하는 리소스와 같은 관련 할당량 공유를 정의하여 팀과 프로젝트 간에 공유 컴퓨팅 리소스를 효율적으로 할당할 수 있습니다. AWS Batch 각 할당량 공유는 연결된 작업 대기열 내에 중첩된 가상 대기열로 작동합니다. 작업 대기열에 대한 작업을 예약할 때 AWS Batch 는 연결된 모든 할당량 공유를 반복합니다.

관리자는 할당량 공유 간에 리소스 공유를 자신 있게 활성화할 수 있습니다. 선점은 모든 할당량 공유가 필요할 때 다른 사람에게 임대한 리소스를 회수할 수 있도록 허용하기 때문입니다. 할당량 공유 내에서 우선 순위가 높은 작업에 대해 실행 중인 작업을 선점할지 아니면 실행 중인 작업을 완료할지 선택할 수 있습니다. 작업 우선 순위는 제출 시 설정하고 나중에 업데이트할 수 있습니다. 선점 결정이 내려지면 업데이트된 우선 순위가 고려됩니다. 용량 사용률은 대기열, 할당량 공유 및 작업 수준 세부 수준에서 모니터링할 수 있습니다.

할당량 관리는 `SAGEMAKER_TRAINING` 서비스 환경에 연결된 작업 대기열에서만 지원됩니다.

주제

- 할당량 공유
- 선점
- 할당량 관리 리소스 생성
- 할당량 공유 생성
- 할당량 공유에 작업 제출

할당량 공유

할당량 공유는 작업 대기열 아래에 중첩되는 가상 대기열입니다. 최대 20개의를 모든 작업 대기열에 연결할 수 있습니다. 할당량 공유를 사용하면 용량 제한을 사용하여 팀 또는 프로젝트에 컴퓨팅 할당량을 할당할 수 있습니다. 할당량 공유는 하나 이상의 용량 제한을 제공해야 하며 최대 5개의 용량 제한을 지원합니다. 각 용량 제한은 지원되는 SageMaker 훈련 작업 인스턴스 유형에 대한 인스턴스 제한으로 표시되어야 합니다.

할당량 공유 리소스 공유 전략

할당량 공유에는 명시적 리소스 공유 구성도 있습니다.

- 할당량 공유의 유휴 컴퓨팅을 해당 작업에 대해서만 예약해야 하는 경우를 선택합니다`RESERVE`.
- 할당량 공유의 유휴 컴퓨팅을 다른 할당량 공유에 임대할 수 있는 경우를 선택합니다`LEND`.
- 할당량 공유의 유휴 컴퓨팅을 다른 할당량 공유에 임대할 수 있고 할당량 공유의 작업이 유휴 컴퓨팅을 임대하도록 허용해야 하는 경우 구성된 임대 한도`LEND_AND_BORROW`로를 선택합니다.

선점으로 용량 복원

AWS Batch 는 교차 공유 선점 작업을 수행하여 작업이 도착할 때 빌린 용량을 할당량 공유로 복원합니다. 선점할 작업을 선택할 때는 작업이 빌리는 인스턴스의 수와 유형, 작업의 상대적 우선순위 및 작업 기간을 AWS Batch 고려하고 사용자 지정 휴리스틱을 적용합니다.

할당량 공유의 용량 제한을 낮추는 관리자는 해당 할당량 공유에서 이미 소비`RESERVE`한 용량이 감소된 용량 제한을 초과하는 경우 리소스 공유 전략으로 `LEND` 또는를 선택했더라도 해당 할당량 내 , `SCHEDULED` `STARTING`또는 `RUNNING` 작업을 선점할 수 있습니다.

선점

할당량 관리는 선점을 통합하는 유일한 AWS Batch 예약 알고리즘으로,는 `SCHEDULED`, `STARTING`또는 `RUNNING` 작업을 AWS Batch 중지하여 `RUNNABLE` 작업 용량을 생성합니다.

교차 공유 선점

할당량 관리는 교차 공유 선점을 사용하여 작업이 도착할 때 빌린 용량을 할당량 공유로 복원합니다.

할당량 공유의 용량 제한을 낮추는 관리자는 해당 할당량 공유가 소비하는 용량이 이제 구성된 용량 제한을 초과하는 경우 해당 할당량 내의 , `SCHEDULED` `STARTING`또는 `RUNNING` 작업을 선점 대상으로 지정할 수도 있습니다.

공유 중 선점

할당량 공유는 공유 중 선점 사항을 활성화하도록 구성할 수 있습니다. 이렇게 하면 우선 순위가 높은 RUNNABLE 작업이 SCHEDULED, STARTING또는에 입력된 동일한 할당량 공유 내에서 우선 순위가 낮은 작업의 선점을 트리거할 수 있습니다RUNNING.

선점 선택 알고리즘

선점할 작업을 선택할 때 작업이 빌리는 인스턴스의 수와 유형, 작업의 상대적 우선순위 및 작업 기간을 AWS Batch 고려하고 사용자 지정 휴리스틱을 적용합니다. UpdateServiceJob API 호출을 사용하여 제출 후 작업schedulingPriority의를 업데이트할 수 있습니다. 이는 RUNNING 작업의 우선순위를 낮추거나(선점 가능성 증가) 공유 선점 기능이 활성화된 할당량 공유에서 RUNNABLE 작업의 우선순위를 높여 작업이 이미 실행 중인 작업을 선점할 수 있도록 하는 데 유용할 수 있습니다.

선점 재시도

기본값은 선점된 작업을 제한 RUNNABLE 없이 다시 대기열에 추가하기 위한 것입니다. 작업 경험의 선점 횟수를 제한하려면 작업 제출 preemptionRetriesBeforeTermination 시를 설정합니다. preemptionRetriesBeforeTermination가 0으로 설정되면 작업은 첫 번째 선점 FAILED 시 로 이동합니다.

최근 선점 시도의 슬라이딩 윈도우는 작업에 저장되며 DescribeServiceJob을 통해 볼 수 있습니다.

할당량 관리 리소스 생성

할당량 관리는 연결된 예약 정책, 서비스 환경 및 작업 대기열을 생성할 때 특정 설정이 필요합니다.

사전 조건

할당량 관리 리소스를 생성하기 전에 다음을 수행해야 합니다.

- IAM 권한 - 작업 대기열, 예약 정책 및 서비스 환경을 생성하고 관리할 AWS Batch 수 있는 권한입니다. 자세한 내용은 AWS Batch IAM 정책, 역할 및 권한 단원을 참조하십시오.

Configure quota management resources (AWS Batch console)

AWS Batch 콘솔은 할당량 관리에 필요한 모든 리소스를 생성하기 위한 통합 워크플로를 제공합니다. 할당량 관리 작업 대기열 생성 워크플로는 할당량 관리 지원 예약 정책 및 서비스 환경도 생성합니다.

1. https://console.aws.amazon.com/batch/ AWS Batch 콘솔을 엽니다.
2. 탐색 창에서 작업 대기열을 선택한 다음 생성을 선택합니다.
3. 오케스트레이션 유형에서 SageMaker 훈련을 선택합니다.
4. 작업 대기열 구성에서 다음을 수행합니다.

   a. 이름에 작업 대기열의 이름을 입력합니다.

   b. Priority에 0에서 1000 사이의 값을 입력합니다. 우선 순위가 높은 작업 대기열에는 서비스 환경에 대한 기본 설정이 지정됩니다.

5. 예약의 경우:

   a. 예약 알고리즘에서 할당량 관리를 선택합니다.

   b. 예약 정책 ARN의 경우:

      - 할당량 관리를 지정하는 예약 정책이 이미 있는 경우 드롭다운에서 선택합니다.
      - 그렇지 않으면 예약 정책 생성을 선택합니다.

        i. 사이드바가 열리고 할당량 관리 예약 정책을 구성합니다.

        ii. 예약 정책의 이름을 입력합니다.

        iii. 생성(Create)을 선택합니다. 이제 예약 정책 ARN 필드가 채워집니다.

6. 서비스 환경 구성의 경우 연결된 서비스 환경에서 다음을 수행합니다.

   > **Note**
   >
   > 할당량 관리가 활성화된 서비스 환경은 단일 할당량 관리가 활성화된 작업 대기열에만 연결할 수 있습니다.

   a. 할당량 관리와 호환되고 아직 할당량 관리가 활성화된 작업 대기열에 연결되지 않은 서비스 환경이 이미 생성된 경우 드롭다운에서 선택합니다.

   b. 그렇지 않으면 서비스 환경 생성을 선택합니다. 사이드바가 열리고 서비스 환경을 구성합니다.

      i. 서비스 환경의 이름을 입력합니다.

      ii. 하나 이상의 용량 제한(및 최대 5개)을 제공합니다. 각 용량 제한에 대해 드롭다운에서 인스턴스 유형과 최대 인스턴스 수를 선택합니다.

7. (선택 사항) 작업 상태 제한의 경우:

a. 구성 오류에서 둘 중 하나를 선택하고 최대 실행 가능 시간(초)을 SERVICE_ENVIRONMENT_MAX_RESOURCE 입력합니다.

b. 용량에서 INSUFFICIENT_INSTANCE_CAPACITY를 선택하고 최대 실행 가능 시간(초)을 입력합니다.

8. 작업 대기열 생성을 선택합니다.

Configure quota management resources (AWS CLI)

AWS CLI를 통해 할당량 관리를 구성하려면 예약 정책, 서비스 환경 및 작업 대기열을 생성합니다. 예약 정책과 서비스 환경 모두 할당량 관리와 호환되어야 하며 작업 대기열을 생성하기 전에 생성되어야 합니다.

예약 정책 생성

`create-scheduling-policy` 명령을 사용하여 할당량 관리 호환 예약 정책을 생성합니다. 생성 중에 할당량 공유 정책을 제공합니다.

```
aws batch create-scheduling-policy \
  --name my-qm-sagemaker-scheduling-policy \
  --quota-share-policy idleResourceAssignmentStrategy="FIFO"
```

예약 정책이 성공적으로 생성되었는지 확인합니다.

```
aws batch describe-scheduling-policies \
  --arns arn-for-my-qm-sagemaker-scheduling-policy
```

서비스 환경 생성

`create-service-environment` 명령을 사용하여 할당량 관리가 활성화된 서비스 환경을 생성합니다. 용량 제한이 `ml.g6.xlarge` 또는와 같이 SageMaker 훈련 작업에서 허용하는 인스턴스 유형을 사용하는지 확인합니다`ml.p4d.24xlarge`.

```
aws batch create-service-environment \
  --service-environment-name my-qm-sagemaker-service-env \
  --service-environment-type SAGEMAKER_TRAINING \
  --capacity-limits capacityUnit=instance_type,maxCapacity=instance_count
```

서비스 환경이 성공적으로 생성되었는지 확인합니다.

```
aws batch describe-service-environments \
  --service-environments my-qm-sagemaker-service-env
```

작업 대기열 생성

create-job-queue 명령을 사용하여 할당량 관리가 활성화된 작업 대기열을 생성합니다. 단, 다음 기준을 만족해야 합니다.

- 현재 다른 작업 대기열에 연결되지 않은 단일 SAGEMAKER_TRAINING 서비스 환경을 제공해야 합니다.
- 서비스 환경은 ml.m6i.xlarge가 아닌와 같은 인스턴스 유형 측면에서 용량 제한을 표시해야 합니다NUM_INSTANCES.
- 가 포함된 예약 정책을 연결해야 합니다quotaSharePolicy.
- jobQueueType이 SAGEMAKER_TRAINING이어야 합니다.

```
aws batch create-job-queue \
  --job-queue-name my-qm-sagemaker-jq \
  --job-queue-type SAGEMAKER_TRAINING \
  --priority 1 \
  --service-environment-order order=1,serviceEnvironment=my-qm-sagemaker-service-env
 \
  --scheduling-policy-arn arn-for-my-qm-sagemaker-scheduling-policy
```

작업 대기열이 성공적으로 생성되었는지 확인합니다.

```
aws batch describe-job-queues \
  --job-queues my-qm-sagemaker-jq
```

다음을 확인하세요.

- state는 ENABLED입니다.
- status는 VALID입니다.
- statusReason은 JobQueue Healthy입니다.

할당량 공유 생성

할당량은 스케줄러가 예약할 때 반복하는 연결된 작업 대기열 내의 가상 대기열로 함수를 AWS Batch 공유합니다. 이를 통해 관리자는 리소스 공유 전략에 대한 명시적 구성을 `capacity limits` 사용하여를 통해 팀 또는 프로젝트에 컴퓨팅 할당량을 할당할 수 있습니다.

사전 조건

할당량 공유를 생성하기 전에 다음을 확인해야 합니다.

- 할당량 관리 예약 정책, 서비스 환경 및 작업 대기열 - 할당량 관리 예약 정책, 서비스 환경 및 할당량 관리가 활성화된 작업 대기열입니다. 자세한 내용은 할당량 관리 리소스 생성 단원을 참조하십시오.
- IAM 권한 - 할당량 공유를 생성하고 관리할 AWS Batch 수 있는 권한입니다. 자세한 내용은 AWS Batch IAM 정책, 역할 및 권한 단원을 참조하십시오.

Configure quota shares (AWS Batch console)

1. https://console.aws.amazon.com/batch/ AWS Batch 콘솔을 엽니다.
2. 탐색 창의 작업 대기열을 선택하고 목록에서 할당량 관리가 활성화된 작업 대기열을 선택합니다. 작업 대기열 이름 링크를 선택합니다.
3. 할당량 공유 섹션에서 할당량 공유 생성을 선택합니다.
4. 할당량 공유의 이름을 입력합니다.
5. 용량 제한에서 용량 제한 추가를 선택합니다. 용량 제한을 하나 이상 지정해야 합니다.

   a. 드롭다운에서 인스턴스 유형을 선택하고이 할당량 공유가 할당한 최대 인스턴스를 설정합니다.

   b. (선택 사항) 용량 제한 추가를 선택하고 반복하여 최대 5개의 용량 제한을 연결합니다.
6. 용량 공유에서이 할당량 공유가 동일한 작업 대기열의 다른 할당량 공유와 용량을 공유하는 방법을 선택합니다.

   - 할당량 공유가 유휴 컴퓨팅을 대여하거나 빌려서는 안 되는 경우 예약을 선택합니다.
   - 할당량 공유가 유휴 컴퓨팅을 다른 할당량 공유에 부여할 수 있는 경우 대출을 선택합니다.
   - 할당량 공유가 유휴 컴퓨팅을 대여 및 대여할 수 있는 경우 대출 및 대여를 선택하고, 작업이 도착하면 교차 공유 선점을 통해 임대 컴퓨팅을 회수합니다.
7. (선택 사항) 공유 중 선점에서 공유 중 선점 활성화 또는 비활성화 여부를 선택합니다. 공유 중 선점 기능을 활성화하면 우선 순위가 높은 작업이 이미 `SCHEDULED`, `STARTING`또는 `RUNNING`

상태인 우선 순위가 낮은 작업을 선점할 수 있습니다. 공유 중 선점을 비활성화하면 우선 순위가 높은 작업이 용량이 사용 가능해질 때까지 대기합니다.

8. 할당량 공유 생성을 선택합니다.

Configure quota shares (AWS CLI)

`create-quota-share` 명령을 사용하여 할당량 공유를 생성합니다. 리소스 공유 전략과 공유 내 선점 활성화 여부를 선택해야 합니다.

대출 및 대여 예제

다음 예제에서는 유휴 용량을 대여하고 대여할 수 있는 할당량 공유를 생성하며, 대여 한도는 구성된 용량 한도의 100%입니다. 또한 공유 중 선점도 활성화하므로 우선 순위가 높은 작업은 SageMaker AI 내에서 예약된 우선 순위가 낮은 작업이 완료될 때까지 기다리지 않습니다.

```
aws batch create-quota-share \
  --quota-share-name lend_and_borrow_qs \
  --job-queue my-qm-sagemaker-jq \
  --capacity-limits maxCapacity=5,capacityUnit=ml.m6i.large \
  --resource-sharing-configuration strategy=LEND_AND_BORROW,borrowLimit=100 \
  --preemption-configuration inSharePreemption=ENABLED
```

대출 전용 예제

할당량 공유는 유휴 용량만 대여하고 직접 빌리지 않도록 구성할 수 있습니다. 다음 예제는 공유 내 선점 비활성화LEND가 있는 페어입니다.

```
aws batch create-quota-share \
  --quota-share-name lend_qs \
  --job-queue my-qm-sagemaker-jq \
  --capacity-limits maxCapacity=8,capacityUnit=ml.m6i.large \
  --resource-sharing-configuration strategy=LEND \
  --preemption-configuration inSharePreemption=DISABLED
```

예약 예제

유휴 용량을 예약하도록 할당량 공유를 구성할 수도 있습니다. 할당량 공유에 유휴 용량이 있는 경우 새로 제출된 작업은 더 빨리 시작될 수 있지만 할당량 공유에 작업이 없는 경우 전체 대기열 사용률이 낮아집니다.

```
aws batch create-quota-share \
  --quota-share-name reserved_qs \
  --job-queue my-qm-sagemaker-jq \
  --capacity-limits maxCapacity=2,capacityUnit=ml.m6i.large \
  --resource-sharing-configuration strategy=RESERVE \
  --preemption-configuration inSharePreemption=DISABLED
```

할당량 공유에 작업 제출

할당량 관리 작업 대기열에서는 모든 작업이 작업 제출 시 할당량 공유를 지정해야 합니다. 할당량 공유에 작업을 제출하려면 SubmitServiceJobquotaShareName에서를 지정합니다. 선택적으로를 제공하여 작업 시도가에 들어가기 전에 선점 시도 횟수를 제한할 preemptionConfiguration 수 있습니다FAILED. 작업 경험의 선점 횟수를 제한하려면 작업 제출 시 ServiceJobPreemptionConfiguration preemptionRetriesBeforeTermination 내에서를 설정합니다.

사전 조건

할당량 공유에 작업을 제출하기 전에 다음을 갖추어야 합니다.

- 할당량 관리 리소스 - 할당량 관리를 위해 구성된 예약 정책, 서비스 환경 및 작업 대기열입니다. 자세한 내용은 할당량 관리 리소스 생성 단원을 참조하십시오.
- 할당량 공유 - 작업 대기열에 하나 이상의 할당량 공유가 생성되었습니다. 자세한 내용은 할당량 공유 생성 단원을 참조하십시오.
- IAM 권한 - 작업을 제출할 수 있는 권한입니다 AWS Batch. 자세한 내용은 AWS Batch IAM 정책, 역할 및 권한 단원을 참조하십시오.

할당량 공유에 서비스 작업 제출

아래 표는 SageMaker Python SDK 또는 AWS CLI를 사용하여 할당량 공유에 서비스 작업을 제출하는 방법을 보여줍니다.

Submit using the SageMaker Python SDK

SageMaker Python SDK는 할당량 관리가 활성화된 작업 대기열에 작업을 제출할 수 있도록 기본적으로 지원됩니다. 다음 예제에서는 모델 트레이너를 생성하고, 훈련 대기열을 생성하고, 할당량 공유에 작업을 제출하는 방법을 보여줍니다. 전체 예제는 GitHub의 전체 샘플 노트북을 참조하세요.

훈련 작업 구성을 ModelTrainer 정의하는를 생성합니다.

```
from sagemaker.train.model_trainer import ModelTrainer
from sagemaker.train.configs import SourceCode, Compute, StoppingCondition

source_code = SourceCode(command="echo 'Hello World'")

model_trainer = ModelTrainer(
    training_image="123456789012.dkr.ecr.us-east-1.amazonaws.com/pytorch-
training:2.5-gpu-py311",
    source_code=source_code,
    base_job_name="my-training-job",
    compute=Compute(instance_type="ml.g5.xlarge", instance_count=1),
    stopping_condition=StoppingCondition(max_runtime_in_seconds=300),
)
```

할당량 관리가 활성화된 작업 대기열을 이름으로 참조하는 TrainingQueue 객체를 생성합니다.

```
from sagemaker.train.aws_batch.training_queue import TrainingQueue

queue = TrainingQueue("my-sagemaker-job-queue")
```

를 호출queue.submit하고를 지정하여 할당량 공유에 작업을 제출합니다quota_share_name. 할당량 공유 내에서 작업 순서에 영향을 미치priority려면를 설정해야 합니다. 실제 환경에서ModelTrainer는 훈련할 데이터가 inputs 있어야 합니다.

```
job = queue.submit(
    job_name="my-training-job",
    training_job=model_trainer,
    quota_share_name="my_quota_share",
    priority=3,
    inputs=None,
)
```

Submit using the AWS CLI

다음 예제에서는 submit-service-job 명령을 사용하여 할당량 공유에 작업을 제출합니다.

```
aws batch submit-service-job \
    --job-name "my-sagemaker-training-job" \
    --job-queue "my-sagemaker-job-queue" \
```

```
    --service-job-type "SAGEMAKER_TRAINING" \
    --quota-share-name "my_quota_share" \
    --timeout-config '{"attemptDurationSeconds":3600}' \
    --scheduling-priority 5 \
    --service-request-payload '{\"TrainingJobName\": \"sagemaker-
training-job-example\", \"AlgorithmSpecification\": {\"TrainingImage\":
 \"123456789012.dkr.ecr.us-east-1.amazonaws.com/pytorch-inference:1.8.0-cpu-
py3\", \"TrainingInputMode\": \"File\", \"ContainerEntrypoint\":  [\"sleep\",
 \"1\"]}, \"RoleArn\":\"arn:aws:iam::123456789012:role/SageMakerExecutionRole\",
 \"OutputDataConfig\": {\"S3OutputPath\": \"s3://example-bucket/model-output/\"},
 \"ResourceConfig\": {\"InstanceType\": \"ml.m5.large\", \"InstanceCount\": 1,
 \"VolumeSizeInGB\": 1}}'"
```

서비스 작업 용량 사용률 추적

AWS Batch 는 대기열의 서비스 작업에 대한 용량 사용률을 추적하는 데 함께 사용할 수 있는 여러 API 작업을 제공합니다. 모니터링 워크플로는 작업 대기열에 연결된 예약 정책의 유형에 따라 달라집니다.

선입선출(FIFO) 예약 정책을 사용하는 작업 대기열의 경우:

1. 총 대기열 사용률(GetJobQueueSnapshot)을 확인합니다.
2. SCHEDULED 및 RUNNING (ListServiceJobs)와 같은 작업을 상태별로 나열합니다.
3. 지정된 작업(DescribeServiceJob)을 검사합니다.

공정 공유(FSS) 또는 할당량 관리(QM) 예약 정책을 사용하는 작업 대기열의 경우:

1. 총 대기열 사용률(GetJobQueueSnapshot)을 확인합니다.
2. 공유당 사용률(GetJobQueueSnapshot)을 봅니다.
3. 상태별로 작업을 나열하고 SCHEDULED 및 RUNNING ()와 같이 사용률에 적극적으로 기여하는 작업을 공유합니다ListServiceJobs.
4. 지정된 작업(DescribeServiceJob)을 검사합니다.

다음 섹션에서는 각 단계를 자세히 안내합니다.

ECS, EKS 및 Fargate 컴퓨팅 작업의 용량 사용률 추적에 대한 자세한 내용은 섹션을 참조하세요컴퓨팅 작업 용량 사용률 추적.

주제

대기열 사용률 확인

GetJobQueueSnapshot 응답의 queueUtilization 필드는 대기열에서 디스패치된 작업에서 소비되는 컴퓨팅 용량의 point-in-time 보기를 제공합니다. 용량은 서비스 작업의 인스턴스 수로 측정됩니다.

공정 공유 또는 할당량 관리 예약 정책을 사용하는 작업 대기열의 경우 응답에는 공유당 분석도 포함되어 있으므로 용량이 공유에 어떻게 분산되는지 확인할 수 있습니다. 자세한 내용은 공유당 사용률 보기 단원을 참조하십시오.

용량 사용률 보기(AWS CLI)

get-job-queue-snapshot 명령을 사용하여 작업 대기열의 용량 사용률 스냅샷을 검색합니다.

```
aws batch get-job-queue-snapshot \
    --job-queue my-job-queue
```

응답은 작업 대기열에 연결된 예약 정책에 따라 달라집니다. 예약 정책 유형의 탭을 선택하여 예제 응답을 확인합니다.

First-in, first-out (FIFO)

다음은 FIFO 작업 대기열에 대한 응답의 예입니다. FIFO 대기열은 예약 정책을 사용하지 않으므로 응답에는 공유당 사용률이 포함되지 않습니다.

```
{
    "frontOfQueue": {
        "jobs": [],
        "lastUpdatedAt": 1700000000000
    },
    "queueUtilization": {
        "totalCapacityUsage": [
            {
                "capacityUnit": "ml.m5.large",
```

```
                "quantity": 9.0
            }
        ],
        "lastUpdatedAt": 1700000000000
    }
}
```

이 예제에서 대기열은 디스패치된 모든 작업에서 총 9개의 인스턴스를 사용합니다.

Fair-share scheduling (FSS)

다음은 공정 공유 작업 대기열에 대한 응답의 예입니다. queueUtilization 객체에는 공유별 분석과 함께 대기열에서 디스패치된 모든 작업이 소비한 총 용량의 point-in-time 스냅샷이 포함되어 있습니다.

```
{
    "frontOfQueue": {
        "jobs": [],
        "lastUpdatedAt": 1700000000000
    },
    "queueUtilization": {
        "totalCapacityUsage": [
            {
                "capacityUnit": "NUM_INSTANCES",
                "quantity": 9.0
            }
        ],
        "fairshareUtilization": {
            "activeShareCount": 2,
            "topCapacityUtilization": [
                {
                    "shareIdentifier": "team-a",
                    "capacityUsage": [
                        {
                            "capacityUnit": "NUM_INSTANCES",
                            "quantity": 5.0
                        }
                    ]
                },
                {
                    "shareIdentifier": "team-b",
                    "capacityUsage": [
                        {
                            "capacityUnit": "NUM_INSTANCES",
```

```
                            "quantity": 4.0
                        }
                    ]
                }
            ]
        },
        "lastUpdatedAt": 1700000000000
    }
}
```

이 예제에서 totalCapacityUsage 필드는 대기열이 총 9개의 인스턴스를 소비함을 보여줍니다. fairshareUtilization 객체는 공유별 분석을 보여줍니다. 공유는 인스턴스 5개를 team-a 사용하고 공유는 인스턴스 4개를 team-b 사용합니다.

Quota management (QM)

다음은 할당량 관리 작업 대기열에 대한 응답의 예입니다. queueUtilization 객체에는 대기열에서 디스패치된 모든 작업이 소비한 총 용량의 point-in-time 스냅샷과 per-quota-share 내역이 포함되어 있습니다. frontOfQuotaShares 객체는 할당량 공유당 첫 번째 RUNNABLE 작업을 보여줍니다.

```
{
    "frontOfQueue": {
        "jobs": [],
        "lastUpdatedAt": 1700000000000
    },
    "frontOfQuotaShares": {
        "quotaShares": {
            "team-a-share": [],
            "team-b-share": []
        },
        "lastUpdatedAt": 1700000000000
    },
    "queueUtilization": {
        "totalCapacityUsage": [
            {
                "capacityUnit": "ml.m5.large",
                "quantity": 9.0
            }
        ],
        "quotaShareUtilization": {
            "topCapacityUtilization": [
                {
```

```
                    "quotaShareName": "team-a-share",
                    "capacityUsage": [
                        {
                            "capacityUnit": "ml.m5.large",
                            "quantity": 5.0
                        }
                    ]
                },
                {
                    "quotaShareName": "team-b-share",
                    "capacityUsage": [
                        {
                            "capacityUnit": "ml.m5.large",
                            "quantity": 4.0
                        }
                    ]
                }
            ]
        },
        "lastUpdatedAt": 1700000000000
    }
}
```

이 예제에서 totalCapacityUsage 필드는 대기열이 총 9개의 인스턴스를 소비함을 보여줍니다. quotaShareUtilization 객체는 per-quota-share 분석을 보여줍니다. 할당량 공유는 인스턴스 5개를 team-a-share 사용하고 할당량 공유는 인스턴스 4개를 team-b-share 사용합니다. frontOfQuotaShares 객체는 RUNNABLE 작업이 해당 위치에 도달한 가장 빠른 시간과 함께 각 할당량 공유의 첫 번째 작업을 표시합니다.

공유당 사용률 보기

공정 공유 또는 할당량 관리 예약 정책이 있는 작업 대기열의 경우의 queueUtilization 응답에는 사용량별로 상위 활성 공유를 나열하는 topCapacityUtilization 배열이 있는 사용률 객체가 GetJobQueueSnapshot 포함됩니다.

이 정보는 다음에 도움이 됩니다.

- 가장 많은 리소스를 소비하는 공유를 식별합니다.
- 리소스가 예상대로 공유에 분산되어 있는지 확인합니다.
- 할당을 포화시키거나 과소 사용할 수 있는 공유를 감지합니다.

- 예약 정책 구성을 조정할지 여부를 결정합니다.

공정 공유 예약 정책에 대한 자세한 내용은 섹션을 참조하세요공정 공유 예약 정책.

할당량 공유에 대한 자세한 내용은 섹션을 참조하세요할당량 공유.

상태별 서비스 작업 나열 및 공유

전체 대기열 및 공유당 사용률을 식별한 후 `ListServiceJobs` API 작업을 사용하여 사용률에 적극적으로 기여하는 서비스 작업을 찾습니다. 작업 상태를 기준으로 필터링하여 RUNNING, SCHEDULED또는 다른 상태의 작업을 볼 수 있습니다. 공정 공유 또는 할당량 관리 예약 정책이 있는 대기열의 경우 공유 식별자를 기준으로 필터링하여 결과를 특정 공유로 좁힐 수도 있습니다.

> **Note**
>
> SHARE_IDENTIFIER 및 QUOTA_SHARE_NAME 필터는 `jobStatus` 파라미터와 결합할 수 있는 유일한 필터입니다. 다른 필터를 사용하면 `jobStatus` 파라미터가 무시됩니다.

서비스 작업 나열(AWS CLI)

list-service-jobs 명령을 `--job-status` 파라미터와 함께 사용하여 상태를 기준으로 필터링합니다.

대기열에서 실행 중인 서비스 작업을 봅니다.

```
aws batch list-service-jobs \
    --job-queue my-job-queue \
    --job-status RUNNING
```

공정 공유 예약 정책이 있는 대기열의 경우 `--filters` 파라미터를와 함께 사용하여 특정 공유SHARE_IDENTIFIER에 대한 작업을 나열합니다. 할당량 관리 예약 정책이 있는 대기열의 경우 QUOTA_SHARE_NAME를 사용하여 특정 할당량 공유에 대한 작업을 나열합니다. 이는 용량 소비가 많은 공유를 식별하고 책임이 있는 작업을 확인하고 싶을 때 유용합니다.

공정 공유 대기열의 공유에 대한 RUNNING 서비스 작업만 나열합니다.

```
aws batch list-service-jobs \
    --job-queue my-job-queue \
    --job-status RUNNING \
    --filters name=SHARE_IDENTIFIER,values="team-a"
```

할당량 관리 예약 정책이 있는 대기열의 경우 QUOTA_SHARE_NAME 필터를 사용합니다.

```
aws batch list-service-jobs \
    --job-queue my-job-queue \
    --job-status RUNNING \
    --filters name=QUOTA_SHARE_NAME,values="my-quota-share"
```

다음은 공정 공유 대기열에서 공유 식별자로 필터링된 실행 중인 서비스 작업을 나열하는 응답의 예입니다.

```
{
    "jobSummaryList": [
        {
            "jobArn": "arn:aws:batch:us-east-1:123456789012:service-job/
a4d6c728-8ee8-4c65-8e2a-9a5e8f4b7c3d",
            "jobId": "a4d6c728-8ee8-4c65-8e2a-9a5e8f4b7c3d",
            "jobName": "my-training-job",
            "serviceJobType": "SAGEMAKER_TRAINING",
            "status": "RUNNING",
            "shareIdentifier": "team-a",
            "createdAt": 1700000000000,
            "scheduledAt": 1700000060000,
            "startedAt": 1700000120000,
            "capacityUsage": [
                {
                    "capacityUnit": "ml.m5.large",
                    "quantity": 5.0
                }
            ],
            "latestAttempt": {
                "serviceResourceId": {
                    "name": "TrainingJobArn",
                    "value": "arn:aws:sagemaker:us-east-1:123456789012:training-job/my-
training-job"
                }
            }
        }
    ]
}
```

이 예제에서 응답에는 작업이 team-a 공유에 속함을 보여주는 shareIdentifier 필드와 작업이 5개의 ml.m5.large 인스턴스를 소비함을 보여주는 capacityUsage 배열이 포함됩니다.

`latestAttempt` 객체에는 대상 서비스에서 추가 세부 정보를 가져오는 데 사용할 수 있는 서비스 리소스 식별자가 포함되어 있습니다.

특정 서비스 작업 검사

관심 있는 서비스 작업을 식별한 후 `DescribeServiceJob` 작업을 사용하여 현재 상태, 서비스 리소스 식별자, 세부 시도 정보 등 작업에 대한 포괄적인 정보를 가져옵니다.

특정 서비스 작업에 대한 세부 정보를 봅니다.

```
aws batch describe-service-job \
    --job-id a4d6c728-8ee8-4c65-8e2a-9a5e8f4b7c3d
```

이 명령은 다음을 포함하여 작업에 대한 포괄적인 정보를 반환합니다.

- 작업 ARN 및 현재 상태
- 서비스 리소스 식별자(예: SageMaker 훈련 작업 ARN)
- 예약 우선 순위 및 재시도 구성
- 원래 서비스 파라미터를 포함하는 서비스 요청 페이로드
- 시작 및 중지 시간이 포함된 자세한 시도 정보
- 대상 서비스의 상태 메시지

기본 SageMaker 훈련 작업 검사

를 통해 SageMaker 훈련 작업을 모니터링할 때 AWS Batch 작업 정보와 기본 SageMaker 훈련 작업 세부 정보에 모두 액세스할 AWS Batch수 있습니다.

작업 세부 정보의 서비스 리소스 식별자에는 SageMaker 훈련 작업 ARN이 포함됩니다.

```
{
    "latestAttempt": {
        "serviceResourceId": {
            "name": "TrainingJobArn",
            "value": "arn:aws:sagemaker:us-east-1:123456789012:training-job/my-
training-job"
        }
    }
}
```

이 ARN을 사용하여 SageMaker에서 추가 세부 정보를 직접 가져올 수 있습니다.

```
aws sagemaker describe-training-job \
    --training-job-name my-training-job
```

AWS Batch 상태와 SageMaker 훈련 작업 상태를 모두 확인하여 작업 진행 상황을 모니터링합니다. AWS Batch 작업 상태에는 전체 작업 수명 주기를 표시하는 반면, SageMaker 훈련 작업 상태는 훈련 프로세스에 대한 서비스별 세부 정보를 제공합니다.

컴퓨팅 작업 용량 사용률 추적

AWS Batch 는 대기열에서 ECS, EKS 및 Fargate 컴퓨팅 작업의 용량 사용률을 추적하는 데 함께 사용할 수 있는 여러 API 작업을 제공합니다. 모니터링 워크플로는 작업 대기열에 연결된 예약 정책의 유형에 따라 달라집니다.

선입선출(FIFO) 예약 정책을 사용하는 작업 대기열의 경우:

1. 총 대기열 사용률(GetJobQueueSnapshot)을 확인합니다.
2. RUNNABLE 및 RUNNING (ListJobs)와 같은 작업을 상태별로 나열합니다.
3. 지정된 작업(DescribeJobs)을 검사합니다.

공정 공유(FSS) 예약 정책을 사용하는 작업 대기열의 경우:

1. 총 대기열 사용률(GetJobQueueSnapshot)을 확인합니다.
2. 공유당 사용률(GetJobQueueSnapshot)을 봅니다.
3. 상태별로 작업을 나열하고 RUNNABLE 및 RUNNING ()와 같이 사용률에 적극적으로 기여하는 작업을 공유합니다ListJobs.
4. 지정된 작업(DescribeJobs)을 검사합니다.

다음 섹션에서는 각 단계를 자세히 안내합니다.

서비스 작업의 용량 사용률 추적에 대한 자세한 내용은 섹션을 참조하세요서비스 작업 용량 사용률 추적.

주제

- 대기열 사용률 확인

- 공유당 사용률 보기
- 상태별 컴퓨팅 작업 나열 및 공유
- 특정 컴퓨팅 작업 검사

대기열 사용률 확인

GetJobQueueSnapshot 응답의 queueUtilization 필드는 대기열에서 디스패치된 작업에서 소비되는 컴퓨팅 용량의 point-in-time 보기를 제공합니다. 용량은 컴퓨팅 작업vCPUs로 측정됩니다.

공정 공유 예약 정책을 사용하는 작업 대기열의 경우 응답에는 공유당 분석도 포함되어 있으므로 용량이 공유에 어떻게 분산되는지 확인할 수 있습니다. 자세한 내용은 공유당 사용률 보기 단원을 참조하십시오.

용량 사용률 보기(AWS CLI)

get-job-queue-snapshot 명령을 사용하여 작업 대기열의 용량 사용률 스냅샷을 검색합니다.

```
aws batch get-job-queue-snapshot \
    --job-queue my-job-queue
```

응답은 작업 대기열에 연결된 예약 정책에 따라 달라집니다. 예약 정책 유형의 탭을 선택하여 예제 응답을 확인합니다.

First-in, first-out (FIFO)

다음은 컴퓨팅 작업을 실행하는 FIFO 작업 대기열에 대한 응답의 예입니다. FIFO 대기열은 예약 정책을 사용하지 않으므로 응답에는 공유당 사용률이 포함되지 않습니다.

```
{
    "frontOfQueue": {
        "jobs": [],
        "lastUpdatedAt": 1700000000000
    },
    "queueUtilization": {
        "totalCapacityUsage": [
            {
                "capacityUnit": "vCPU",
                "quantity": 96.0
            }
        ],
```

```
        "lastUpdatedAt": 1700000000000
    }
}
```

이 예제에서 대기열은 모든 디스패치된 작업에서 총 96vCPUs를 사용합니다.

Fair-share scheduling (FSS)

다음은 공정 공유 작업 대기열에 대한 응답의 예입니다. queueUtilization 객체에는 공유별 분석과 함께 대기열에서 디스패치된 모든 작업이 소비한 총 용량의 point-in-time 스냅샷이 포함되어 있습니다.

```
{
    "frontOfQueue": {
        "jobs": [],
        "lastUpdatedAt": 1700000000000
    },
    "queueUtilization": {
        "totalCapacityUsage": [
            {
                "capacityUnit": "vCPU",
                "quantity": 192.0
            }
        ],
        "fairshareUtilization": {
            "activeShareCount": 2,
            "topCapacityUtilization": [
                {
                    "shareIdentifier": "team-a",
                    "capacityUsage": [
                        {
                            "capacityUnit": "vCPU",
                            "quantity": 128.0
                        }
                    ]
                },
                {
                    "shareIdentifier": "team-b",
                    "capacityUsage": [
                        {
                            "capacityUnit": "vCPU",
                            "quantity": 64.0
                        }
                    ]
```

```
                }
            ]
        },
        "lastUpdatedAt": 1700000000000
    }
}
```

이 예제에서 `totalCapacityUsage` 필드는 대기열이 총 192vCPUs를 소비함을 보여줍니다. `fairshareUtilization` 객체는 공유별 분석을 보여줍니다. 공유는 128개의 vCPUs `team-a` 소비하고 공유는 64개의 vCPU를 `team-b` 소비합니다. vCPUs

공유당 사용률 보기

공정 공유 예약 정책이 있는 작업 대기열의 경우의 `queueUtilization` 응답에는 사용량별로 상위 활성 공유를 나열하는 `topCapacityUtilization` 배열이 있는 `fairshareUtilization` 객체가 `GetJobQueueSnapshot` 포함됩니다.

이 정보는 다음에 도움이 됩니다.

- 가장 많은 리소스를 소비하는 공유를 식별합니다.
- 공정 공유 일정이 예상대로 리소스를 배포하고 있는지 확인합니다.
- 할당을 포화시키거나 과소 사용할 수 있는 공유를 감지합니다.
- 예약 정책에서 공유 가중치를 조정할지 여부를 결정합니다.

공정 공유 예약 정책에 대한 자세한 내용은 섹션을 참조하세요공정 공유 예약 정책.

상태별 컴퓨팅 작업 나열 및 공유

전체 대기열 및 공유당 사용률을 식별한 후 `ListJobs` API 작업을 사용하여 사용률에 적극적으로 기여하는 컴퓨팅 작업을 찾습니다. 작업 상태를 기준으로 필터링하여 `RUNNING`, `RUNNABLE`또는 다른 상태의 작업을 볼 수 있습니다. 공정 공유 예약 정책이 있는 대기열의 경우 공유 식별자를 기준으로 필터링하여 결과를 특정 공유로 좁힐 수도 있습니다.

> **Note**
>
> `SHARE_IDENTIFIER` 필터는 `jobStatus` 파라미터와 결합할 수 있는 유일한 필터입니다. 다른 필터를 사용하면 `jobStatus` 파라미터가 무시됩니다.

컴퓨팅 작업 나열(AWS CLI)

list-jobs 명령을 `--job-status` 파라미터와 함께 사용하여 상태를 기준으로 필터링합니다.

대기열에서 실행 중인 컴퓨팅 작업을 봅니다.

```
aws batch list-jobs \
    --job-queue my-job-queue \
    --job-status RUNNING
```

발송 대기 중인 컴퓨팅 작업 보기:

```
aws batch list-jobs \
    --job-queue my-job-queue \
    --job-status RUNNABLE
```

공정 공유 예약 정책이 있는 대기열의 경우 `--filters` 파라미터를와 함께 사용하여 특정 공유SHARE_IDENTIFIER에 대한 작업을 나열합니다. 이는 용량 소비가 많은 공유를 식별하고 책임이 있는 작업을 확인하고 싶을 때 유용합니다.

공정 공유 대기열의 공유에 대한 RUNNING 컴퓨팅 작업만 나열합니다.

```
aws batch list-jobs \
    --job-queue my-job-queue \
    --job-status RUNNING \
    --filters name=SHARE_IDENTIFIER,values="team-a"
```

다음은 실행 중인 컴퓨팅 작업을 나열하기 위한 응답의 예입니다.

```
{
    "jobSummaryList": [
        {
            "jobArn": "arn:aws:batch:us-east-1:123456789012:job/
b5e7d839-9ff9-5d76-9f3b-0b6f9g5c8e4f",
            "jobId": "b5e7d839-9ff9-5d76-9f3b-0b6f9g5c8e4f",
            "jobName": "my-data-processing-job",
            "status": "RUNNING",
            "shareIdentifier": "team-a",
            "createdAt": 1700000000000,
            "startedAt": 1700000120000,
            "capacityUsage": [
```

```
                {
                    "capacityUnit": "vCPU",
                    "quantity": 4.0
                }
            ],
            "container": {
                "exitCode": null
            },
            "jobDefinition": "arn:aws:batch:us-east-1:123456789012:job-definition/my-
job-def:1"
        }
    ]
}
```

특정 컴퓨팅 작업 검사

관심 있는 컴퓨팅 작업을 식별한 후 DescribeJobs 작업을 사용하여 현재 상태, 컨테이너 세부 정보 및 리소스 구성을 포함하여 작업에 대한 포괄적인 정보를 가져옵니다.

특정 컴퓨팅 작업에 대한 세부 정보를 봅니다.

```
aws batch describe-jobs \
    --jobs b5e7d839-9ff9-5d76-9f3b-0b6f9g5c8e4f
```

이 명령은 다음을 포함하여 작업에 대한 포괄적인 정보를 반환합니다.

- 작업 ARN 및 현재 상태
- 컨테이너 구성 및 리소스 요구 사항(vCPUs 및 메모리)
- 작업 정의 및 컴퓨팅 환경 세부 정보
- 예약 우선 순위 및 재시도 구성
- 시작 및 중지 시간이 포함된 자세한 시도 정보
- 컨테이너 로그에 액세스하기 위한 로그 스트림 정보

작업 정의

AWS Batch 작업 정의는 작업 실행 방법을 지정합니다. 각 작업은 작업 정의를 참조해야 하는 반면, 작업 정의에 지정된 대부분의 파라미터는 실행 시간에 재정의될 수 있습니다.

작업 정의에 지정되는 일부 속성은 다음과 같습니다.

- 작업의 컨테이너에 사용할 도커 이미지
- 컨테이너에 사용할 vCPU 수와 메모리 크기
- 컨테이너 시작 시 컨테이너가 실행할 명령
- 컨테이너 시작 시 컨테이너로 전달할 환경 변수(있는 경우)
- 컨테이너에 사용해야 할 데이터 볼륨
- 작업에서 AWS 권한에 사용해야 하는 IAM 역할(있는 경우).

내용

단일 노드 작업 정의 생성

에서 작업을 실행하려면 먼저 작업 정의를 생성 AWS Batch해야 합니다. 이 프로세스는 단일 노드 작업과 다중 노드 병렬 작업 간에 약간 다릅니다. 이 주제에서는 특히 다중 노드 병렬 작업(그룹 스케줄링이라고도 함)이 아닌 AWS Batch 작업에 대한 작업 정의를 생성하는 방법을 다룹니다.

Amazon Elastic Container 리소스에서 다중 노드 병렬 작업 정의를 생성하려면 다음을 수행합니다. 자세한 내용은 the section called “다중 노드 병렬 작업 정의 생성” 단원을 참조하십시오.

주제

- Amazon EC2 리소스에 단일 노드 작업 정의 생성
- Fargate 리소스에 단일 노드 작업 정의 생성
- Amazon EKS 리소스에 단일 노드 작업 정의 생성
- Amazon EC2 리소스에 다중 컨테이너에 대한 단일 노드 작업 정의 생성

Amazon EC2 리소스에 단일 노드 작업 정의 생성

Amazon Elastic Compute Cloud(Amazon EC2) 리소스에서 단일 노드 작업 정의를 생성하려면 다음 단계를 완료합니다.

Amazon EC2 리소스에 새 작업 정의를 생성하려면

1. https://console.aws.amazon.com/batch/ AWS Batch 콘솔을 엽니다.
2. 탐색 모음에서 사용할 AWS 리전 를 선택합니다.
3. 왼쪽 탐색 창에서 작업 정의를 선택합니다.
4. 생성(Create)을 선택합니다.
5. 오케스트레이션 유형으로 Amazon Elastic Compute Cloud(Amazon EC2)를 선택합니다.
6. EC2 플랫폼 구성의 경우 다중 노드 병렬 처리 활성화를 끕니다.
7. 이름(Name)에 고유한 작업 정의 이름을 입력합니다. 각 이름의 최대 길이는 128자입니다. 대문자 및 소문자, 숫자, 하이픈(-) 및 밑줄(_)을 포함할 수 있습니다.
8. (선택 사항) 실행 제한 시간에 제한 시간 값(초)을 입력합니다. 실행 제한 시간은 완료되지 않은 작업이 종료되기까지의 시간입니다. 시도가 제한 시간을 초과하면 중지되고 상태가 `FAILED`(으)로 변경됩니다. 자세한 내용은 작업 제한 시간 단원을 참조하십시오. 최솟값은 60초입니다.
9. (선택 사항) 예약 우선 순위를 켭니다. 0에서 100 사이의 예약 우선 순위 값을 입력합니다. 값이 높을수록 우선 순위가 높습니다.
10. (선택 사항) 작업 시도의 경우 작업을 `RUNNABLE` 상태로 이동하려는 AWS Batch 시도 횟수를 입력합니다. 1~10 사이의 숫자를 입력합니다.
11. (선택 사항) 재시도 전략 조건의 경우 종료 시 평가 추가를 선택합니다. 파라미터 값을 하나 이상 입력한 다음 작업을 선택합니다. 각 조건 세트에 대해 작업을 재시도 또는 종료로 설정해야 합니다. 이러한 작업은 다음을 의미합니다.

- 재시도 - 지정한 작업 시도 횟수에 도달할 때까지 AWS Batch 재시도합니다.
- 종료 - 작업 재시도를 AWS Batch 중지합니다.

> **Important**
>
> 종료 시 평가 추가를 선택한 경우 하나 이상의 파라미터를 구성하고 작업을 선택하거나 종료 시 평가 제거를 선택해야 합니다.

12. (선택 사항) 태그를 확장한 다음 태그 추가를 선택하여 리소스에 태그를 추가합니다. 키와 선택 값을 입력하고 태그 추가를 선택합니다.
13. (선택 사항) 작업 및 작업 정의에서 Amazon ECS 작업으로 태그를 전파하려면 태그 전파를 활성화합니다.
14. 다음 페이지를 선택합니다.
15. 컨테이너 구성 섹션에서:

    a. 이미지에서 작업에 사용할 Docker 이미지를 선택합니다. 기본적으로 Docker Hub 레지스트리 내 이미지는 사용 가능합니다. 또한 *repository-url*/*image*:*tag*(을)를 사용하여 다른 리포지토리를 지정할 수도 있습니다. 각 이름의 최대 길이는 225자입니다. 여기에는 대문자와 소문자, 숫자, 하이픈(-), 밑줄(_), 콜론(:), 슬래시(/) 및 숫자 기호(#)를 사용할 수 있습니다. 이 파라미터는 Docker 원격 API(Docker Remote API)의 컨테이너 생성(Create a container) 섹션에 있는 `Image`(와)과 docker run의 `IMAGE` 파라미터로 매핑됩니다.

    > **Note**
    >
    > 도커 이미지 아키텍처는 예정된 컴퓨팅 리소스의 프로세서 아키텍처와 일치해야 합니다. 예를 들어 ARM 기반 Docker 이미지는 ARM 기반 컴퓨팅 리소스에서만 실행할 수 있습니다.

    - Amazon ECR Public 리포지토리에 있는 이미지는 전체 `registry/repository[:tag]` 또는 `registry/repository[@digest]` 명명 규칙을 사용합니다(예: `public.ecr.aws/`*registry_alias*`/`*my-web-app*`:`*latest*).
    - Amazon ECR 리포지토리에 있는 이미지는 전체 `registry/repository[:tag]` 명명 규칙을 사용합니다 (예: *aws_account_id*`.dkr.ecr.`*region*`.amazonaws.com/`*my-web-app*`:`*latest*).

- Docker Hub의 공식 리포지토리 안의 이미지는 단일 이름을 사용합니다(예: `ubuntu` 또는 `mongo`).
- Docker Hub 다른 리포지토리에 저장된 이미지는 조직 이름으로 한정됩니다(예: `amazon/amazon-ecs-agent`).
- 다른 온라인 리포지토리 안의 이미지는 도메인 이름을 사용하여 추가로 한정됩니다(예: `quay.io/assemblyline/ubuntu`).

b. 명령에서 필드에 명령을 JSON 문자열 배열 형식으로 입력합니다.

이 파라미터는 도커 원격 API(Docker Remote API)의 컨테이너 생성(Create a container) 섹션에 있는 `Cmd`(와)과 docker run의 `COMMAND` 파라미터로 매핑됩니다. Docker `CMD` 파라미터에 대한 자세한 정보는 https://docs.docker.com/engine/reference/builder/#cmd를 참조하세요.

> **Note**
>
> 사용자는 명령에 파라미터 대체 및 자리 표시자 기본값을 사용할 수 있습니다. 자세한 내용은 Parameters 단원을 참조하십시오.

c. (선택 사항) 실행 역할에서 Amazon ECS 컨테이너 에이전트에게 사용자를 대신하여 AWS API를 호출할 수 있는 권한을 부여하는 IAM 역할을 지정합니다. 이 기능은 Amazon ECS IAM 역할을 작업에 사용합니다. 자세한 내용을 알아보려면 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS 작업 실행 IAM 역할 섹션을 참조하세요.

d. 작업 역할 구성에서 AWS APIs에 대한 권한이 있는 IAM 역할을 선택합니다. 이 기능은 Amazon ECS IAM 역할을 작업에 사용합니다. 자세한 내용은 Amazon Elastic Container Service 개발자 안내서의 태스크에 대한 IAM 역할을 참조하세요.

> **Note**
>
> Amazon Elastic Container Service 태스크 역할 신뢰 관계를 보유한 역할만 여기 표시됩니다. AWS Batch 작업에 대한 IAM 역할 생성에 대한 자세한 내용은 Amazon Elastic Container Service 개발자 안내서의 작업에 대한 IAM 역할 및 정책 생성을 참조하세요.

16. 파라미터의 경우 파라미터 추가를 선택하여 파라미터 대체 플레이스홀더를 키 및 값(선택 사항) 페어로 추가합니다.
17. 환경 구성 섹션에서:

a. vCPU에서 컨테이너에 예약할 vCPU 수를 지정합니다. 이 파라미터는 Docker 원격 API(Docker Remote API)의 컨테이너 생성(Create a container) 섹션에 있는 `CpuShares`(와)과 docker run에 대한 `--cpu-shares` 옵션에 매핑됩니다. 각 vCPU는 1,024개의 CPU 공유와 동일합니다. vCPU를 최소 하나 이상 지정해야 합니다.

b. 메모리에는 컨테이너에 사용할 수 있는 메모리 한도를 입력합니다. 컨테이너가 여기서 지정된 메모리 용량을 초과하면 해당 컨테이너가 중지됩니다. 이 파라미터는 Docker 원격 API(Docker Remote API)의 컨테이너 생성(Create a container) 섹션에 있는 `Memory`(와)과 docker run에 대한 `--memory` 옵션에 매핑됩니다. 한 작업에 대해 메모리를 최소한 4MiB 지정해야 합니다.

> **Note**
>
> 리소스 사용률을 극대화하려면 특정 인스턴스 유형의 작업에 메모리 우선 순위를 지정합니다. 자세한 내용은 컴퓨팅 리소스 메모리 관리 단원을 참조하십시오.

c. GPU 개수에서 컨테이너에 예약할 GPU의 개수를 선택합니다.

d. (선택 사항) 환경 변수의 경우 환경 변수 추가를 선택하여 환경 변수를 이름-값 쌍으로 추가합니다. 이러한 변수는 컨테이너로 전달됩니다.

e. (선택 사항) 암호의 경우 암호 추가를 선택하여 암호를 이름-값 쌍으로 추가합니다. 이러한 보안 암호는 컨테이너에 노출됩니다. 자세한 내용은 LogConfiguration:secretOptions를 참조하세요.

18. 다음 페이지를 선택합니다.

19. Linux 구성 섹션에서:

a. User(사용자)에서 컨테이너 내부에서 사용할 사용자 이름을 입력합니다. 이 파라미터는 Docker 원격 API(Docker Remote API)의 컨테이너 생성(Create a container) 섹션에 있는 `User`(와)과 docker run에 대한 `--user` 옵션에 매핑됩니다.

b. (선택 사항) 호스트 인스턴스에 대한 상위 권한을 작업 컨테이너에 부여하려면 (`root` 사용자와 유사) 권한이 있음 슬라이더를 오른쪽으로 드래그합니다. 이 파라미터는 Docker 원격 API(Docker Remote API)의 컨테이너 생성(Create a container) 섹션에 있는 `Privileged`(와)과 docker run에 대한 `--privileged` 옵션에 매핑됩니다.

c. (선택 사항) 컨테이너 내에서 `init` 프로세스를 실행하려면 init 프로세스 활성화를 켭니다. 이 프로세스는 신호를 전달하고 결과를 받아들입니다.

20. (선택 사항) 파일 시스템 구성 섹션에서:

a. 읽기 전용 파일 시스템 활성화를 켜서 볼륨에 대한 쓰기 권한을 제거합니다.

b. 공유 메모리 크기에 `/dev/shm` 볼륨의 크기(MiB)를 입력합니다.

c. 최대 스왑 크기에는 컨테이너가 사용할 수있는 총 스왑 메모리 양(MiB)을 입력합니다.

d. 스왑 활용도의 경우 컨테이너의 스왑 동작을 나타내는 값을 0에서 100 사이의 값으로 입력합니다. 값을 지정하지 않고 스와핑이 활성화된 경우 기본값 60이 사용됩니다. 자세한 내용은 LinuxParameters:swappiness를 참조하세요.

e. (선택 사항) 추가 구성을 확장합니다.

f. (선택 사항) Tmpfs의 경우 tmpfs 추가를 선택하여 `tmpfs` 마운트를 추가합니다.

g. (선택 사항) 디바이스의 경우 디바이스 추가를 선택하여 장치를 추가합니다.

   i. 컨테이너 경로에 호스트 인스턴스에 매핑된 디바이스를 노출할 컨테이너 인스턴스의 경로를 지정합니다. 이 필드를 비워두면 호스트 경로가 컨테이너에 사용됩니다.

   ii. 호스트 경로에 호스트 인스턴스의 디바이스 경로를 지정합니다.

   iii. 권한에서 디바이스에 적용할 권한을 하나 이상 선택합니다. 사용 가능한 권한은 읽기, 쓰기 및 MKNOD입니다.

h. (선택 사항) 볼륨 구성의 경우 볼륨 추가를 선택하여 컨테이너에 전달할 볼륨 목록을 생성합니다. 볼륨의 이름 및 소스 경로를 입력한 다음 볼륨 추가를 선택합니다. 또한 EFS 활성화를 켜도록 선택할 수도 있습니다.

i. (선택 사항) 마운트 포인트의 경우 마운트 포인트 구성 추가를 선택하여 데이터 볼륨의 마운트 포인트를 추가합니다. 소스 볼륨과 컨테이너 경로를 지정해야 합니다. 이러한 마운트 포인트는 컨테이너 인스턴스의 Docker daemon에 전달됩니다. 볼륨을 읽기 전용으로 설정할 수도 있습니다.

j. (선택 사항) Ulimits 구성의 경우 ulimit 추가를 선택하여 컨테이너에 `ulimits` 값을 추가합니다. 이름, 소프트 제한, 하드 제한 값을 입력한 다음 ulimit 추가를 선택합니다.

21. 작업 속성 섹션에서:

a. 실행 역할 - 조건부에서 Amazon ECS 에이전트가 사용자를 대신하여 AWS API를 호출할 수 있도록 허용하는 역할을 선택합니다. 실행 역할 생성에 대한 자세한 내용은 자습서: IAM 실행 역할 생성 섹션을 참조하세요.

b. ECS 실행 명령을 선택하여 Amazon ECS 컨테이너 쉘에 대한 직접 액세스를 허용하고 호스트 OS를 우회합니다. 태스크 역할을 선택해야 합니다.

> **Important**
>
> ECS 실행 명령을 사용하려면 파일 시스템에 대한 쓰기 권한이 있어야 합니다.

c. 작업 역할에서 Amazon ECS Identity and Access Management(IAM) 역할을 선택하여 컨테이너가 사용자를 대신하여 AWS API를 호출하도록 허용합니다. 자세한 내용은 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS 태스크 IAM 역할을 참조하세요.

22. (선택 사항)로깅 구성 섹션에서:

a. 로그 드라이버에서 사용할 로그 드라이버를 선택합니다. 사용 가능한 로그 드라이버에 대한 자세한 내용은 LogConfiguration:logDriver를 참조하세요.

> **Note**
>
> 기본적으로 `awslogs` 로그 드라이버가 사용됩니다.

b. 옵션에서 옵션 추가를 선택하여 옵션을 추가합니다. 이름-값 쌍을 입력한 다음 옵션 추가를 선택합니다.

c. 암호에서 암호 추가를 선택합니다. 이름-값 페어를 입력한 다음 암호 추가를 선택하여 암호를 추가합니다.

> **Tip**
>
> 자세한 내용은 LogConfiguration:secretOptions를 참조하세요.

23. 다음 페이지를 선택합니다.

24. 작업 정의 검토에서 구성 단계를 검토하세요. 변경해야 하는 경우 편집을 선택합니다 작업을 마쳤으면 작업 정의 생성을 선택합니다.

Fargate 리소스에 단일 노드 작업 정의 생성

다음 단계를 완료하여 AWS Fargate 리소스에 대한 단일 노드 작업 정의를 생성합니다.

Fargate 리소스에서 새 작업 정의를 생성하려면

1. https://console.aws.amazon.com/batch/ AWS Batch 콘솔을 엽니다.
2. 상단 탐색 모음에서 사용할 AWS 리전 를 선택합니다.

3. 왼쪽 탐색 창에서 작업 정의를 선택합니다.
4. 생성을 선택합니다.
5. 오케스트레이션 유형에서 Fargate를 선택합니다. 자세한 내용은 Fargate 컴퓨팅 환경 단원을 참조하십시오.
6. 이름(Name)에 고유한 작업 정의 이름을 입력합니다. 각 이름의 최대 길이는 128자입니다. 대문자 및 소문자, 숫자, 하이픈(-) 및 밑줄(_)을 포함할 수 있습니다.
7. (선택 사항) 실행 제한 시간에 제한 시간 값(초)을 입력합니다. 실행 제한 시간은 완료되지 않은 작업이 종료되기까지의 시간입니다. 시도가 제한 시간을 초과하면 중지되고 상태가 `FAILED`(으)로 변경됩니다. 자세한 내용은 작업 제한 시간 단원을 참조하십시오. 최솟값은 60초입니다.
8. (선택 사항) 예약 우선 순위를 켭니다. 0에서 100 사이의 예약 우선 순위 값을 입력합니다. 값이 높을수록 낮은 값보다 우선 순위가 높습니다.
9. (선택 사항) 태그를 확장한 다음, 태그 추가를 선택하여 리소스에 태그를 추가합니다. 작업 및 작업 정의의 태그를 전파하려면 태그 전파 시작을 켭니다.
10. Fargate 플랫폼 구성 섹션에서:
    a. 런타임 플랫폼에서 컴퓨팅 환경 아키텍처를 선택합니다.
    b. 운영 체제 제품군에서 컴퓨팅 환경의 운영 체제를 선택합니다.
    c. CPU 아키텍처에서 vCPU 아키텍처를 선택합니다.
    d. Fargate 플랫폼 버전에서 `LATEST` 또는 특정 런타임 환경 버전을 입력합니다.
    e. (선택 사항)퍼블릭 IP 할당을 켜서 Fargate 작업 네트워크 인터페이스에 퍼블릭 IP 주소를 할당합니다. 프라이빗 서브넷에서 실행 중인 작업의 경우 프라이빗 서브넷에 인터넷으로 아웃바운드 트래픽을 라우팅할 NAT 게이트웨이가 연결되어 있어야 합니다. 컨테이너 이미지를 가져올 수 있도록 이 작업을 수행하는 것이 좋습니다. 자세한 내용은 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS 작업 네트워킹을 참조하세요.
    f. (선택 사항)임시 스토리지에 태스크에 할당할 임시 스토리지의 양을 입력합니다. 임시 스토리지의 용량은 21GiB에서 200GiB 사이여야 합니다. 값을 입력하지 않으면 기본적으로 20GiB의 임시 스토리지가 할당됩니다.

       > **Note**
       > 임시 스토리지에는 Fargate 플랫폼 버전 1.4 이상이 필요합니다.

    g. 실행 역할에서 Amazon ECS 컨테이너 및 Fargate 에이전트에게 사용자를 대신하여 AWS API를 호출할 수 있는 권한을 부여하는 IAM 역할을 지정합니다. 이 기능은 태스크 기능에 Amazon ECS IAM 역할을 사용합니다. 구성 사전 조건을 포함하여 자세한 내용을 알아보려면

Amazon Elastic Container Service 개발자 안내서의 Amazon ECS 작업 실행 IAM 역할을 참조하세요.

h. 작업 시도에 AWS Batch 이 작업을 RUNNABLE 상태로 전환하기 시도하는 횟수를 입력합니다. 1~10 사이의 숫자를 입력합니다.

i. (선택 사항) 재시도 전략 조건의 경우 종료 시 평가 추가를 선택합니다. 파라미터 값을 하나 이상 입력한 다음 작업을 선택합니다. 각 조건 세트에 대해 작업을 재시도 또는 종료로 설정해야 합니다. 이러한 작업은 다음을 의미합니다.

- 재시도 - 지정한 작업 시도 횟수에 도달할 때까지 AWS Batch 재시도합니다.
- 종료 - 작업 재시도를 AWS Batch 중지합니다.

> **Important**
>
> 종료 시 평가 추가를 선택한 경우 하나 이상의 파라미터를 구성하고 작업을 선택하거나 종료 시 평가 제거를 선택해야 합니다.

11. 다음 페이지를 선택합니다.

12. 컨테이너 구성 섹션에서:

a. 이미지에서 작업에 사용할 도커 이미지를 선택합니다. 기본적으로 Docker Hub 레지스트리 내 이미지는 사용 가능합니다. 또한 *repository-url*/*image*:*tag*(을)를 사용하여 다른 리포지토리를 지정할 수도 있습니다. 각 이름의 최대 길이는 225자입니다. 대문자와 소문자, 숫자, 하이픈(-), 밑줄(_), 콜론(:), 마침표(.), 슬래시(/) 및 숫자 기호(#)를 포함할 수 있습니다. 이 파라미터는 Docker 원격 API(Docker Remote API)의 컨테이너 생성(Create a container) 섹션에 있는 Image(와)과 docker run의 IMAGE 파라미터로 매핑됩니다.

> **Note**
>
> 도커 이미지 아키텍처는 예정된 컴퓨팅 리소스의 프로세서 아키텍처와 일치해야 합니다. 예를 들어 ARM 기반 Docker 이미지는 ARM 기반 컴퓨팅 리소스에서만 실행할 수 있습니다.

- Amazon ECR Public 리포지토리에 있는 이미지는 전체 `registry/repository[:tag]` 또는 `registry/repository[@digest]` 명명 규칙을 사용합니다(예: `public.ecr.aws/`*registry_alias*`/`*my-web-app*`:`*latest*).

- Amazon ECR 리포지토리에 있는 이미지는 전체 `registry/repository[:tag]` 명명 규칙을 사용합니다 (예: *`aws_account_id`*`.dkr.ecr.`*`region`*`.amazonaws.com/`*`my-web-app`*`:`*`latest`*).
- Docker Hub의 공식 리포지토리 안의 이미지는 단일 이름을 사용합니다(예: `ubuntu` 또는 `mongo`).
- Docker Hub 다른 리포지토리에 저장된 이미지는 조직 이름으로 한정됩니다(예: `amazon/amazon-ecs-agent`).
- 다른 온라인 리포지토리 안의 이미지는 도메인 이름을 사용하여 추가로 한정됩니다(예: `quay.io/assemblyline/ubuntu`).

b. 명령에서 필드에 명령을 JSON 문자열 배열 형식으로 입력합니다.

이 파라미터는 도커 원격 API(Docker Remote API)의 컨테이너 생성(Create a container) 섹션에 있는 `Cmd`(와)과 docker run의 `COMMAND` 파라미터로 매핑됩니다. Docker `CMD` 파라미터에 대한 자세한 정보는 https://docs.docker.com/engine/reference/builder/#cmd를 참조하세요.

> **Note**
>
> 사용자는 명령에 파라미터 대체 및 자리 표시자 기본값을 사용할 수 있습니다. 자세한 내용은 Parameters 단원을 참조하십시오.

c. (선택 사항)작업 정의에 파라미터를 이름-값 매핑으로 추가하여 작업 정의 기본값을 재정의합니다. 파라미터를 추가하려면

- 파라미터에서 파라미터 추가를 선택하고 이름-값 쌍을 입력한 다음 파라미터 추가를 선택합니다.

> **Important**
>
> 파라미터 추가를 선택한 경우 하나 이상의 파라미터를 구성하거나 파라미터 제거를 선택해야 합니다.

d. 환경 구성 섹션에서:

i. 작업 역할 구성에서 AWS APIs에 대한 권한이 있는 IAM 역할을 선택합니다. 이 기능은 태스크 기능에 Amazon ECS IAM 역할을 사용합니다. 자세한 내용은 Amazon Elastic Container Service 개발자 안내서의 태스크에 대한 IAM 역할을 참조하세요.

> **Note**
>
> Amazon Elastic Container Service 태스크 역할 신뢰 관계를 보유한 역할만 여기 표시됩니다. AWS Batch 작업에 대한 IAM 역할을 생성하는 방법에 대한 자세한 내용은 Amazon Elastic Container Service 개발자 안내서의 작업에 대한 IAM 역할 및 정책 생성을 참조하세요.

ii. vCPU에서 컨테이너에 예약할 vCPU 수를 지정합니다. 이 파라미터는 Docker 원격 API(Docker Remote API)의 컨테이너 생성(Create a container) 섹션에 있는 `CpuShares`(와)과 docker run에 대한 `--cpu-shares` 옵션에 매핑됩니다. 각 vCPU는 1,024개의 CPU 공유와 동일합니다. vCPU를 최소 하나 이상 지정해야 합니다.

iii. 메모리에는 컨테이너에 사용할 수 있는 메모리 한도를 입력합니다. 컨테이너가 여기에 지정된 메모리를 초과하려 하면 해당 컨테이너가 중지됩니다. 이 파라미터는 Docker 원격 API(Docker Remote API)의 컨테이너 생성(Create a container) 섹션에 있는 `Memory`(와)과 docker run에 대한 `--memory` 옵션에 매핑됩니다. 한 작업에 대해 메모리를 최소한 4MiB 지정해야 합니다.

GuardDuty Runtime Monitoring을 사용하는 경우 GuardDuty 보안 에이전트에 약간의 메모리 오버헤드가 있습니다. 따라서 메모리 제한에 GuardDuty 보안 에이전트의 크기가 포함되어야 합니다. GuardDuty 보안 에이전트 메모리 제한에 대한 자세한 내용은 GuardDuty 사용 설명서의 CPU and memory limits를 참조하세요. 모범 사례에 대한 자세한 내용은 Amazon ECS 개발자 안내서의 런타임 모니터링을 활성화한 후 Fargate 작업에서 메모리 오류를 해결하는 방법을 참조하세요.

> **Note**
>
> 리소스 사용률을 극대화하려면 특정 인스턴스 유형의 작업에 메모리 우선 순위를 지정합니다. 자세한 내용은 컴퓨팅 리소스 메모리 관리 단원을 참조하십시오.

e. (선택 사항) 환경 변수의 경우 환경 변수 추가를 선택하여 환경 변수를 이름-값 쌍으로 추가합니다. 이러한 변수는 컨테이너로 전달됩니다.

f. (선택 사항) 암호의 경우 암호 추가를 선택하여 암호를 이름-값 쌍으로 추가합니다. 이러한 보안 암호는 컨테이너에 노출됩니다. 자세한 내용은 LogConfiguration:secretOptions를 참조하세요.

g. 다음 페이지를 선택합니다.

13. (선택 사항) Linux 구성 섹션에서:

a. 사용자에서 컨테이너 내부에서 사용할 사용자 이름을 입력합니다.

b. 컨테이너 내에서 init 프로세스를 실행하려면 init 프로세스 활성화를 켭니다. 이 프로세스는 신호를 전달하고 결과를 받아들입니다.

c. 읽기 전용 파일 시스템 활성화를 켜서 볼륨에 대한 쓰기 권한을 제거합니다.

d. (선택 사항) 추가 구성을 확장합니다.

e. 마운트 포인트 구성에서 마운트 포인트 구성 추가를 선택하여 데이터 볼륨의 마운트 포인트를 추가합니다. 소스 볼륨과 컨테이너 경로를 지정해야 합니다. 이러한 마운트 포인트는 컨테이너 인스턴스의 Docker daemon에 전달됩니다.

f. 볼륨 구성에서 볼륨 추가를 선택하여 컨테이너에 전달할 볼륨 목록을 생성합니다. 볼륨의 이름 및 소스 경로를 입력한 다음 볼륨 추가를 선택합니다.

g. 작업 속성 섹션에서:

   i. 실행 역할 - 조건부에서 Amazon ECS 에이전트가 사용자를 대신하여 AWS API를 호출할 수 있도록 허용하는 역할을 선택합니다. 실행 역할 생성에 대한 자세한 내용은 자습서: IAM 실행 역할 생성 섹션을 참조하세요.

   ii. ECS 실행 명령을 선택하여 Amazon ECS 컨테이너 쉘에 대한 직접 액세스를 허용하고 호스트 OS를 우회합니다. 태스크 역할을 선택해야 합니다.

   > **Important**
   >
   > ECS 실행 명령을 사용하려면 파일 시스템에 대한 쓰기 권한이 있어야 합니다.

   iii. 작업 역할에서 Amazon ECS Identity and Access Management(IAM) 역할을 선택하여 컨테이너가 사용자를 대신하여 AWS API를 호출하도록 허용합니다. 자세한 내용은 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS 태스크 IAM 역할을 참조하세요.

h. 로깅 구성 섹션에서:

   i. (선택 사항)로그 드라이버에서 사용할 로그 드라이버를 선택합니다. 사용 가능한 로그 드라이버에 대한 자세한 내용은 LogConfiguration:logDriver를 참조하세요.

   > **Note**
   >
   > 기본적으로 `awslogs` 로그 드라이버가 사용됩니다.

ii. (선택 사항)옵션에서 옵션 추가를 선택하여 옵션을 추가합니다. 이름-값 쌍을 입력한 다음 옵션 추가를 선택합니다.

iii. (선택 사항) 암호에서 암호 추가를 선택하여 암호를 추가합니다. 이름-값 쌍을 입력한 다음 암호 추가를 선택합니다.

> **Tip**
>
> 자세한 내용은 LogConfiguration:secretOptions를 참조하세요.

14. 다음 페이지를 선택합니다.
15. 작업 정의 검토에서 구성 단계를 검토하세요. 변경해야 하는 경우 편집을 선택합니다 작업을 마쳤으면 작업 정의 생성을 선택합니다.

Amazon EKS 리소스에 단일 노드 작업 정의 생성

Amazon Elastic Kubernetes Service(Amazon EKS)에서 단일 노드 작업 정의를 생성하려면 다음 단계를 완료합니다.

Amazon EKS 리소스에 새 작업 정의를 생성하려면

1. https://console.aws.amazon.com/batch/ AWS Batch 콘솔을 엽니다.
2. 상단 탐색 모음에서 사용할 AWS 리전 를 선택합니다.
3. 왼쪽 탐색 창에서 작업 정의를 선택합니다.
4. 생성(Create)을 선택합니다.
5. 오케스트레이션 유형으로는 Elastic Kubernetes Service(EKS)를 선택합니다.
6. 이름(Name)에 고유한 작업 정의 이름을 입력합니다. 각 이름의 최대 길이는 128자입니다. 대문자 및 소문자, 숫자, 하이픈(-) 및 밑줄(_)을 포함할 수 있습니다.
7. (선택 사항) 실행 제한 시간에 제한 시간 값(초)을 입력합니다. 실행 제한 시간은 완료되지 않은 작업이 종료되기까지의 시간입니다. 시도가 제한 시간을 초과하면 중지되고 상태가 `FAILED`(으)로 변경됩니다. 자세한 내용은 작업 제한 시간 단원을 참조하십시오. 최솟값은 60초입니다.
8. (선택 사항) 예약 우선 순위를 켭니다. 0에서 100 사이의 예약 우선 순위 값을 입력합니다. 값이 높을수록 낮은 값보다 우선 순위가 높습니다.
9. (선택 사항) 태그를 확장한 다음, 태그 추가를 선택하여 리소스에 태그를 추가합니다.
10. 다음 페이지를 선택합니다.
11. EKS pod 속성 섹션에서:

a. 서비스 계정 이름에는 pod에서 실행되는 프로세스의 ID를 제공하는 계정을 입력합니다.

b. Kubernetes pod 네트워크 모델을 사용하려면 호스트 네트워크를 켜고 수신 연결을 위한 수신 포트를 엽니다. 발신 통신에만 이 설정을 끕니다.

c. DNS 정책에서 다음 중 하나를 선택합니다.

- 값 없음(null) - pod가 Kubernetes 환경의 DNS 설정을 무시합니다.
- 기본값 - pod가 실행 중인 노드의 이름 확인 구성을 상속합니다.

> **Note**
>
> DNS 정책이 지정되지 않은 경우 기본값은 기본 DNS 정책이 아닙니다. 대신 ClusterFirst가 사용됩니다.

- ClusterFirst - 구성된 클러스터 도메인 접미사와 일치하지 않는 모든 DNS 쿼리가 노드에서 상속된 업스트림 이름 서버로 전달됩니다.
- ClusterFirstWithHostNet – 호스트 네트워크가 켜져 있는 경우에 사용합니다.

d. (선택 사항) 볼륨에서 볼륨 추가를 선택한 후 다음을 수행합니다.

i. 볼륨의 이름을 추가합니다.

ii. (선택 사항) 호스트의 디렉터리에 대한 호스트 경로를 추가합니다.

iii. (선택 사항) 중간 및 크기 제한을 추가하여 Kubernetes emptyDir을 구성합니다.

iv. (선택 사항) 포드의 보안 암호 이름과 보안 암호가 선택 사항인지 여부를 제공합니다.

v. (선택 사항) Kubernetes 영구 볼륨 클레임을 포드에 연결할 클레임 이름을 정의하고 읽기 전용인지 여부를 정의합니다.

e. (선택 사항) 포드 레이블에서는 포드 레이블 추가를 선택한 다음 이름-값 쌍을 입력합니다.

> **Important**
>
> 포드 레이블의 접두사는`kubernetes.io/`, `k8s.io/` 또는 `batch.amazonaws.com/`을 포함할 수 없습니다.

f. (선택 사항) 포드 주석에서 주석 추가를 선택한 다음 이름-값 페어를 입력합니다.

> **Important**
>
> 포드 주석의 접두사는 `kubernetes.io/`, `k8s.io/` 또는 `batch.amazonaws.com/`을 포함할 수 없습니다.

g. 다음 페이지를 선택합니다.

h. 컨테이너 구성 섹션에서:

i. 이름에 컨테이너의 고유 이름을 입력합니다. 이름은 문자 또는 숫자로 시작되어야 하며 최대 63자여야 합니다. 소문자 및 대문자, 숫자, 하이픈(-)을 포함할 수 있습니다.

ii. 이미지에서 작업에 사용할 Docker 이미지를 선택합니다. 기본적으로 Docker Hub 레지스트리 내 이미지는 사용 가능합니다. 또한 *`repository-url`*`/`*`image`*`:`*`tag`*(을)를 사용하여 다른 리포지토리를 지정할 수도 있습니다. 이름은 최대 255자입니다. 대문자와 소문자, 숫자, 하이픈(-), 밑줄(_), 콜론(:), 마침표(.), 슬래시(/) 및 숫자 기호(#)를 포함할 수 있습니다. 이 파라미터는 Docker 원격 API의 컨테이너 생성 섹션에 있는 `Image`와 docker run의 `IMAGE` 파라미터로 매핑됩니다.

> **Note**
>
> 도커 이미지 아키텍처는 예정된 컴퓨팅 리소스의 프로세서 아키텍처와 일치해야 합니다. 예를 들어 ARM 기반 Docker 이미지는 ARM 기반 컴퓨팅 리소스에서만 실행할 수 있습니다.

- Amazon ECR Public 리포지토리에 있는 이미지는 전체 `registry/repository[:tag]` 또는 `registry/repository[@digest]` 명명 규칙을 사용합니다(예: `public.ecr.aws/`*`registry_alias`*`/`*`my-web-app`*`:`*`latest`*).
- Amazon ECR 리포지토리에 있는 이미지는 전체 `registry/repository[:tag]` 명명 규칙을 사용합니다 (예: *`aws_account_id`*`.dkr.ecr.`*`region`*`.amazonaws.com/`*`my-web-app`*`:`*`latest`*).
- Docker Hub의 공식 리포지토리 안의 이미지는 단일 이름을 사용합니다(예: `ubuntu` 또는 `mongo`).
- Docker Hub 다른 리포지토리에 저장된 이미지는 조직 이름으로 한정됩니다(예: `amazon/amazon-ecs-agent`).

- 다른 온라인 리포지토리 안의 이미지는 도메인 이름을 사용하여 추가로 한정됩니다 (예: `quay.io/assemblyline/ubuntu`).

iii. (선택 사항) 이미지 풀 정책에는 이미지를 가져오는 시기를 선택합니다.

iv. (선택 사항) 명령에는 컨테이너에 전달할 JSON 명령을 입력합니다.

v. (선택 사항) 인수에는 컨테이너로 전달할 인수를 입력합니다. 인수가 제공되지 않으면 컨테이너 이미지 명령이 사용됩니다.

i. (선택 사항) 작업 정의에 파라미터를 이름-값 매핑으로 추가하여 작업 정의 기본값을 재정의할 수 있습니다. 파라미터를 추가하려면

- 파라미터에 이름-값 쌍을 입력한 다음, 파라미터 추가를 선택합니다.

> **Important**
>
> 파라미터 추가를 선택한 경우 하나 이상의 파라미터를 구성하거나 파라미터 제거를 선택해야 합니다.

j. 환경 구성 섹션에서:

i. vCPU에서 컨테이너에 예약할 vCPU 수를 지정합니다. 이 파라미터는 Docker 원격 API(Docker Remote API)의 컨테이너 생성(Create a container) 섹션에 있는 `CpuShares`(와)과 docker run에 대한 `--cpu-shares` 옵션에 매핑됩니다. 각 vCPU는 1,024개의 CPU 공유와 동일합니다. vCPU를 최소 하나 이상 지정해야 합니다.

ii. 메모리에는 컨테이너에 사용할 수 있는 메모리 한도를 입력합니다. 컨테이너가 여기에 지정된 메모리를 초과하려 하면 해당 컨테이너가 중지됩니다. 이 파라미터는 Docker 원격 API(Docker Remote API)의 컨테이너 생성(Create a container) 섹션에 있는 `Memory`(와)과 docker run에 대한 `--memory` 옵션에 매핑됩니다. 한 작업에 대해 메모리를 최소한 4MiB 지정해야 합니다.

> **Note**
>
> 리소스 사용률을 극대화하려면 특정 인스턴스 유형의 작업에 메모리 우선 순위를 지정합니다. 자세한 내용은 컴퓨팅 리소스 메모리 관리 단원을 참조하십시오.

k. (선택 사항) 환경 변수의 경우 환경 변수 추가를 선택하여 환경 변수를 이름-값 쌍으로 추가합니다. 이러한 변수는 컨테이너로 전달됩니다.

l. (선택 사항) 볼륨 마운트에서

i. 볼륨 마운트 추가를 선택합니다.

ii. 이름을 입력한 다음, 볼륨이 마운트된 컨테이너의 마운트 경로를 입력합니다. SubPath를 입력하여 루트 대신 참조된 볼륨 내부에서 하위 경로를 지정합니다.

iii. 볼륨에 대한 쓰기 권한을 제거하려면 읽기 전용을 선택합니다.

iv. 볼륨 마운트 추가를 선택합니다.

m. (선택 사항) 사용자로 실행에 컨테이너 프로세스를 실행할 사용자 ID를 입력합니다.

> **Note**
>
> 컨테이너를 실행하려면 이미지에 사용자 ID가 있어야 합니다.

n. (선택 사항) 그룹으로 실행에 컨테이너 프로세스 런타임을 실행할 그룹 ID를 입력합니다.

> **Note**
>
> 컨테이너를 실행하려면 이미지에 그룹 ID가 있어야 합니다.

o. (선택 사항) 작업 컨테이너에 호스트 인스턴스에 대한 승격된 권한(`root` 사용자와 비슷함)을 부여하려면 권한이 있음 슬라이더를 오른쪽으로 드래그합니다. 이 파라미터는 Docker 원격 API(Docker Remote API)의 컨테이너 생성(Create a container) 섹션에 있는 `Privileged`(와)과 docker run에 대한 `--privileged` 옵션에 매핑됩니다.

p. (선택 사항) 루트 파일 시스템에 대한 쓰기 액세스를 제거하려면 읽기 전용 루트 파일 시스템을 켭니다.

q. (선택 사항) 루트가 아닌 사용자로 pod에서 컨테이너를 실행하려면 루트가 아닌 상태로 실행을 켭니다.

> **Note**
>
> 루트가 아닌 상태로 실행이 켜져 있는 경우, kubelet은 런타임 시 이미지의 유효성을 검사하여 이미지가 UID 0으로 실행되지 않는지 확인합니다.

r. 다음 페이지를 선택합니다.

12. 작업 정의 검토에서 구성 단계를 검토하세요. 변경해야 하는 경우 편집을 선택합니다 작업을 마쳤으면 작업 정의 생성을 선택합니다.

Amazon EC2 리소스에 다중 컨테이너에 대한 단일 노드 작업 정의 생성

Amazon Elastic Compute Cloud(Amazon EC2) 리소스에서 여러 컨테이너에 대해 단일 노드 작업 정의를 생성하려면 다음 단계를 완료합니다.

Amazon EC2 리소스에 새 작업 정의를 생성하려면

1. https://console.aws.amazon.com/batch/ AWS Batch 콘솔을 엽니다.
2. 탐색 모음에서 사용할 AWS 리전 를 선택합니다.
3. 왼쪽 탐색 창에서 작업 정의를 선택합니다.
4. 생성(Create)을 선택합니다.
5. 오케스트레이션 유형으로 Amazon Elastic Compute Cloud(Amazon EC2)를 선택합니다.
6. 작업 정의 구조에서 레거시 containerProperties 구조 사용 프로세싱을 끕니다.
7. EC2 플랫폼 구성의 경우 다중 노드 병렬 처리 활성화를 끕니다.
8. 다음을 선택합니다.
9. 일반 구성 섹션에 다음을 입력합니다.

    a. 이름(Name)에 고유한 작업 정의 이름을 입력합니다. 각 이름의 최대 길이는 128자입니다. 대문자 및 소문자, 숫자, 하이픈(-) 및 밑줄(_)을 포함할 수 있습니다.

    b. 실행 제한 시간 - 선택 사항에 제한 시간 값(초)을 입력합니다. 실행 제한 시간은 완료되지 않은 작업이 종료되기까지의 시간입니다. 시도가 제한 시간을 초과하면 중지되고 상태가 FAILED(으)로 변경됩니다. 자세한 내용은 작업 제한 시간 단원을 참조하십시오. 최솟값은 60초입니다.

    c. 예약 우선 순위 - 선택 사항을 켭니다. 0에서 100 사이의 예약 우선 순위 값을 입력합니다. 값이 높을수록 우선 순위가 높습니다.

    d. 태그 - 선택 사항을 확장한 다음 태그 추가를 선택하여 리소스에 태그를 추가합니다. 키와 선택 값을 입력하고 태그 추가를 선택합니다.

    e. 작업 및 작업 정의에서 Amazon ECS 태스크로 태그를 전파하려면 태그 전파를 켭니다.

10. 재시도 전략 - 선택 사항에서 다음을 입력합니다.

    a. 작업 시도에 AWS Batch 가 작업 상태를 RUNNABLE로 전환하기 시도하는 횟수를 입력합니다. 1~10 사이의 숫자를 입력합니다.

    b. 재시도 전략 조건의 경우 종료 시 평가 추가를 선택합니다. 파라미터 값을 하나 이상 입력한 다음 작업을 선택합니다. 각 조건 세트에 대해 작업을 재시도 또는 종료로 설정해야 합니다. 이러한 작업은 다음을 의미합니다.

- 재시도 - 지정한 작업 시도 횟수에 도달할 때까지 AWS Batch 재시도합니다.
- 종료 - 작업 재시도를 AWS Batch 중지합니다.

> **Important**
>
> 종료 시 평가 추가를 선택한 경우 하나 이상의 파라미터를 구성하고 작업을 선택하거나 종료 시 평가 제거를 선택해야 합니다.

11. 태스크 속성 섹션에서 다음을 입력합니다.

    a. 실행 역할 - 조건부에서 Amazon ECS 에이전트가 사용자를 대신하여 AWS API를 호출할 수 있도록 허용하는 역할을 선택합니다. 실행 역할 생성에 대한 자세한 내용은 자습서: IAM 실행 역할 생성 섹션을 참조하세요.

    b. ECS 실행 명령을 선택하여 Amazon ECS 컨테이너 쉘에 대한 직접 액세스를 허용하고 호스트 OS를 우회합니다. 태스크 역할을 선택해야 합니다.

    > **Important**
    >
    > ECS 실행 명령을 사용하려면 파일 시스템에 대한 쓰기 권한이 있어야 합니다.

    c. 태스크 역할에서 Amazon ECS Identity and Access Management(IAM) 역할을 선택하여 컨테이너가 사용자를 대신하여 AWS API를 호출하도록 허용합니다. 자세한 내용은 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS 태스크 IAM 역할을 참조하세요.

    d. IPC 모드에서 `host`, `task` 또는 `none`을 선택합니다. `host`를 지정하면 동일한 컨테이너 인스턴스에서 호스트 IPC 모드를 지정한 태스크 내 모든 컨테이너가 동일한 IPC 리소스를 호스트 Amazon EC2 인스턴스와 공유합니다. 태스크를 지정하면 지정된 태스크 내 모든 컨테이너가 동일한 IPC 리소스를 공유합니다. 아무 것도 지정되지 않은 경우, 태스크 컨테이너 내에 있는 IPC 리소스는 프라이빗이며, 태스크 또는 컨테이너 인스턴스의 다른 컨테이너와 공유되지 않습니다. 값을 지정하지 않을 경우, IPC 리소스 네임스페이스 공유는 컨테이너 인스턴스의 Docker 대몬 설정에 따라 달라집니다.

    e. PID 모드에서 `host` 또는 `task`를 선택합니다. 예를 들어 사이드카 모니터링에서는 동일한 작업에서 실행 중인 다른 컨테이너에 대한 정보에 액세스하기 위해 `pidMode`가 필요할 수 있습니다. `host`를 지정하면 동일한 컨테이너 인스턴스에서 PID 모드를 지정한 태스크 내 모든 컨테이너가 동일한 프로세스 네임스페이스를 호스트 Amazon EC2 인스턴스와 공유합니다. `task`을 지정하면 지정된 태스크 내 모든 컨테이너가 동일한 프로세스 네임스페이스를 공유합니다. 값을 지정하지 않을 경우, 기본값은 각 컨테이너의 프라이빗 네임스페이스입니다.

12. 사용 가능한 리소스 섹션에서 다음을 입력합니다.

    a. 고유한 이름과 요청된 값을 입력합니다.

    b. 소모성 리소스 추가를 선택하여 소모성 리소스를 더 추가할 수 있습니다.

13. 스토리지 섹션에서 다음을 입력합니다.

    a. 볼륨의 이름 및 소스 경로를 입력한 다음 볼륨 추가를 선택합니다. 또한 EFS 활성화를 켜도록 선택할 수도 있습니다.

    b. 볼륨 추가를 선택하여 볼륨을 더 추가할 수 있습니다.

14. 파라미터의 경우 파라미터 추가를 선택하여 파라미터 대체 플레이스홀더를 키 및 값(선택 사항) 페어로 추가합니다.

15. 다음 페이지를 선택합니다.

16. 컨테이너 구성 섹션에서:

    a. 이름(Name)에 컨테이너 이름을 입력합니다.

    b. 컨테이너가 필수인 경우 필수 컨테이너를 활성화합니다.

    c. 이미지에서 작업에 사용할 Docker 이미지를 선택합니다. 기본적으로 Docker Hub 레지스트리 내 이미지는 사용 가능합니다. 또한 *repository-url*/*image*:*tag*(을)를 사용하여 다른 리포지토리를 지정할 수도 있습니다. 각 이름의 최대 길이는 225자입니다. 여기에는 대문자와 소문자, 숫자, 하이픈(-), 밑줄(_), 콜론(:), 슬래시(/) 및 숫자 기호(#)를 사용할 수 있습니다. 이 파라미터는 Docker 원격 API(Docker Remote API)의 컨테이너 생성(Create a container) 섹션에 있는 `Image`(와)과 docker run의 `IMAGE` 파라미터로 매핑됩니다.

    > **Note**
    >
    > 도커 이미지 아키텍처는 예정된 컴퓨팅 리소스의 프로세서 아키텍처와 일치해야 합니다. 예를 들어 ARM 기반 Docker 이미지는 ARM 기반 컴퓨팅 리소스에서만 실행할 수 있습니다.

    - Amazon ECR Public 리포지토리에 있는 이미지는 전체 `registry/repository[:tag]` 또는 `registry/repository[@digest]` 명명 규칙을 사용합니다(예: `public.ecr.aws/`*`registry_alias`*`/`*`my-web-app`*`:`*`latest`*).
    - Amazon ECR 리포지토리에 있는 이미지는 전체 `registry/repository[:tag]` 명명 규칙을 사용합니다 (예: *`aws_account_id`*`.dkr.ecr.`*`region`*`.amazonaws.com/`*`my-web-app`*`:`*`latest`*).

- Docker Hub의 공식 리포지토리 안의 이미지는 단일 이름을 사용합니다(예: `ubuntu` 또는 `mongo`).
- Docker Hub 다른 리포지토리에 저장된 이미지는 조직 이름으로 한정됩니다(예: `amazon/amazon-ecs-agent`).
- 다른 온라인 리포지토리 안의 이미지는 도메인 이름을 사용하여 추가로 한정됩니다(예: `quay.io/assemblyline/ubuntu`).

d. 리소스 요구 사항에서 다음의 각 항목을 구성합니다.

i. vCPU에서 컨테이너의 CPU 수를 선택합니다.

ii. 메모리에서 컨테이너의 메모리 양을 선택합니다.

iii. GPU - 선택 사항에서 컨테이너에 대한 GPU 수를 선택합니다.

e. User(사용자)에서 컨테이너 내부에서 사용할 사용자 이름을 입력합니다.

f. 읽기 전용 파일 시스템 활성화를 켜서 볼륨에 대한 쓰기 권한을 제거합니다.

g. 권한 부여를 켜서 루트 사용자와 유사하게 호스트 인스턴스에서 작업 컨테이너에 승격된 권한을 부여합니다.

h. 명령에서 필드에 명령을 JSON 문자열 배열 형식으로 입력합니다.

이 파라미터는 도커 원격 API(Docker Remote API)의 컨테이너 생성(Create a container) 섹션에 있는 `Cmd`(와)과 docker run의 `COMMAND` 파라미터로 매핑됩니다. Docker `CMD` 파라미터에 대한 자세한 정보는 https://docs.docker.com/engine/reference/builder/#cmd를 참조하세요.

> **Note**
>
> 사용자는 명령에 파라미터 대체 및 자리 표시자 기본값을 사용할 수 있습니다. 자세한 내용은 Parameters 단원을 참조하십시오.

i. 리포지토리 자격 증명 - 선택 사항에서 자격 증명을 포함하는 보안 암호의 ARN을 입력합니다.

j. 환경 변수 - 선택 사항에서 환경 변수 추가를 선택하여 컨테이너에 전달할 환경 변수를 추가합니다.

k. Linux 파라마티 - 선택 사항 섹션에서:

i. 컨테이너 내에서 init 프로세스를 실행하려면 init 프로세스 활성화를 켭니다.

ii. 공유 메모리 크기에 /dev/shm 볼륨의 크기(MiB)를 입력합니다.

iii. 최대 스왑 크기에는 컨테이너가 사용할 수있는 총 스왑 메모리 양(MiB)을 입력합니다.

iv. 스왑 활용도의 경우 컨테이너의 스왑 동작을 나타내는 값을 0에서 100 사이의 값으로 입력합니다. 값을 지정하지 않고 스와핑이 활성화된 경우 기본값 60이 사용됩니다.

v. 디바이스의 경우 의 경우 를 선택하여 장치를 추가합니다.

A. 컨테이너 경로에 호스트 인스턴스에 매핑된 디바이스를 노출할 컨테이너 인스턴스의 경로를 지정합니다. 이 필드를 비워두면 호스트 경로가 컨테이너에 사용됩니다.

B. 호스트 경로에 호스트 인스턴스의 디바이스 경로를 지정합니다.

C. 권한에서 디바이스에 적용할 권한을 하나 이상 선택합니다. 사용 가능한 권한은 읽기, 쓰기 및 MKNOD입니다.

vi. Tmpfs의 경우 tmpfs 추가를 선택하여 `tmpfs` 마운트를 추가합니다.

l.

> **Note**
>
> Firelens 로깅은 전용 컨테이너에서 수행되어야 합니다. Firelens 로깅을 구성하려면:
>
> - 전용 Firelens 컨테이너를 제외한 모든 컨테이너에서 로깅 드라이버를 `awsfirelens`로 설정합니다.
> - Firelens 컨테이너에서 Firelens 구성 - 선택 사항 및 로깅 구성 - 선택 사항을 로깅 대상으로 설정합니다.

Firelens 구성 - 선택 사항 섹션에서:

> **Important**
>
> AWS Batch 는 비 MNP, 비 FARGATE Amazon ECS 작업에 `host` 네트워크 모드를 적용합니다. Amazon ECS Firelens에는 루트 사용자가 필요합니다. `host` 네트워크 모드를 사용하여 태스크를 실행할 때 Amazon ECS는 더 나은 보안을 위해 루트 사용자(UID 0)를 사용하여 컨테이너를 실행하지 않을 것을 권고합니다. 따라서 Firelens 로깅이 있는 모든 비 MNP, 비 FARGATE ECS 작업은 보안 모범 사례를 충족하지 않습니다.

i. 유형에서 `fluentd` 또는 `fluentbit`를 선택합니다.

ii. 옵션에 옵션의 이름/값 페어를 입력합니다. 추가 옵션을 사용하여 옵션을 더 추가할 수 있습니다.

m. 로깅 구성 - 선택 사항 섹션에서:

i. 로그 드라이버에서 사용할 로그 드라이버를 선택합니다. 사용 가능한 로그 드라이버에 대한 자세한 내용은 LogConfiguration:logDriver를 참조하세요.

> **Note**
>
> 기본적으로 `awslogs` 로그 드라이버가 사용됩니다.

ii. 옵션에서 옵션 추가를 선택하여 옵션을 추가합니다. 이름-값 쌍을 입력한 다음 옵션 추가를 선택합니다.

iii. 암호에서 암호 추가를 선택합니다. 이름-값 페어를 입력한 다음 암호 추가를 선택하여 암호를 추가합니다.

> **Tip**
>
> 자세한 내용은 LogConfiguration:secretOptions를 참조하세요.

n. 마운트 포인트 - 선택 사항에서 마운트 포인트 추가를 선택하여 데이터 볼륨의 마운트 포인트를 추가합니다. 소스 볼륨과 컨테이너 경로를 지정해야 합니다.

o. 보안 암호 - 선택 사항에서 보안 암호 추가를 선택하여 보안 암호를 추가합니다. 이름-값 쌍을 입력한 다음 암호 추가를 선택합니다.

> **Tip**
>
> 자세한 내용은 LogConfiguration:secretOptions를 참조하세요.

p. Ulimits - 선택 사항에서 ulimit 추가를 선택하여 컨테이너에 `ulimits` 값을 추가합니다. 이름, 소프트 제한, 하드 제한 값을 입력한 다음 ulimit 추가를 선택합니다.

q. 종속성 - 선택 사항에서 컨테이너 종속성 추가를 선택합니다. 컨테이너의 이름과 상태를 선택하여 언제 이 컨테이너가 시작되는지를 결정합니다.

17. 컨테이너가 하나만 구성된 경우에는 컨테이너 추가를 선택하고 새 컨테이너 구성을 완료해야 합니다. 그렇지 않으면 다음을 선택하여 검토합니다.

다중 노드 병렬 작업 정의 생성

에서 작업을 실행하려면 먼저 작업 정의를 생성 AWS Batch해야 합니다. 이 프로세스는 단일 노드 작업과 다중 노드 병렬 작업 간에 약간 다릅니다. 이 주제에서는 특히 AWS Batch 다중 노드 병렬 작업(그룹

일정이라고도 함)에 대한 작업 정의를 생성하는 방법을 다룹니다. 자세한 내용은 다중 노드 병렬 작업 단원을 참조하십시오.

> **Note**
>
> AWS Fargate는 다중 노드 병렬 작업을 지원하지 않습니다.

내용

- 자습서: Amazon EC2 리소스의 다중 노드 병렬 작업 정의 생성

자습서: Amazon EC2 리소스의 다중 노드 병렬 작업 정의 생성

Amazon Elastic Compute Cloud(Amazon EC2) 리소스에서 다중 노드 병렬 작업 정의를 생성합니다.

> **Note**
>
> 단일 노드 작업 정의를 생성하려면 Amazon EC2 리소스에 단일 노드 작업 정의 생성 섹션을 참조하세요.

Amazon EC2 리소스의 다중 노드 병렬 작업 정의를 생성하려면

1. https://console.aws.amazon.com/batch/ AWS Batch 콘솔을 엽니다.
2. 탐색 모음에서 사용할 AWS 리전 를 선택합니다.
3. 탐색 창에서 작업 정의를 선택합니다.
4. 생성을 선택합니다.
5. 오케스트레이션 유형으로 Amazon Elastic Compute Cloud(Amazon EC2)를 선택합니다.
6. 다중 노드 병렬 활성화에서 다중 노드 병렬을 켭니다.
7. 이름(Name)에 고유한 작업 정의 이름을 입력합니다. 이름은 최대 128자까지 포함할 수 있으며, 대문자와 소문자, 숫자, 하이픈(-), 밑줄(_)을 포함할 수 있습니다.
8. (선택 사항) 실행 제한 시간에서 작업 시도의 실행을 허용할 최대 시간(초)을 지정합니다. 시도가 제한 시간을 초과하면 중지되고 상태가 `FAILED`(으)로 변경됩니다. 자세한 내용은 작업 제한 시간 단원을 참조하십시오.
9. (선택 사항) 예약 우선 순위를 켭니다. 0에서 100 사이의 예약 우선 순위 값을 입력합니다. 값이 높을수록 낮은 값보다 우선 순위가 높습니다.

10. (선택 사항) 작업 시도의 경우 작업을 RUNNABLE 상태로 이동하려는 AWS Batch 시도 횟수를 입력합니다. 1~10 사이의 숫자를 입력합니다.
11. (선택 사항) 재시도 전략 조건의 경우 종료 시 평가 추가를 선택합니다. 파라미터 값을 하나 이상 입력한 다음 작업을 선택합니다. 각 조건 세트에 대해 작업을 재시도 또는 종료로 설정해야 합니다. 이러한 작업은 다음을 의미합니다.

    - 재시도 - 지정한 작업 시도 횟수에 도달할 때까지 AWS Batch 재시도합니다.
    - 종료 - 작업 재시도를 AWS Batch 중지합니다.

    > **Important**
    >
    > 종료 시 평가 추가를 선택한 경우 하나 이상의 파라미터를 구성하고 작업을 선택하거나 종료 시 평가 제거를 선택해야 합니다.

12. (선택 사항) 태그를 확장한 다음 태그 추가를 선택하여 리소스에 태그를 추가합니다. 태그 추가를 선택하고 키 및 값(선택 사항)을 입력합니다. 또한 작업 및 작업 정의에서 Amazon ECS 태스크로 태그를 전파하려면 태그 전파 시작을 켭니다.
13. 다음 페이지를 선택합니다.
14. 노드 수에 작업에 사용할 총 노드 수를 입력합니다.
15. 기본 노드에 메인 노드에 사용할 노드 인덱스를 입력합니다. 기본 노드 인덱스는 `0`입니다.
16. 인스턴스 유형에서 인스턴스 유형을 선택합니다.

    > **Note**
    >
    > 선택한 인스턴스 유형이 모든 노드에 적용됩니다.

17. 파라미터의 경우 파라미터 추가를 선택하여 파라미터 대체 플레이스홀더를 키 및 값(선택 사항) 페어로 추가합니다.
18. 노드 범위 섹션에서:

    a. 노드 범위 추가를 선택합니다. 그러면 노드 범위 섹션이 생성됩니다.

    b. 대상 노드에서 *range_start*:*range_end* 표기법을 사용하여 노드 그룹의 범위를 지정합니다.

    작업에 지정한 노드 수에 대해 최대 다섯 개의 노드 범위를 생성할 수 있습니다. 노드 범위는 노드의 인덱스 값을 사용하며 노드 인덱스는 0에서 시작됩니다. 최종 노드 그룹의 범위 끝 인

덱스 값이 지정한 노드 수보다 하나 작아야 합니다. 예를 들어, 10개의 노드를 지정하고 단일 노드 그룹을 사용하려고 한다고 가정해 보겠습니다. 그러면 최종 범위는 9입니다.

c. 이미지에서 작업에 사용할 Docker 이미지를 선택합니다. 기본적으로 Docker Hub 레지스트리 내 이미지는 사용 가능합니다. 또한 *repository-url*/*image*:*tag*(을)를 사용하여 다른 리포지토리를 지정할 수도 있습니다. 이름은 최대 225자입니다. 여기에는 대문자와 소문자, 숫자, 하이픈(-), 밑줄(_), 콜론(:), 슬래시(/) 및 숫자 기호(#)를 사용할 수 있습니다. 이 파라미터는 Docker 원격 API(Docker Remote API)의 컨테이너 생성(Create a container) 섹션에 있는 `Image`(와)과 docker run의 `IMAGE` 파라미터로 매핑됩니다.

> **Note**
>
> 도커 이미지 아키텍처는 예정된 컴퓨팅 리소스의 프로세서 아키텍처와 일치해야 합니다. 예를 들어 ARM 기반 Docker 이미지는 ARM 기반 컴퓨팅 리소스에서만 실행할 수 있습니다.

- Amazon ECR Public 리포지토리에 있는 이미지는 전체 `registry/repository[:tag]` 또는 `registry/repository[@digest]` 명명 규칙을 사용합니다(예: `public.ecr.aws/`*`registry_alias`*`/`*`my-web-app`*`:`*`latest`*).
- Amazon ECR 리포지토리에 있는 이미지는 전체 `registry/repository[:tag]` 명명 규칙을 사용합니다. 예: *`aws_account_id`*`.dkr.ecr.`*`region`*`.amazonaws.com/`*`my-web-app`*`:`*`latest`*
- Docker Hub의 공식 리포지토리 안의 이미지는 단일 이름을 사용합니다(예: `ubuntu` 또는 `mongo`).
- Docker Hub 다른 리포지토리에 저장된 이미지는 조직 이름으로 한정됩니다(예: `amazon/amazon-ecs-agent`).
- 다른 온라인 리포지토리 안의 이미지는 도메인 이름을 사용하여 추가로 한정됩니다(예: `quay.io/assemblyline/ubuntu`).

d. 명령에서 필드에 명령을 JSON 문자열 배열 형식으로 입력합니다.

이 파라미터는 도커 원격 API(Docker Remote API)의 컨테이너 생성(Create a container) 섹션에 있는 `Cmd`(와)과 docker run의 `COMMAND` 파라미터로 매핑됩니다. Docker `CMD` 파라미터에 대한 자세한 정보는 https://docs.docker.com/engine/reference/builder/#cmd를 참조하세요.

> **Note**
>
> 사용자는 명령에 파라미터 대체 및 자리 표시자 기본값을 사용할 수 있습니다. 자세한 내용은 Parameters 단원을 참조하십시오.

e. vCPU(vCPUs)에서 컨테이너에 예약할 vCPU 수를 지정합니다. 이 파라미터는 Docker 원격 API(Docker Remote API)의 컨테이너 생성(Create a container) 섹션에 있는 `CpuShares`(와)과 docker run에 대한 `--cpu-shares` 옵션에 매핑됩니다. 각 vCPU는 1,024개의 CPU 공유와 동일합니다. vCPU를 최소 하나 이상 지정해야 합니다.

f. 메모리에서 작업 컨테이너에 제공할 메모리의 하드 제한(MiB)을 지정합니다. 컨테이너가 여기에 지정된 메모리를 초과하려 하면 해당 컨테이너가 중지됩니다. 이 파라미터는 Docker 원격 API(Docker Remote API)의 컨테이너 생성(Create a container) 섹션에 있는 `Memory`(와)과 docker run에 대한 `--memory` 옵션에 매핑됩니다. 한 작업에 대해 메모리를 최소한 4MiB 지정해야 합니다.

> **Note**
>
> 리소스 사용률을 최대화하기 위해 특정 인스턴스 유형에 대해 작업에 최대한 많은 메모리를 제공할 수 있습니다. 자세한 내용은 컴퓨팅 리소스 메모리 관리 단원을 참조하십시오.

g. (선택 사항)GPU 수에서 작업에서 사용할 GPU 수를 지정합니다. 작업은 컨테이너에서 해당 컨테이너에 고정된 특정 GPU 수로 실행됩니다.

h. (선택 사항) 작업 역할에서 작업의 컨테이너에 AWS APIs. 이 기능은 태스크 기능에 Amazon ECS IAM 역할을 사용합니다. 자세한 내용은 Amazon Elastic Container Service 개발자 안내서에서 작업에 대한 IAM 역할 섹션을 참조하세요.

> **Note**
>
> Fargate 리소스에서 실행되는 작업의 경우, 작업 역할이 필요합니다.

> **Note**
>
> Amazon Elastic Container Service 태스크 역할 신뢰 관계를 보유한 역할만 여기 표시됩니다. AWS Batch 작업에 대한 IAM 역할 생성에 대한 자세한 내용은 Amazon

Elastic Container Service 개발자 안내서의 작업에 대한 IAM 역할 및 정책 생성을 참조하세요.

i. (선택 사항) 실행 역할에서 Amazon ECS 컨테이너 에이전트에게 사용자를 대신하여 AWS API를 호출할 수 있는 권한을 부여하는 IAM 역할을 지정합니다. 이 기능은 작업 기능에 Amazon ECS IAM 역할을 사용합니다. 자세한 내용을 알아보려면 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS 작업 실행 IAM 역할 섹션을 참조하세요.

19. (선택 사항) 추가 구성을 확장합니다.

a. 환경 변수에서 환경 변수 추가를 선택하여 환경 변수를 이름-값 쌍으로 추가합니다. 이러한 변수는 컨테이너로 전달됩니다.

b. 작업 역할 구성의 경우 작업의 컨테이너에 AWS APIs. 이 기능은 태스크 기능에 Amazon ECS IAM 역할을 사용합니다. 자세한 내용은 Amazon Elastic Container Service 개발자 안내서에서 작업에 대한 IAM 역할 섹션을 참조하세요.

> **Note**
>
> Fargate 리소스에서 실행되는 작업의 경우, 작업 역할이 필요합니다.

> **Note**
>
> Amazon Elastic Container Service 태스크 역할 신뢰 관계를 보유한 역할만 여기 표시됩니다. AWS Batch 작업에 대한 IAM 역할을 생성하는 자세한 내용은 Amazon Elastic Container Service 개발자 안내서에서 작업에 대한 IAM 역할 및 정책 생성 섹션을 참조하세요.

c. 실행 역할에서 Amazon ECS 컨테이너 에이전트에게 사용자를 대신하여 AWS API를 호출할 수 있는 권한을 부여하는 IAM 역할을 지정합니다. 이 기능은 작업 기능에 Amazon ECS IAM 역할을 사용합니다. 자세한 내용을 알아보려면 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS 작업 실행 IAM 역할 섹션을 참조하세요.

20. 보안 구성 섹션에서:

a. (선택 사항)작업 컨테이너에 호스트 인스턴스에 대한 승격된 권한(`root` 사용자와 비슷함)을 부여하려면 권한이 있음(Privileged)을 선택합니다. 이 파라미터는 Docker 원격 API(Docker Remote API)의 컨테이너 생성(Create a container) 섹션에 있는 `Privileged`(와)과 docker run에 대한 `--privileged` 옵션에 매핑됩니다.

b. (선택 사항)사용자에서 컨테이너 내부에서 사용할 사용자 이름을 입력합니다. 이 파라미터는 Docker 원격 API(Docker Remote API)의 컨테이너 생성(Create a container) 섹션에 있는 `User`(와)과 docker run에 대한 `--user` 옵션에 매핑됩니다.

c. (선택 사항) 암호의 경우 암호 추가를 선택하여 암호를 이름-값 쌍으로 추가합니다. 이러한 보안 암호는 컨테이너에 노출됩니다. 자세한 내용은 LogConfiguration:secretOptions를 참조하세요.

21. Linux 구성 섹션에서:

a. 읽기 전용 파일 시스템 활성화를 켜서 볼륨에 대한 쓰기 권한을 제거합니다.

b. (선택 사항)컨테이너 내에서 `init` 프로세스를 실행하려면 init 프로세스 활성화를 켭니다. 이 프로세스는 신호를 전달하고 결과를 받아들입니다.

c. 공유 메모리 크기에 `/dev/shm` 볼륨의 크기(MiB)를 입력합니다.

d. 최대 스왑 크기에는 컨테이너가 사용할 수있는 총 스왑 메모리 양(MiB)을 입력합니다.

e. 스왑 활용도의 경우 컨테이너의 스왑 동작을 나타내는 값을 0에서 100 사이의 값으로 입력합니다. 값을 지정하지 않고 스왑을 활성화한 경우 기본적으로 60으로 설정됩니다. 자세한 내용은 LinuxParameters:swappiness를 참조하세요.

f. (선택 사항) 디바이스의 경우 디바이스 추가를 선택하여 장치를 추가합니다.

i. 컨테이너 경로에 호스트 인스턴스에 매핑된 디바이스를 노출할 컨테이너 인스턴스의 경로를 지정합니다. 이 필드를 비워두면 호스트 경로가 컨테이너에 사용됩니다.

ii. 호스트 경로에 호스트 인스턴스의 디바이스 경로를 지정합니다.

iii. 권한에서 디바이스에 적용할 권한을 하나 이상 선택합니다. 사용 가능한 권한은 읽기, 쓰기 및 MKNOD입니다.

22. (선택 사항) 마운트 포인트의 경우 마운트 포인트 구성 추가를 선택하여 데이터 볼륨의 마운트 포인트를 추가합니다. 소스 볼륨과 컨테이너 경로를 지정해야 합니다. 이러한 마운트 포인트는 컨테이너 인스턴스의 Docker 대몬(daemon)으로 전달됩니다. 볼륨을 읽기 전용으로 설정할 수도 있습니다.

23. (선택 사항) Ulimits 구성의 경우 ulimit 추가를 선택하여 컨테이너에 `ulimits` 값을 추가합니다. 이름, 소프트 제한, 하드 제한 값을 입력한 다음 ulimit 추가를 선택합니다.

24. (선택 사항) 볼륨 구성의 경우 볼륨 추가를 선택하여 컨테이너에 전달할 볼륨 목록을 생성합니다. 볼륨의 이름 및 소스 경로를 입력한 다음 볼륨 추가를 선택합니다. 또한 EFS 활성화를 켜도록 선택할 수도 있습니다.

25. (선택 사항) Tmpfs의 경우 tmpfs 추가를 선택하여 `tmpfs` 마운트를 추가합니다.

26. 태스크 속성 섹션에서:

a. 실행 역할 - 조건부에서 Amazon ECS 에이전트가 사용자를 대신하여 AWS API를 호출할 수 있도록 허용하는 역할을 선택합니다. 실행 역할 생성에 대한 자세한 내용은 자습서: IAM 실행 역할 생성 섹션을 참조하세요.

b.

> **Important**
>
> ECS 실행 명령을 사용하려면 컴퓨팅 환경이 다중 노드 병렬 작업을 위한 컴퓨팅 환경 고려 사항을 충족해야 합니다.

ECS 실행 명령을 선택하여 Amazon ECS 컨테이너 쉘에 대한 직접 액세스를 허용하고 호스트 OS를 우회합니다. 태스크 역할을 선택해야 합니다.

> **Important**
>
> ECS 실행 명령을 사용하려면 파일 시스템에 대한 쓰기 권한이 있어야 합니다.

c. 작업 역할에서 Amazon ECS Identity and Access Management(IAM) 역할을 선택하여 컨테이너가 사용자를 대신하여 AWS API를 호출하도록 허용합니다. 자세한 내용은 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS 태스크 IAM 역할을 참조하세요.

27. (선택 사항)로깅 구성 섹션에서:

a. 로그 드라이버에서 사용할 로그 드라이버를 선택합니다. 사용 가능한 로그 드라이버에 대한 자세한 내용은 LogConfiguration:logDriver를 참조하세요.

> **Note**
>
> 기본적으로 `awslogs` 로그 드라이버가 사용됩니다.

b. 옵션에서 옵션 추가를 선택하여 옵션을 추가합니다. 이름-값 쌍을 입력한 다음 옵션 추가를 선택합니다.

c. 암호에서 암호 추가를 선택합니다. 이름-값 페어를 입력한 다음 암호 추가를 선택하여 암호를 추가합니다.

> **Tip**
>
> 자세한 내용은 LogConfiguration:secretOptions를 참조하세요.

28. 다음 페이지를 선택합니다.
29. 작업 정의 검토에서 구성 단계를 검토하세요. 변경해야 하는 경우 편집을 선택합니다 작업을 마쳤으면 작업 정의 생성을 선택합니다.

ContainerProperties를 사용하는 작업 정의 템플릿

다음은 단일 컨테이너가 포함된 빈 작업 정의 템플릿입니다. 이 템플릿을 사용하여 작업 정의를 생성한 다음, 파일에 저장하여 AWS CLI `--cli-input-json` 옵션과 함께 사용할 수 있습니다. 이러한 파라미터에 대한 자세한 내용은 JobDefinition을 참조하세요.

> **Note**
>
> 다음 AWS CLI 명령을 사용하여 단일 컨테이너 작업 정의 템플릿을 생성할 수 있습니다.
>
> ```
> $ aws batch register-job-definition --generate-cli-skeleton
> ```

```
{
    "jobDefinitionName": "",
    "type": "container",
    "parameters": {
        "KeyName": ""
    },
    "schedulingPriority": 0,
    "containerProperties": {
        "image": "",
        "vcpus": 0,
        "memory": 0,
        "command": [
            ""
        ],
        "jobRoleArn": "",
        "executionRoleArn": "",
        "volumes": [
            {
                "host": {
                    "sourcePath": ""
                },
                "name": "",
                "efsVolumeConfiguration": {
```

```
                "fileSystemId": "",
                "rootDirectory": "",
                "transitEncryption": "ENABLED",
                "transitEncryptionPort": 0,
                "authorizationConfig": {
                    "accessPointId": "",
                    "iam": "DISABLED"
                }
            }
        }
    ],
    "environment": [
        {
            "name": "",
            "value": ""
        }
    ],
    "mountPoints": [
        {
            "containerPath": "",
            "readOnly": true,
            "sourceVolume": ""
        }
    ],
    "readonlyRootFilesystem": true,
    "privileged": true,
    "ulimits": [
        {
            "hardLimit": 0,
            "name": "",
            "softLimit": 0
        }
    ],
    "user": "",
    "instanceType": "",
    "resourceRequirements": [
        {
            "value": "",
            "type": "MEMORY"
        }
    ],
    "linuxParameters": {
        "devices": [
            {
```

```
                "hostPath": "",
                "containerPath": "",
                "permissions": [
                    "WRITE"
                ]
            }
        ],
        "initProcessEnabled": true,
        "sharedMemorySize": 0,
        "tmpfs": [
            {
                "containerPath": "",
                "size": 0,
                "mountOptions": [
                    ""
                ]
            }
        ],
        "maxSwap": 0,
        "swappiness": 0
    },
    "logConfiguration": {
        "logDriver": "syslog",
        "options": {
            "KeyName": ""
        },
        "secretOptions": [
            {
                "name": "",
                "valueFrom": ""
            }
        ]
    },
    "secrets": [
        {
            "name": "",
            "valueFrom": ""
        }
    ],
    "networkConfiguration": {
        "assignPublicIp": "DISABLED"
    },
    "fargatePlatformConfiguration": {
        "platformVersion": ""
```

```
        }
    },
    "nodeProperties": {
        "numNodes": 0,
        "mainNode": 0,
        "nodeRangeProperties": [
            {
                "targetNodes": "",
                "container": {
                    "image": "",
                    "vcpus": 0,
                    "memory": 0,
                    "command": [
                        ""
                    ],
                    "jobRoleArn": "",
                    "executionRoleArn": "",
                    "volumes": [
                        {
                            "host": {
                                "sourcePath": ""
                            },
                            "name": "",
                            "efsVolumeConfiguration": {
                                "fileSystemId": "",
                                "rootDirectory": "",
                                "transitEncryption": "DISABLED",
                                "transitEncryptionPort": 0,
                                "authorizationConfig": {
                                    "accessPointId": "",
                                    "iam": "ENABLED"
                                }
                            }
                        }
                    ],
                    "environment": [
                        {
                            "name": "",
                            "value": ""
                        }
                    ],
                    "mountPoints": [
                        {
                            "containerPath": "",
```

```
                "readOnly": true,
                "sourceVolume": ""
            }
        ],
        "readonlyRootFilesystem": true,
        "privileged": true,
        "ulimits": [
            {
                "hardLimit": 0,
                "name": "",
                "softLimit": 0
            }
        ],
        "user": "",
        "instanceType": "",
        "resourceRequirements": [
            {
                "value": "",
                "type": "MEMORY"
            }
        ],
        "linuxParameters": {
            "devices": [
                {
                    "hostPath": "",
                    "containerPath": "",
                    "permissions": [
                        "WRITE"
                    ]
                }
            ],
            "initProcessEnabled": true,
            "sharedMemorySize": 0,
            "tmpfs": [
                {
                    "containerPath": "",
                    "size": 0,
                    "mountOptions": [
                        ""
                    ]
                }
            ],
            "maxSwap": 0,
            "swappiness": 0
```

```
                },
                "logConfiguration": {
                    "logDriver": "awslogs",
                    "options": {
                        "KeyName": ""
                    },
                    "secretOptions": [
                        {
                            "name": "",
                            "valueFrom": ""
                        }
                    ]
                },
                "secrets": [
                    {
                        "name": "",
                        "valueFrom": ""
                    }
                ],
                "networkConfiguration": {
                    "assignPublicIp": "DISABLED"
                },
                "fargatePlatformConfiguration": {
                    "platformVersion": ""
                }
            }
        }
    ]
},
"retryStrategy": {
    "attempts": 0,
    "evaluateOnExit": [
        {
            "onStatusReason": "",
            "onReason": "",
            "onExitCode": "",
            "action": "RETRY"
        }
    ]
},
"propagateTags": true,
"timeout": {
    "attemptDurationSeconds": 0
},
```

```
    "tags": {
        "KeyName": ""
    },
    "platformCapabilities": [
        "EC2"
    ],
    "eksProperties": {
        "podProperties": {
            "serviceAccountName": "",
            "hostNetwork": true,
            "dnsPolicy": "",
            "containers": [
                {
                    "name": "",
                    "image": "",
                    "imagePullPolicy": "",
                    "command": [
                        ""
                    ],
                    "args": [
                        ""
                    ],
                    "env": [
                        {
                            "name": "",
                            "value": ""
                        }
                    ],
                    "resources": {
                        "limits": {
                            "KeyName": ""
                        },
                        "requests": {
                            "KeyName": ""
                        }
                    },
                    "volumeMounts": [
                        {
                            "name": "",
                            "mountPath": "",
                            "readOnly": true
                        }
                    ],
                    "securityContext": {
```

```
                        "runAsUser": 0,
                        "runAsGroup": 0,
                        "privileged": true,
                        "readOnlyRootFilesystem": true,
                        "runAsNonRoot": true
                    }
                }
            ],
            "volumes": [
                {
                    "name": "",
                    "hostPath": {
                        "path": ""
                    },
                    "emptyDir": {
                        "medium": "",
                        "sizeLimit": ""
                    },
                    "secret": {
                        "secretName": "",
                        "optional": true
                    }
                }
            ]
        }
    }
}
```

ContainerProperties에 대한 작업 정의 파라미터

ContainerProperties를 사용하는 작업 정의는 여러 부분으로 나뉩니다.

- 작업 정의 이름
- 작업 정의의 유형
- 파라미터 대체 자리 표시자 기본값
- 작업에 대한 컨테이너 속성
- Amazon EKS 리소스에서 실행되는 작업에 필요한 작업 정의의 Amazon EKS 속성
- 다중 노드 병렬 작업에 필요한 노드 속성
- Fargate 리소스에서 작업을 실행하는 데 필요한 플랫폼 기능
- 작업 정의의 기본 태그 전파 세부 정보

- 작업 정의의 기본 재시도 전략
- 작업 정의의 기본 예약 우선 순위
- 작업 정의의 기본 태그
- 작업 정의의 기본 제한 시간

목차

작업 정의 이름

`jobDefinitionName`

작업 정의를 등록할 때 이름을 지정합니다. 각 이름의 최대 길이는 128자입니다. 대문자 및 소문자, 숫자, 하이픈(-) 및 밑줄(_)을 포함할 수 있습니다. 이 이름으로 등록된 첫 번째 작업 정의에는 버전 번호 1이 부여됩니다. 그런 다음 같은 이름으로 등록되는 작업 정의는 버전 번호가 하나씩 증가하여 부여됩니다.

유형: 문자열

필수 항목 여부: 예

유형

`type`

작업 정의를 등록할 때 작업 유형을 지정합니다. 작업이 Fargate 리소스에서 실행되는 경우 `multinode`는 지원되지 않습니다. 다중 노드 병렬 작업에 대한 자세한 내용은 다중 노드 병렬 작업 정의 생성 섹션을 참조하세요.

유형: 문자열

유효한 값: `container` | `multinode`

필수 여부: 예

Parameters

`parameters`

작업을 제출할 때 자리 표시자를 교체하거나 기본 작업 정의 파라미터를 재정의하는 파라미터를 지정할 수 있습니다. 작업 제출 요청의 파라미터는 작업 정의의 기본값보다 우선합니다. 즉, 동일한 형식을 사용하는 여러 작업에 동일한 작업 정의를 사용할 수 있습니다. 또한 제출 시 명령의 값을 프로그래밍 방식으로 변경할 수 있습니다.

유형: 문자열 대 문자열 맵

필수 여부: 아니요

작업 정의를 등록할 때 작업 컨테이너 속성의 command 필드에 파라미터 대체 자리 표시자를 사용할 수 있습니다. 구문은 다음과 같습니다.

```
"command": [
    "ffmpeg",
    "-i",
    "Ref::inputfile",
    "-c",
    "Ref::codec",
    "-o",
    "Ref::outputfile"
]
```

위의 예에서는 명령에 *Ref::inputfile*, *Ref::codec*, *Ref::outputfile* 파라미터 대체 자리 표시자가 있습니다. 작업 정의의 `parameters` 객체를 사용하여 이러한 자리 표시자의 기본값을 설

정할 수 있습니다. 예를 들어, *Ref::codec* 자리 표시자의 기본값을 설정하려면 작업 정의에 다음을 지정합니다.

```
"parameters" : {"codec" : "mp4"}
```

이 작업 정의를 실행하기 위해 제출하면 컨테이너 명령의 *Ref::codec* 인수가 기본값인 mp4로 대체됩니다.

컨테이너 속성

작업 정의를 등록할 때, 작업 배치 시 컨테이너 인스턴스의 도커 대몬(daemon)으로 전달되는 컨테이너 속성 목록을 지정합니다. 다음과 같은 컨테이너 속성을 작업 정의에 사용할 수 있습니다. 단일 노드 작업의 경우 이러한 컨테이너 속성은 작업 정의 수준에서 설정됩니다. 다중 노드 병렬 작업의 경우 컨테이너 속성은 각 노드 그룹에 대해 노드 속성 수준에서 설정됩니다.

`command`

컨테이너로 전달되는 명령입니다. 이 파라미터는 도커 원격 API(Docker Remote API)의 컨테이너 생성(Create a container) 섹션에 있는 Cmd(와)과 docker run의 COMMAND 파라미터로 매핑됩니다. 도커 CMD 파라미터에 대한 자세한 정보는 https://docs.docker.com/engine/reference/builder/#cmd를 참조하세요.

```
"command": ["string", ...]
```

유형: 문자열 배열

필수 여부: 아니요

`environment`

컨테이너로 전달할 환경 변수입니다. 이 파라미터는 Docker 원격 API의 컨테이너 생성 섹션에 있는 Env와 docker run에 대한 `--env` 옵션에 매핑됩니다.

> **Important**
>
> 자격 증명 데이터와 같은 민감한 정보에 대해서는 일반 텍스트 환경 변수를 사용하지 않는 것이 좋습니다.

> **Note**
>
> 환경 변수는 'AWS_BATCH'로 시작할 수 없습니다. 이 이름 지정 규칙은 AWS Batch 서비스가 설정하는 변수용으로 예약되어 있습니다.

유형: 키-값 쌍 배열

필수 여부: 아니요

`name`

환경 변수의 이름입니다.

유형: 문자열

필수 항목 여부: `environment` 사용 시, 예

`value`

환경 변수의 값입니다.

유형: 문자열

필수 항목 여부: `environment` 사용 시, 예

```
"environment" : [
    { "name" : "envName1", "value" : "envValue1" },
    { "name" : "envName2", "value" : "envValue2" }
]
```

`executionRoleArn`

작업 정의를 등록할 때 IAM 역할을 지정할 수 있습니다. 역할은 해당 정책에 지정된 API 작업을 호출할 수 있는 권한을 Amazon ECS 컨테이너 에이전트에 제공합니다. Fargate 리소스에서 실행되는 작업은 실행 역할을 제공해야 합니다. 자세한 내용은 AWS Batch IAM 실행 역할 섹션을 참조하세요.

유형: 문자열

필수 여부: 아니요

`fargatePlatformConfiguration`

Fargate 리소스에서 실행되는 작업에 대한 플랫폼 구성입니다. EC2 리소스에서 실행되는 작업에는 이 파라미터를 지정하지 않아야 합니다.

유형: FargatePlatformConfiguration 객체

필수 여부: 아니요

`platformVersion`

AWS Fargate 플랫폼 버전은 작업에 사용하거나 승인된 최신 버전의 AWS Fargate 플랫폼을 사용하는 데 LATEST가 사용됩니다.

유형: 문자열

기본값: LATEST

필수 여부: 아니요

`image`

작업을 시작하는 데 사용되는 이미지입니다. 이 문자열은 Docker 대몬으로 직접 전달됩니다. Docker Hub 레지스트리 내 이미지는 기본적으로 사용 가능합니다. 또한 *repository-url*/*image*:*tag*(을)를 사용하여 다른 리포지토리를 지정할 수도 있습니다. 최대 255개의 문자(대문자 및 소문자), 숫자, 하이픈, 밑줄, 콜론, 마침표, 슬래시 및 부호가 허용됩니다. 이 파라미터는 Docker 원격 API의 컨테이너 생성 섹션에 있는 Image와 docker run의 IMAGE 파라미터로 매핑됩니다.

> **Note**
>
> 도커 이미지 아키텍처는 예정된 컴퓨팅 리소스의 프로세서 아키텍처와 일치해야 합니다. 예를 들어, ARM 기반 도커 이미지는 ARM 기반 컴퓨팅 리소스에서만 실행될 수 있습니다.

- Amazon ECR Public 리포지토리에 있는 이미지는 전체 `registry/repository[:tag]` 또는 `registry/repository[@digest]` 명명 규칙을 사용합니다(예: `public.ecr.aws/`*registry_alias*`/`*my-web-app*`:`*latest*).
- Amazon ECR 리포지토리에 있는 이미지는 전체 `registry/repository:[tag]` 명명 규칙을 사용합니다. 예를 들면 *aws_account_id*`.dkr.ecr.`*region*`.amazonaws.com/`*my-web-app*`:`*latest*와 같습니다.
- Docker Hub 공식 리포지토리 안의 이미지는 단일 이름을 사용합니다(예: ubuntu 또는 mongo).

- Docker Hub의 다른 리포지토리에 저장된 이미지는 조직 이름으로 한정됩니다(예: `amazon/amazon-ecs-agent`).
- 다른 온라인 리포지토리 안의 이미지는 도메인 이름을 사용하여 추가로 한정됩니다(예: `quay.io/assemblyline/ubuntu`).

유형: 문자열

필수 항목 여부: 예

`instanceType`

다중 노드 병렬 작업에 사용할 인스턴스 유형입니다. 다중 노드 병렬 작업에 있는 모든 노드 그룹은 동일한 인스턴스 유형을 사용해야 합니다. 이 파라미터는 단일 노드 컨테이너 작업이나 Fargate 리소스에서 실행되는 작업에는 적용할 수 없습니다.

유형: 문자열

필수 여부: 아니요

`jobRoleArn`

작업 정의를 등록할 때 IAM 역할을 지정할 수 있습니다. 역할은 해당 정책에 지정된 API 작업을 호출할 수 있는 권한을 작업 컨테이너에 제공합니다. 자세한 내용은 Amazon Elastic Container Service 개발자 안내서의 태스크에 대한 IAM 역할을 참조하세요.

유형: 문자열

필수 여부: 아니요

`linuxParameters`

컨테이너에 적용되는 Linux 수정(예: 디바이스 매핑에 대한 세부 정보)

```
"linuxParameters": {
    "devices": [
        {
            "hostPath": "string",
            "containerPath": "string",
            "permissions": [
                "READ", "WRITE", "MKNOD"
            ]
        }
    ],
    "initProcessEnabled": true|false,
    "sharedMemorySize": 0,
```

```
    "tmpfs": [
        {
            "containerPath": "string",
            "size": integer,
            "mountOptions": [
                "string"
            ]
        }
    ],
    "maxSwap": integer,
    "swappiness": integer
}
```

유형: LinuxParameters 객체

필수 여부: 아니요

`devices`

컨테이너에 매핑되는 디바이스 목록 이 파라미터는 도커 원격 APIhttps://docs.docker.com/engine/api/v1.38/의 컨테이너 생성 섹션에 있는 `Devices` 및 docker run에 대한 `--device` 옵션에 매핑됩니다.

> **Note**
>
> 이 파라미터는 Fargate 리소스에서 실행되는 작업에는 적용되지 않습니다.

유형: Device 객체 배열

필수 여부: 아니요

`hostPath`

호스트 컨테이너 인스턴스에서 사용 가능한 디바이스가 있는 경로입니다.

유형: 문자열

필수 항목 여부: 예

`containerPath`

디바이스가 컨테이너에 노출되는 경로입니다. 지정하지 않으면 호스트 경로와 동일한 경로에 노출됩니다.

유형: 문자열

필수 여부: 아니요

`permissions`

컨테이너의 디바이스에 대한 권한입니다. 지정하지 않으면 권한이READ, WRITE, 및 MKNOD로 설정됩니다.

유형: 문자열 배열

필수 여부: 아니요

유효한 값: READ | WRITE | MKNOD

`initProcessEnabled`

true로 설정된 경우 신호를 전달하고 프로세스의 결과를 받아들이는 컨테이너 내에서 init 프로세스를 실행합니다. 이 파라미터는 docker run에 대한 --init 옵션에 매핑됩니다. 이 파라미터를 사용하려면 컨테이너 인스턴스에서 Docker 원격 API 버전 1.25 이상을 사용해야 합니다. 컨테이너 인스턴스의 도커 원격 API 버전을 확인하려면, 컨테이너 인스턴스에 로그인한 후 sudo docker version | grep "Server API version" 명령을 실행합니다.

유형: 부울

필수 여부: 아니요

`maxSwap`

작업이 사용할 수 있는 총 스왑 메모리 양(MiB)입니다. 이 파라미터는 docker run에 대한 --memory-swap 옵션으로 변환되며 컨테이너 메모리의 합계에 maxSwap 값을 더한 값이 됩니다. 자세한 내용을 알아보려면 도커 설명서의 --memory-swap 세부 정보를 참조하세요.

maxSwap의 0 값이 지정되면 컨테이너는 스왑을 사용하지 않습니다. 허용되는 값은 0 또는 양수입니다. maxSwap 파라미터를 생략하면 컨테이너는 실행되는 컨테이너 인스턴스에 대한 스왑 구성을 사용합니다. maxSwap 매개 변수를 사용하려면 swappiness 값을 설정해야 합니다.

> **Note**
>
> 이 파라미터는 Fargate 리소스에서 실행되는 작업에는 적용되지 않습니다.

유형: Integer

필수 여부: 아니요

`sharedMemorySize`

`/dev/shm` 볼륨의 크기 값(MiB)입니다. 이 파라미터는 docker run에 대한 `--shm-size` 옵션에 매핑됩니다.

> **Note**
>
> 이 파라미터는 Fargate 리소스에서 실행되는 작업에는 적용되지 않습니다.

유형: Integer

필수 여부: 아니요

`swappiness`

이를 통해 컨테이너의 메모리 스왑 동작을 조정할 수 있습니다. `0`의 `swappiness` 값은 절대적으로 필요한 경우가 아니면 스왑이 일어나지 않도록 합니다. `100`의 `swappiness` 값은 페이지가 적극적으로 스와핑되도록 합니다. 허용되는 값은 `0`과 `100` 사이의 숫자입니다. `swappiness` 파라미터를 지정하지 않으면 `60`의 기본값이 사용됩니다. `maxSwap` 값이 지정되지 않은 경우 이 파라미터는 무시됩니다. `maxSwap`가 0으로 설정된 경우 컨테이너는 스왑을 사용하지 않습니다. 이 파라미터는 docker run에 대한 `--memory-swappiness` 옵션에 매핑됩니다.

컨테이너별 스왑 구성을 사용하는 경우 다음을 고려하세요.

- 컨테이너를 사용하려면 컨테이너 인스턴스에서 스왑 공간을 활성화하고 할당해야 합니다.

  > **Note**
  >
  > Amazon ECS에 최적화된 AMI에는 기본적으로 스왑이 활성화되어 있지 않습니다. 이 기능을 사용하려면 인스턴스에서 스왑을 활성화해야합니다. 자세한 내용은 Amazon EC2사용자 안내서의 인스턴스 스토어 스왑 볼륨 또는 스왑 파일을 사용하여 Amazon EC2 인스턴스에서 스왑 스페이스로 작동하도록 메모리를 할당하려면 어떻게 해야 하나요?를 참조하세요.

- 스왑 공간 파라미터는 EC2 리소스를 사용하는 작업 정의에 대해서만 지원됩니다.
- `maxSwap` 및 `swappiness` 파라미터가 작업 정의에서 생략된 경우 각 컨테이너의 기본 `swappiness` 값은 60입니다. 총 스왑 사용량은 컨테이너 메모리 예약의 두 배로 제한됩니다.

> **Note**
> 이 파라미터는 Fargate 리소스에서 실행되는 작업에는 적용되지 않습니다.

유형: Integer

필수 여부: 아니요

`tmpfs`

tmpfs 마운트의 컨테이너 경로, 마운트 옵션 및 크기입니다.

유형: Tmpfs 객체 배열

> **Note**
> 이 파라미터는 Fargate 리소스에서 실행되는 작업에는 적용되지 않습니다.

필수 여부: 아니요

`containerPath`

tmpfs 볼륨을 마운트할 컨테이너의 절대 파일 경로입니다.

유형: 문자열

필수 항목 여부: 예

`mountOptions`

tmpfs 볼륨 마운트 옵션의 목록입니다.

유효한 값: `"defaults" | "ro" | "rw" | "suid" | "nosuid" | "dev" | "nodev" | "exec" | "noexec" | "sync" | "async" | "dirsync" | "remount" | "mand" | "nomand" | "atime" | "noatime" | "diratime" | "nodiratime" | "bind" | "rbind" | "unbindable" | "runbindable" | "private" | "rprivate" | "shared" | "rshared" | "slave" | "rslave" | "relatime" | "norelatime" | "strictatime" | "nostrictatime" | "mode" | "uid" | "gid" | "nr_inodes" | "nr_blocks" | "mpol"`

유형: 문자열 배열

필수 여부: 아니요

`size`

tmpfs 볼륨의 크기(MiB)입니다.

유형: 정수

필수 여부: 예

`logConfiguration`

작업의 로그 구성 사양입니다.

이 파라미터는 도커 원격 APIhttps://docs.docker.com/engine/api/v1.38/의 컨테이너 생성 섹션에 있는 `LogConfig` 및 docker run에 대한 `--log-driver` 옵션에 매핑됩니다. 기본적으로 컨테이너는 도커 대몬이 사용하는 것과 동일한 로깅 드라이버를 사용합니다. 하지만 컨테이너는 이 파라미터를 사용하여 컨테이너 정의에 로그 드라이버를 지정함으로써 도커 대몬(daemon)과 다른 로깅 드라이버를 사용할 수 있습니다. 컨테이너에 다른 로깅 드라이버를 사용하려면 컨테이너 인스턴스 또는 원격 로그 서버에 로그 시스템이 올바르게 구성되어야 원격 로깅 옵션을 제공할 수 있습니다. 지원되는 다양한 로그 드라이버 옵션에 대한 자세한 정보는 Docker 설명서의 로깅 드라이버 구성을 참조하세요.

> **Note**
>
> 현재 AWS Batch는 도커 대몬(daemon)에서 사용 가능한 로깅 드라이버의 하위 세트를 지원합니다(LogConfiguration 데이터 유형에 표시됨).

이 파라미터를 사용하려면 컨테이너 인스턴스에서 Docker 원격 API 버전 1.18 이상을 사용해야 합니다. 컨테이너 인스턴스의 도커 원격 API 버전을 확인하려면, 컨테이너 인스턴스에 로그인한 후 `sudo docker version | grep "Server API version"` 명령을 실행합니다.

```
"logConfiguration": {
    "devices": [
        {
            "logDriver": "string",
            "options": {
                "optionName1" : "optionValue1",
                "optionName2" : "optionValue2"
            }
            "secretOptions": [
```

```
            {
                "name" : "secretOptionName1",
                "valueFrom" : "secretOptionArn1"
            },
            {
                "name" : "secretOptionName2",
                "valueFrom" : "secretOptionArn2"
            }
        ]
    }
]
}
```

유형: LogConfiguration 객체

필수 여부: 아니요

`logDriver`

작업에 사용할 로그 드라이버입니다. 기본적으로 AWS Batch는 `awslogs` 로그 드라이버를 활성화합니다. 이 파라미터에 대해 나열된 유효한 값은 Amazon ECS 컨테이너 에이전트가 기본적으로 통신할 수 있는 로그 드라이버입니다.

이 파라미터는 도커 원격 APIhttps://docs.docker.com/engine/api/v1.38/의 컨테이너 생성 섹션에 있는 `LogConfig` 및 docker run에 대한 `--log-driver` 옵션에 매핑됩니다. 기본적으로 작업은 도커 대몬(daemon)이 사용하는 것과 동일한 로깅 드라이버를 사용합니다. 하지만 작업은 이 파라미터를 사용하여 작업 정의에 로그 드라이버를 지정함으로써 도커 대몬(daemon)과 다른 로깅 드라이버를 사용할 수 있습니다. 작업에 다른 로깅 드라이버를 지정하려면 컴퓨팅 환경의 컨테이너 인스턴스에 로그 시스템이 구성되어야 합니다. 또는 원격 로깅 옵션을 제공하도록 다른 로그 서버에 구성할 수도 있습니다. 지원되는 다양한 로그 드라이버 옵션에 대한 자세한 정보는 Docker 설명서의 로깅 드라이버 구성을 참조하세요.

> **Note**
>
> 현재 AWS Batch는 Docker 대몬(daemon)에서 사용 가능한 로깅 드라이버의 하위 집합을 지원합니다. 향후의 Amazon ECS 컨테이너 에이전트 릴리스에서 로그 드라이버가 추가될 예정입니다.

지원되는 로그 드라이버는 `awslogs`, `fluentd`, `gelf`, `json-file`, `journald`, `logentries`, `syslog`, `splunk`입니다.

> **Note**
>
> Fargate 리소스에서 실행되는 작업은 `awslogs` 및 `splunk` 로그 드라이버로 제한됩니다.

이 파라미터를 사용하려면 컨테이너 인스턴스에서 Docker 원격 API 버전 1.18 이상을 사용해야 합니다. 컨테이너 인스턴스의 도커 원격 API 버전을 확인하려면, 컨테이너 인스턴스에 로그인한 후 `sudo docker version | grep "Server API version"` 명령을 실행합니다.

> **Note**
>
> 컨테이너 인스턴스에서 실행되는 Amazon ECS 컨테이너 에이전트는 `ECS_AVAILABLE_LOGGING_DRIVERS` 환경 변수를 사용하여 해당 인스턴스에서 사용 가능한 로깅 드라이버를 등록해야 합니다. 그렇지 않으면 해당 인스턴스에 배치된 컨테이너가 이러한 로그 구성 옵션을 사용할 수 없습니다. 자세한 내용을 알아보려면 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS 컨테이너 에이전트 구성을 참조하세요.

`awslogs`

Amazon CloudWatch Logs 로깅 드라이버를 지정합니다. 자세한 내용은 awslogs 로그 드라이버 사용 섹션 및 도커 설명서의 Amazon CloudWatch Logs 로깅 드라이버를 참조하세요.

`fluentd`

Fluentd 로깅 드라이버를 지정합니다. 사용법 및 옵션을 포함한 자세한 내용은 도커 설명서의 Fluentd 로깅 드라이버를 참조하세요.

`gelf`

GELF(Graylog Extended Format) 로깅 드라이버를 지정합니다. 사용법 및 옵션을 포함한 자세한 내용은 도커 설명서의 Graylog Extended Format 로깅 드라이버를 참조하세요.

`journald`

저널드 로깅 드라이버를 지정합니다. 사용법 및 옵션을 포함한 자세한 내용은 도커 설명서의 Journald 로깅 드라이버를 참조하세요.

`json-file`

JSON 파일 로깅 드라이버를 지정합니다. 사용법 및 옵션을 포함한 자세한 내용은 도커 설명서의 JSON 파일 로깅 드라이버를 참조하세요.

`splunk`

Splunk 로깅 드라이버를 지정합니다. 사용법 및 옵션을 포함한 자세한 내용은 도커 설명서의 Splunk 로깅 드라이버를 참조하세요.

`syslog`

syslog 로깅 드라이버를 지정합니다. 사용법 및 옵션을 포함한 자세한 내용은 도커 설명서의 Syslog 로깅 드라이버를 참조하세요.

유형: 문자열

필수 항목 여부: 예

유효한 값: `awslogs` | `fluentd` | `gelf` | `journald` | `json-file` | `splunk` | `syslog`

> **Note**
>
> 위 목록에는 포함되지 않았지만 Amazon ECS 컨테이너 에이전트와 함께 사용하려는 사용자 지정 드라이버가 있는 경우 GitHub에서 사용 가능한 Amazon ECS 컨테이너 에이전트 프로젝트를 해당 드라이버와 함께 작동하도록 사용자 지정할 수 있습니다. 포함하고 싶은 변경에 대해서는 풀 요청을 제출할 것을 권장합니다. 하지만 Amazon Web Services는 현재 이 소프트웨어의 변경된 사본을 실행하는 요청을 지원하지 않습니다.

`options`

작업의 지정 로그 드라이버에 전송하는 로그 구성 옵션입니다.

이 파라미터를 사용하려면 컨테이너 인스턴스에서 Docker 원격 API 버전 1.19 이상을 사용해야 합니다.

유형: 문자열 대 문자열 맵

필수 여부: 아니요

`secretOptions`

로그 구성에 전달할 암호를 나타내는 객체입니다. 자세한 내용은 민감한 데이터 지정 섹션을 참조하세요.

유형: 객체 배열

필수 여부: 아니요

`name`

작업에서 설정할 로그 드라이버 옵션의 이름입니다.

유형: 문자열

필수 항목 여부: 예

`valueFrom`

컨테이너의 로그 구성에 노출할 암호의 Amazon 리소스 이름(ARN)입니다. 지원되는 값은 Secrets Manager 암호의 전체 ARN이거나 혹은 SSM Parameter Store 내 파라미터의 전체 ARN입니다.

> **Note**
>
> SSM Parameter Store 파라미터가 현재 실행 중인 태스크와 동일한 AWS 리전에 있는 경우 파라미터의 전체 ARN 또는 이름을 사용할 수 있습니다. 파라미터가 다른 리전에 있다면 전체 ARN을 지정해야 합니다.

유형: 문자열

필수 항목 여부: 예

`memory`

이 파라미터는 더 이상 사용되지 않으므로 대신 *resourceRequirements*를 사용합니다.

작업에 예약된 메모리의 MiB 수입니다.

resourceRequirements를 사용하는 방법의 예로, 작업 정의에 다음과 유사한 구문이 포함되어 있는지 살펴보겠습니다.

```
"containerProperties": {
  "memory": 512
}
```

resourceRequirements를 사용하는 동등한 구문은 다음과 같습니다.

```
"containerProperties": {
  "resourceRequirements": [
    {
      "type": "MEMORY",
      "value": "512"
    }
  ]
}
```

유형: 정수

필수 여부: 예

`mountPoints`

컨테이너에서 데이터 볼륨의 탑재 지점입니다. 이 파라미터는 Docker 원격 API의 컨테이너 생성 섹션에 있는 `Volumes`와 docker run에 대한 `--volume` 옵션에 매핑됩니다.

```
"mountPoints": [
                {
                  "sourceVolume": "string",
                  "containerPath": "string",
                  "readOnly": true|false
                }
              ]
```

유형: 객체 배열

필수 여부: 아니요

`sourceVolume`

탑재할 볼륨의 이름입니다.

유형: 문자열

필수 항목 여부: `mountPoints` 사용 시, 예

`containerPath`

호스트 볼륨을 마운트할 컨테이너의 경로입니다.

유형: 문자열

필수 항목 여부: mountPoints 사용 시, 예

readOnly

이 값이 true일 경우 컨테이너에는 볼륨에 대한 읽기 전용 액세스가 부여됩니다. 이 값이 false일 경우 컨테이너는 볼륨에 쓸 수 있습니다.

유형: 부울

필수 여부: 아니요

기본값: False

networkConfiguration

Fargate 리소스에서 실행되는 작업에 대한 네트워크 구성입니다. EC2 리소스에서 실행되는 작업에는 이 파라미터를 지정하지 않아야 합니다.

```
"networkConfiguration": {
   "assignPublicIp": "string"
}
```

유형: 객체 배열

필수 여부: 아니요

assignPublicIp

작업에 퍼블릭 IP 주소가 있는지 여부를 나타냅니다. 이는 작업에 아웃바운드 네트워크 액세스가 필요한 경우 필요합니다.

유형: 문자열

유효한 값: ENABLED | DISABLED

필수 여부: 아니요

기본값: DISABLED

privileged

이 파라미터가 true일 경우 컨테이너에는 호스트 컨테이너 인스턴스에 대한 승격된 권한을 부여받습니다(root 사용자와 비슷함). 이 파라미터는 Docker 원격 API(Docker Remote API)의 컨테이너

생성(Create a container) 섹션에 있는 Privileged(와)과 docker run에 대한 --privileged 옵션에 매핑됩니다. 이 파라미터는 Fargate 리소스에서 실행되는 작업에는 적용되지 않습니다. 입력하지 않거나 false로 지정합니다.

```
"privileged": true|false
```

유형: 부울

필수 여부: 아니요

readonlyRootFilesystem

이 파라미터가 true일 경우 컨테이너에는 루트 파일 시스템에 대한 읽기 전용 액세스가 부여됩니다. 이 파라미터는 Docker 원격 API의 컨테이너 생성 섹션에 있는 ReadonlyRootfs와 docker run에 대한 --read-only 옵션에 매핑됩니다.

```
"readonlyRootFilesystem": true|false
```

유형: 부울

필수 여부: 아니요

resourceRequirements

컨테이너에 할당할 리소스의 유형 및 양입니다. 지원되는 리소스에는 GPUMEMORY 및 VCPU이 포함됩니다.

```
"resourceRequirements" : [
  {
    "type": "GPU",
    "value": "number"
  }
]
```

유형: 객체 배열

필수 여부: 아니요

type

컨테이너에 할당할 리소스 유형입니다. 지원되는 리소스에는 GPUMEMORY 및 VCPU이 포함됩니다.

유형: 문자열

필수 항목 여부: `resourceRequirements` 사용 시, 예

`value`

컨테이너에 대해 예약할 지정된 자원의 수량입니다. 값은 지정된 `type`에 따라 다릅니다.

type="GPU"

컨테이너용으로 예약할 물리적 GPU의 개수입니다. 작업의 모든 컨테이너에 예약된 GPU 수는 작업이 실행되는 컴퓨팅 리소스에서 사용할 수 있는 GPU 수를 초과할 수 없습니다.

type="MEMORY"

컨테이너에 표시할 메모리의 하드 제한(MiB)입니다. 컨테이너가 여기서 지정된 메모리를 초과하려 하면 해당 컨테이너가 중지됩니다. 이 파라미터는 도커 원격 APIhttps://docs.docker.com/engine/api/v1.38/의 컨테이너 생성 섹션에 있는 `Memory` 및 docker run에 대한 `--memory` 옵션에 매핑됩니다. 한 작업에 대해 메모리를 최소한 4MiB 지정해야 합니다. 이것은 필수이지만 다중 노드 병렬(MNP) 작업의 경우 여러 위치에서 지정할 수 있습니다. 각 노드에 대해 한 번 이상 지정해야 합니다. 이 파라미터는 도커 원격 APIhttps://docs.docker.com/engine/api/v1.38/의 컨테이너 생성 섹션에 있는 `Memory` 및 docker run에 대한 `--memory` 옵션에 매핑됩니다.

> **Note**
>
> 특정 인스턴스 유형에 대해 작업에 가능한 한 많은 메모리를 제공하여 리소스 사용률을 최대화하려는 경우 컴퓨팅 리소스 메모리 관리 섹션을 참조하세요.

Fargate 리소스에서 실행되는 작업의 경우 `value`는 지원되는 값 중 하나와 일치해야 합니다. 또한 `VCPU` 값은 해당 메모리 값에 지원되는 값 중 하나여야 합니다.

VCPU	MEMORY
0.25 vCPU	512, 1024, 2048MiB
0.5 vCPU	1024~4096 MiB(1024MiB 증분)
1 vCPU	2048~8192MiB(1024MiB 증분)
2 vCPU	4096~16384MiB(1024MiB 증분)

VCPU	MEMORY
4 vCPU	8192~30720MiB(1024MiB 증분)
8 vCPU	16384~61440MiB(4096MiB 증분)
16 vCPU	32768~122880MiB(8192MiB 증분)

type="VCPU"

작업에 예약된 vCPU 개수입니다. 이 파라미터는 도커 원격 API의 컨테이너 생성 섹션에 있는 CpuShares 및 docker run에 대한 --cpu-shares 옵션에 매핑됩니다. 각 vCPU는 1,024개의 CPU 공유와 동일합니다. EC2 리소스에서 실행되는 작업의 경우 하나 이상의 vCPU를 지정해야 합니다. 이것은 필수이지만 여러 위치에서 지정할 수 있습니다. 각 노드에 대해 한 번 이상 지정해야 합니다.

Fargate 리소스에서 실행되는 작업의 경우 value는 지원되는 값 중 하나와 일치해야 하며 MEMORY 값은 해당 VCPU 값에 대해 지원되는 값 중 하나여야 합니다. 지원되는 값은 0.25, 0.5, 1, 2, 4, 8 및 16입니다.

Fargate 온디맨드 vCPU 리소스 수 할당량의 기본값은 vCPU 6개입니다. Fargate 할당량에 대한 자세한 내용은 Amazon Web Services 일반 참조의 AWS Fargate 할당량을 참조하세요.

유형: 문자열

필수 항목 여부: resourceRequirements 사용 시, 예

secrets

환경 변수로 노출되는 작업의 암호입니다. 자세한 내용은 민감한 데이터 지정 섹션을 참조하세요.

```
"secrets": [
    {
      "name": "secretName1",
      "valueFrom": "secretArn1"
    },
    {
      "name": "secretName2",
      "valueFrom": "secretArn2"
    }
```

```
    ...
]
```

유형: 객체 배열

필수 여부: 아니요

`name`

암호가 포함된 환경 변수의 이름입니다.

유형: 문자열

필수 항목 여부: `secrets` 사용 시, 예

`valueFrom`

컨테이너에 노출될 암호입니다. 지원되는 값은 Secrets Manager 암호의 전체 Amazon 리소스 이름(ARN)이거나 혹은 SSM Parameter Store 내 파라미터의 전체 ARN입니다.

> **Note**
>
> SSM Parameter Store 파라미터가 시작 중인 작업과 동일한 AWS 리전에 있는 경우 파라미터의 전체 ARN 또는 이름을 사용할 수 있습니다. 파라미터가 다른 리전에 있다면 전체 ARN을 지정해야 합니다.

유형: 문자열

필수 항목 여부: `secrets` 사용 시, 예

`ulimits`

컨테이너에 설정할 `ulimits` 값의 목록입니다. 이 파라미터는 Docker 원격 API의 컨테이너 생성 섹션에 있는 `Ulimits`와 docker run에 대한 `--ulimit` 옵션에 매핑됩니다.

```
"ulimits": [
  {
    "name": string,
    "softLimit": integer,
    "hardLimit": integer
  }
```

```
  ...
]
```

유형: 객체 배열

필수 여부: 아니요

`name`

`type`의 `ulimit`입니다.

유형: 문자열

필수 항목 여부: `ulimits` 사용 시, 예

`hardLimit`

`ulimit` 유형의 하드 제한입니다.

유형: 정수

필수 항목 여부: `ulimits` 사용 시, 예

`softLimit`

`ulimit` 유형의 소프트 제한입니다.

유형: 정수

필수 항목 여부: `ulimits` 사용 시, 예

`user`

컨테이너 내부에서 사용할 사용자 이름입니다. 이 파라미터는 Docker 원격 API의 컨테이너 생성 섹션에 있는 `User`와 docker run에 대한 `--user` 옵션에 매핑됩니다.

```
"user": "string"
```

유형: 문자열

필수 여부: 아니요

vcpus

이 파라미터는 더 이상 사용되지 않으므로 대신 *resourceRequirements*를 사용합니다.

컨테이너에 예약된 vCPU 개수입니다.

resourceRequirements를 사용하는 방법의 예로, 작업 정의에 다음과 비슷한 줄이 포함되어 있는 경우

```
"containerProperties": {
  "vcpus": 2
}
```

resourceRequirements를 사용하는 동등한 줄은 다음과 같습니다.

```
"containerProperties": {
  "resourceRequirements": [
    {
      "type": "VCPU",
      "value": "2"
    }
  ]
}
```

유형: 정수

필수 여부: 예

volumes

작업 정의를 등록할 때 컨테이너 인스턴스의 도커 대몬(daemon)으로 전달되는 볼륨 목록을 지정할 수 있습니다. 컨테이너 속성에는 다음 파라미터를 사용할 수 있습니다.

```
"volumes": [
  {
    "name": "string",
    "host": {
      "sourcePath": "string"
    },
    "efsVolumeConfiguration": {
      "authorizationConfig": {
        "accessPointId": "string",
```

```
        "iam": "string"
      },
      "fileSystemId": "string",
      "rootDirectory": "string",
      "transitEncryption": "string",
      "transitEncryptionPort": number
    }
  }
]
```

name

볼륨의 이름입니다. 최대 255개의 문자(대문자 및 소문자), 숫자, 하이픈 및 밑줄이 허용됩니다. 이 이름은 컨테이너 정의 mountPoints의 sourceVolume 파라미터에서 참조됩니다.

유형: 문자열

필수 여부: 아니요

host

host 파라미터의 콘텐츠는 데이터 볼륨이 호스트 컨테이너 인스턴스에서 지속되는지 여부와 저장 위치를 결정합니다. host 파라미터가 비어 있으면 도커 대몬(daemon)이 데이터 볼륨의 호스트 경로를 할당합니다. 그러나 연결된 컨테이너의 실행이 중지된 후에는 데이터 유지가 보장되지 않습니다.

> **Note**
>
> 이 파라미터는 Fargate 리소스에서 실행되는 작업에는 적용되지 않습니다.

유형: 객체

필수 여부: 아니요

sourcePath

컨테이너에 제공되는 호스트 컨테이너 인스턴스의 경로입니다. 이 파라미터가 비어 있으면 Docker 대몬이 사용자 대신 호스트 경로를 할당합니다.

host 파라미터에 sourcePath 파일 위치가 들어 있으면, 사용자가 수동으로 삭제하지 않는 한 데이터 볼륨이 호스트 컨테이너 인스턴스 상에 지정된 위치를 유지합니다. sourcePath

값이 호스트 컨테이너 인스턴스에 없을 경우 도커 대몬(daemon)이 해당 경로를 생성합니다. 해당 위치가 있을 경우 소스 경로 폴더의 콘텐츠를 내보냅니다.

유형: 문자열

필수 여부: 아니요

`efsVolumeConfiguration`

이 파라미터는 태스크 저장을 위해 Amazon Elastic File System 파일 시스템을 사용할 때 지정됩니다. 자세한 내용은 Amazon EFS 볼륨 섹션을 참조하세요.

유형: 객체

필수 여부: 아니요

`authorizationConfig`

Amazon EFS 파일 시스템에 대한 권한 부여 구성 세부 정보입니다.

유형: 문자열

필수 여부: 아니요

`accessPointId`

사용할 Amazon EFS 액세스 포인트 ID입니다. 액세스 포인트를 지정하는 경우 `EFSVolumeConfiguration`에 지정된 루트 디렉터리 값을 생략하거나 `/`로 설정해야 합니다. 그러면 EFS 액세스 포인트에 설정된 경로가 적용됩니다. 액세스 포인트를 사용하는 경우 `EFSVolumeConfiguration`에서 전송 중 데이터 암호화를 활성화해야 합니다. 자세한 정보는 Amazon Elastic File System 사용 설명서의 Amazon EFS 액세스 포인트 태스크를 참조세요.

유형: 문자열

필수 여부: 아니요

`iam`

Amazon EFS 파일 시스템을 탑재할 때 작업 정의에 정의된 AWS Batch 작업 IAM 역할을 사용할지 여부를 결정합니다. 활성화된 경우 `EFSVolumeConfiguration`에서 전송 중 데이터 암호화를 활성화해야 합니다. 이 파라미터가 누락되면 `DISABLED`의 기본값이 사용됩니다. 자세한 내용은 Amazon EFS 액세스 포인트 사용 섹션을 참조하세요.

유형: 문자열

유효한 값: `ENABLED | DISABLED`

필수 여부: 아니요

`fileSystemId`

사용할 Amazon EFS 파일 시스템 ID입니다.

유형: 문자열

필수 여부: 아니요

`rootDirectory`

호스트 내의 루트 디렉터리로 탑재할 Amazon EFS 파일 시스템 내 디렉터리입니다. 이 파라미터가 생략되면 Amazon EFS 볼륨의 루트가 사용됩니다. `/`를 지정하면 이 파라미터를 생략하는 것과 동일한 효과가 있습니다. 최대 길이는 4,096자입니다.

> **Important**
>
> `authorizationConfig`에서 EFS 액세스 포인트를 지정하는 경우 루트 디렉터리 파라미터를 생략하거나 `/`로 설정해야 합니다. 그러면 Amazon EFS 액세스 포인트에 설정된 경로가 적용됩니다.

유형: 문자열

필수 여부: 아니요

`transitEncryption`

암호화에서 Amazon ECS 호스트와 Amazon EFS 서버 간 전송 중 Amazon EFS 데이터를 활성화할지 여부를 결정합니다. Amazon EFS IAM 권한 부여를 사용하는 경우 전송 중 데이터 암호화를 활성화해야 합니다. 이 파라미터가 누락되면 `DISABLED`의 기본값이 사용됩니다. 자세한 내용을 알아보려면 Amazon Elastic File System 사용 설명서의 전송 중 데이터 암호화를 참조하세요.

유형: 문자열

유효한 값: `ENABLED | DISABLED`

필수 여부: 아니요

`transitEncryptionPort`

Amazon ECS 호스트와 Amazon EFS 서버 간에 암호화된 데이터를 전송할 때 사용할 포트입니다. 전송 중 데이터 암호화 포트를 지정하지 않으면 Amazon EFS 탑재 헬퍼가 사용하는 포트 선택 전략이 사용됩니다. 이 값은 0~65,535여야 합니다. 자세한 정보는 Amazon Elastic File System 사용 설명서의 EFS 탑재 헬퍼를 참조하세요.

유형: Integer

필수 여부: 아니요

Amazon EKS 속성

Amazon EKS 기반 작업과 관련된 다양한 속성이 있는 객체입니다. Amazon ECS 기반 작업 정의에 지정해서는 안 됩니다.

`podProperties`

작업의 Kubernetes 포드 리소스에 대한 속성입니다.

유형: EksPodProperties 객체

필수 여부: 아니요

`containers`

Amazon EKS 포드에서 사용되는 컨테이너의 속성입니다.

유형: EksContainer 객체

필수 여부: 아니요

`args`

진입점에 대한 인수 배열입니다. 지정하지 않으면 컨테이너 이미지의 CMD가 사용됩니다. 이는 Kubernetes에서 포드의 진입점 부분에 있는 `args` 멤버에 해당합니다. 환경 변수 참조는 컨테이너의 환경을 사용하여 확장됩니다.

참조된 환경 변수가 존재하지 않는 경우 명령의 참조는 변경되지 않습니다. 예를 들어, 참조가 '$(NAME1)'이고 NAME1 환경 변수가 존재하지 않는 경우 명령 문자열은 '$(NAME1)'로 유지됩니다 $$는 $로 바뀌고 결과 문자열은 확장되지 않습니다. 예를 들어, $$(VAR_NAME)은

VAR_NAME 환경 변수의 존재 여부에 관계없이 $(VAR_NAME)으로 전달됩니다. 자세한 내용은 Dockerfile 참조의 CMD와 Kubernetes 문서의 포드에 대한 명령 및 인자 정의를 참조하세요.

유형: 문자열 배열

필수 여부: 아니요

command

컨테이너의 진입점입니다. 쉘 내에서는 실행되지 않습니다. 지정하지 않으면 컨테이너 이미지의 ENTRYPOINT가 사용됩니다. 환경 변수 참조는 컨테이너의 환경을 사용하여 확장됩니다.

참조된 환경 변수가 존재하지 않는 경우 명령의 참조는 변경되지 않습니다. 예를 들어, 참조가 '$(NAME1)'이고 NAME1 환경 변수가 존재하지 않는 경우 명령 문자열은 '$(NAME1)'로 유지됩니다 '$$'는 $로 바뀌고 결과 문자열은 확장되지 않습니다. 예를 들어, $$(VAR_NAME)은 VAR_NAME 환경 변수의 존재 여부에 관계없이 $(VAR_NAME)으로 전달됩니다. 진입점은 업데이트할 수 없습니다. 자세한 내용은 Dockerfile 참조의 진입점과 Kubernetes 문서의 컨테이너에 대한 명령 및 인자 정의와 진입점을 참조하세요.

유형: 문자열 배열

필수 여부: 아니요

env

컨테이너로 전달할 환경 변수입니다.

> **Note**
>
> 환경 변수는 'AWS_BATCH'로 시작할 수 없습니다. 이 이름 지정 규칙은 AWS Batch가 설정하는 변수용으로 예약되어 있습니다.

유형: EksContainerEnvironmentVariable 객체 배열

필수 여부: 아니요

name

환경 변수의 이름입니다.

유형: 문자열

필수 항목 여부: 예

`value`

환경 변수의 값입니다.

유형: 문자열

필수 여부: 아니요

`image`

컨테이너를 시작하는 데 사용되는 도커 이미지입니다.

유형: 문자열

필수 항목 여부: 예

`imagePullPolicy`

컨테이너에 대한 이미지 가져오기 정책입니다. 지원되는 값은 `Always`, `IfNotPresent` 및 `Never`입니다. 이 파라미터의 기본값은 `IfNotPresent`입니다. 그러나 `:latest` 태그가 지정된 경우 `Always`로 기본 설정됩니다. 자세한 내용은 Kubernetes 문서의 이미지 업데이트를 참조하세요.

유형: 문자열

필수 여부: 아니요

`name`

컨테이너의 이름입니다. 이름을 지정하지 않으면 기본 이름 '`Default`'가 사용됩니다. 포드의 컨테이너마다 고유한 이름이 있어야 합니다.

유형: 문자열

필수 여부: 아니요

`resources`

컨테이너에 할당할 리소스의 유형 및 양입니다. 지원되는 리소스에는 `memory`cpu 및 `nvidia.com/gpu`이 포함됩니다. 자세한 내용은 Kubernetes 문서의 포드 및 컨테이너 리소스 관리를 참조하세요.

유형: EksContainerResourceRequirements 객체

필수 여부: 아니요

`limits`

컨테이너용으로 예약할 리소스의 유형 및 수량입니다. 값은 지정된 name에 따라 다릅니다. `limits` 또는 `requests` 객체를 사용하여 리소스를 요청할 수 있습니다.

메모리

정수를 사용하는 컨테이너의 메모리 하드 제한(MiB)으로 접미사는 'Mi'입니다. 컨테이너가 지정된 메모리를 초과하려 하면 해당 컨테이너가 중지됩니다. 작업에 대해 최소 4MiB의 메모리를 지정해야 합니다. `limits`, `requests` 또는 둘 다에 `memory`를 지정할 수 있습니다. `memory`가 두 곳 모두에 지정된 경우 `limits`에 지정된 값이 `requests`에 지정된 값과 같아야 합니다.

> **Note**
>
> 리소스 활용도를 극대화하려면 사용 중인 특정 인스턴스 유형에 대해 가능한 한 많은 메모리를 작업에 제공합니다. 자세한 방법은 컴퓨팅 리소스 메모리 관리을 참조하세요.

cpu

컨테이너용으로 예약된 CPU 개수입니다. 값은 `0.25`의 짝수 배수여야 합니다. `limits`, `requests` 또는 둘 다에 `cpu`를 지정할 수 있습니다. `cpu`가 두 곳 모두에 지정된 경우 `limits`에 지정된 값이 `requests`에 지정된 값 이상이어야 합니다.

nvidia.com/gpu

컨테이너용으로 예약된 GPU 개수입니다. 값은 정수여야 합니다. `limits`, `requests` 또는 둘 다에 `memory`를 지정할 수 있습니다. `memory`가 두 곳 모두에 지정된 경우 `limits`에 지정된 값이 `requests`에 지정된 값과 같아야 합니다.

유형: 문자열 간 맵

값 길이 제약 조건: 최소 길이는 1. 최대 길이는 256.

필수 여부: 아니요

`requests`

컨테이너에 요청할 리소스의 유형 및 수량입니다. 값은 지정된 `name`에 따라 다릅니다. `limits` 또는 `requests` 객체를 사용하여 리소스를 요청할 수 있습니다.

메모리

정수를 사용하는 컨테이너의 메모리 하드 제한(MiB)으로 접미사는 'Mi'입니다. 컨테이너가 지정된 메모리를 초과하려 하면 해당 컨테이너가 중지됩니다. 작업에 대해 최소 4MiB의 메모리를 지정해야 합니다. `limits`, `requests` 또는 둘 다에 `memory`를 지정할 수 있습니다. `memory`가 둘 다에 지정된 경우 `limits`에 지정된 값이 `requests`에 지정된 값과 같아야 합니다.

> **Note**
>
> 특정 인스턴스 유형에 대해 작업에 가능한 한 많은 메모리를 제공하여 리소스 사용률을 최대화하려는 경우 컴퓨팅 리소스 메모리 관리 섹션을 참조하세요.

cpu

컨테이너용으로 예약된 CPU 개수입니다. 값은 `0.25`의 짝수 배수여야 합니다. `limits`, `requests` 또는 둘 다에 `cpu`를 지정할 수 있습니다. `cpu`가 둘 다에 지정된 경우 `limits`에 지정된 값이 `requests`에 지정된 값 이상이어야 합니다.

nvidia.com/gpu

컨테이너용으로 예약된 GPU 개수입니다. 값은 정수여야 합니다. `limits`, `requests` 또는 둘 다에 `nvidia.com/gpu`를 지정할 수 있습니다. `nvidia.com/gpu`가 둘 다에 지정된 경우 `limits`에 지정된 값이 `requests`에 지정된 값과 같아야 합니다.

유형: 문자열 간 맵

값 길이 제약 조건: 최소 길이는 1. 최대 길이는 256.

필수 여부: 아니요

`securityContext`

작업의 보안 컨텍스트입니다. 자세한 내용은 Kubernetes 문서의 포드 또는 컨테이너에 대한 보안 컨텍스트 구성을 참조하세요.

유형: EksContainerSecurityContext 객체

필수 여부: 아니요

`privileged`

이 파라미터가 `true`일 경우 컨테이너에는 호스트 컨테이너 인스턴스에 대한 승격된 권한이 부여됩니다. 권한 수준은 `root` 사용자 권한과 유사합니다. 기본값은 `false`입니다. 이 파라미터는 Kubernetes 문서의 권한이 있는 포드 보안 정책에 있는 `privileged` 정책에 매핑됩니다.

유형: 부울

필수 여부: 아니요

`readOnlyRootFilesystem`

이 파라미터가 `true`일 경우 컨테이너에는 루트 파일 시스템에 대한 읽기 전용 액세스가 부여됩니다. 기본값은 `false`입니다. 이 파라미터는 Kubernetes 문서의 볼륨 및 파일 시스템 포드 보안 정책에 있는 `ReadOnlyRootFilesystem` 정책에 매핑됩니다.

유형: 부울

필수 여부: 아니요

`runAsGroup`

이 파라미터를 지정하면 지정된 그룹 ID(`gid`)로 컨테이너가 실행됩니다. 이 파라미터를 지정하지 않으면 기본값은 이미지 메타데이터에 지정된 그룹입니다. 이 파라미터는 Kubernetes 문서의 사용자 및 그룹 포드 보안 정책에 있는 `RunAsGroup` 및 `MustRunAs` 정책에 매핑됩니다.

유형: Long

필수 여부: 아니요

`runAsNonRoot`

이 파라미터를 지정하면 컨테이너는 `uid`가 `0`이 아닌 사용자로 실행됩니다. 이 파라미터를 지정하지 않으면 해당 규칙이 적용됩니다. 이 파라미터는 Kubernetes 문서의 사용자 및 그룹 포드 보안 정책에 있는 `RunAsUser` 및 `MustRunAsNonRoot` 정책에 매핑됩니다.

유형: Long

필수 여부: 아니요

`runAsUser`

이 파라미터를 지정하면 지정된 사용자 ID(`uid`)로 컨테이너가 실행됩니다. 이 파라미터를 지정하지 않으면 기본값은 이미지 메타데이터에 지정된 사용자입니다. 이 파라미터는 Kubernetes 문서의 사용자 및 그룹 포드 보안 정책에 있는 `RunAsUser` 및 `MustRanAs` 정책에 매핑됩니다.

유형: Long

필수 여부: 아니요

`volumeMounts`

Amazon EKS 작업의 컨테이너에 볼륨이 마운트됩니다. Kubernetes의 볼륨 및 볼륨 마운트에 대한 자세한 내용은 Kubernetes 문서의 볼륨을 참조하세요.

유형: EksContainerVolumeMount 객체 배열

필수 여부: 아니요

`mountPath`

볼륨이 마운트되는 컨테이너의 경로입니다.

유형: 문자열

필수 여부: 아니요

`name`

볼륨 마운트의 이름입니다. 포드에 있는 볼륨 중 하나의 이름과 일치해야 합니다.

유형: 문자열

필수 여부: 아니요

`readOnly`

이 값이 `true`일 경우 컨테이너에는 볼륨에 대한 읽기 전용 액세스가 부여됩니다. 그렇지 않으면 컨테이너가 볼륨에 쓸 수 있습니다. 기본값은 `false`입니다.

유형: 부울

필수 여부: 아니요

`dnsPolicy`

포드의 DNS 정책입니다. 기본값은 `ClusterFirst`입니다. `hostNetwork` 파라미터가 지정되지 않은 경우 기본값은 `ClusterFirstWithHostNet`입니다. `ClusterFirst`는 구성된 클러스터 도메인 접미사와 일치하지 않는 모든 DNS 쿼리가 노드에서 상속된 업스트림 이름 서버로 전달됨을 나타냅니다. RegisterJobDefinition API 작업에서 `dnsPolicy`에 대해 지정된 값이 없으면 DescribeJobDefinitions 또는 DescribeJobs API 작업에서 `dnsPolicy`에 대한 값이 반환되지 않습니다. 포드 사양 설정에는 `hostNetwork` 파라미터 값에 따라 `ClusterFirst` 또는 `ClusterFirstWithHostNet`이 포함됩니다. 자세한 내용은 Kubernetes 문서의 포드 DNS 정책을 참조하세요.

유효한 값: `Default` | `ClusterFirst` | `ClusterFirstWithHostNet`

유형: 문자열

필수 여부: 아니요

`hostNetwork`

포드가 호스트의 네트워크 IP 주소를 사용하는지 여부를 나타냅니다. 기본값은 `true`입니다. 이를 `false`로 설정하면 Kubernetes 포드 네트워킹 모델이 활성화됩니다. 대부분의 AWS Batch 워크로드는 송신 전용이며 수신 연결에 대해 포드마다 IP 할당 오버헤드가 필요하지 않습니다. 자세한 내용은 Kubernetes 문서의 호스트 네임스페이스와 포드 네트워킹을 참조하세요.

유형: 부울

필수 여부: 아니요

`serviceAccountName`

포드를 실행하는 데 사용되는 서비스 계정의 이름입니다. 자세한 내용은 Amazon EKS 사용 설명서의 Kubernetes 서비스 계정과 IAM 역할을 수임하도록 Kubernetes 서비스 계정 구성 및 Kubernetes 문서의 포드에 대한 서비스 계정 구성을 참조하세요.

유형: 문자열

필수 여부: 아니요

`volumes`

Amazon EKS 리소스를 사용하는 작업 정의에 대한 볼륨을 지정합니다.

유형: EKSVolume 객체 배열

필수 여부: 아니요

emptyDir

Kubernetes emptyDir 볼륨의 구성을 지정합니다. emptyDir 볼륨은 포드가 노드에 할당될 때 처음 생성됩니다. 포드가 해당 노드에서 실행되는 한 존재합니다. emptyDir 볼륨은 처음에는 비어 있습니다. 포드의 모든 컨테이너는 emptyDir 볼륨의 파일을 읽고 쓸 수 있습니다. 그러나 emptyDir 볼륨은 각 컨테이너에서 동일하거나 다른 경로에 마운트될 수 있습니다. 어떤 이유로든 포드가 노드에서 제거되면 emptyDir의 데이터는 영구적으로 삭제됩니다. 자세한 내용은 Kubernetes 문서의 emptyDir를 참조하세요.

유형: EksEmptyDir 객체

필수 여부: 아니요

medium

볼륨을 저장할 매체입니다. 기본값은 노드의 스토리지를 사용하는 빈 문자열입니다.

""

(기본값) 노드의 디스크 스토리지를 사용합니다.

"Memory"

노드의 RAM이 지원하는 tmpfs 볼륨을 사용합니다. 노드가 재부팅되면 볼륨의 콘텐츠가 손실되고 볼륨의 모든 스토리지는 컨테이너의 메모리 제한에 포함됩니다.

유형: 문자열

필수 여부: 아니요

sizeLimit

볼륨의 최대 크기입니다. 기본적으로 최대 크기는 정의되어 있지 않습니다.

유형: 문자열

길이 제약 조건: 최소 길이는 1입니다. 최대 길이는 256.

필수 여부: 아니요

hostPath

Kubernetes `hostPath` 볼륨의 구성을 지정합니다. `hostPath` 볼륨은 호스트 노드의 파일 시스템에 있는 기존 파일 또는 디렉터리를 포드에 마운트합니다. 자세한 내용은 Kubernetes 문서의 hostPath를 참조하세요.

유형: EksHostPath 객체

필수 여부: 아니요

경로

포드의 컨테이너에 마운트할 호스트의 파일 또는 디렉터리 경로입니다.

유형: 문자열

필수 여부: 아니요

name

볼륨의 이름입니다. 이름은 DNS 하위 도메인 이름으로 허용되어야 합니다. 자세한 내용은 Kubernetes 문서의 DNS 서브도메인 이름을 참조하세요.

유형: 문자열

필수 항목 여부: 예

보안 암호

Kubernetes `secret` 볼륨의 구성을 지정합니다. 자세한 내용은 Kubernetes 설명서의 암호를 참조하세요.

유형: EksSecret 객체

필수 여부: 아니요

선택 사항

암호 또는 암호 키를 정의해야 하는지 여부를 지정합니다.

유형: 부울

필수 여부: 아니요

secretName

암호의 이름입니다. 이름은 DNS 하위 도메인 이름으로 허용되어야 합니다. 자세한 내용은 Kubernetes 문서의 DNS 서브도메인 이름을 참조하세요.

유형: 문자열

필수 항목 여부: 예

플랫폼 기능

`platformCapabilities`

작업 정의에 필요한 플랫폼 기능입니다. 값을 지정하지 않은 경우 기본값은 `EC2`입니다. Fargate 리소스에서 실행되는 작업에는 `FARGATE`가 지정됩니다.

> **Note**
>
> 작업이 Amazon EKS 리소스에서 실행되는 경우, `platformCapabilities`를 지정하지 않아야 합니다.

유형: 문자열

유효한 값: `EC2` | `FARGATE`

필수 여부: 아니요

태그 전파

`propagateTags`

작업 또는 작업 정의에서 해당 Amazon ECS 작업으로 태그를 전파할지 여부를 지정합니다. 값을 지정하지 않으면 태그가 전파되지 않습니다. 태그는 작업이 생성될 때만 작업에 전파될 수 있습니다. 이름이 같은 태그의 경우 작업 태그가 작업 정의 태그보다 우선합니다. 작업 및 작업 정의의 결합된 태그의 총 수가 50개를 넘으면 작업이 `FAILED` 상태로 전환됩니다.

> **Note**
>
> 작업이 Amazon EKS 리소스에서 실행되는 경우, `propagateTags`를 지정하지 않아야 합니다.

유형: 부울

필수 여부: 아니요

노드 속성

`nodeProperties`

다중 노드 병렬 작업 정의를 등록할 때 노드 속성의 목록을 지정해야 합니다. 이러한 노드 속성은 작업에 사용할 노드 수, 주 노드 인덱스 및 사용할 다양한 노드 범위를 정의합니다. 작업이 Fargate 리소스에서 실행되는 경우, `nodeProperties`를 지정할 수 없습니다. 대신 `containerProperties`를 사용합니다. 다음과 같은 노드 속성을 작업 정의에 사용할 수 있습니다. 자세한 내용은 다중 노드 병렬 작업 섹션을 참조하세요.

> **Note**
>
> 작업이 Amazon EKS 리소스에서 실행되는 경우, `nodeProperties`를 지정하지 않아야 합니다.

유형: NodeProperties 객체

필수 여부: 아니요

`mainNode`

다중 노드 병렬 작업의 기본 노드에 대한 노드 인덱스를 지정합니다. 이 노드 인덱스 값은 노드 수보다 적어야 합니다.

유형: 정수

필수 여부: 예

`numNodes`

다중 노드 병렬 작업과 연결된 노드 수입니다.

유형: 정수

필수 여부: 예

`nodeRangeProperties`

노드 범위 및 다중 노드 병렬 작업과 연결된 해당 속성의 목록입니다.

> **Note**
>
> 노드 그룹은 모두 동일한 컨테이너 속성을 공유하는 작업 노드의 동일한 그룹입니다. AWS Batch를 사용하여 각 작업에 대해 최대 5개의 개별 노드 그룹을 지정할 수 있습니다.

유형: NodeRangeProperty 객체의 배열

필수 여부: 예

`targetNodes`

노드 인덱스 값을 사용하는 노드의 범위입니다. `0:3`의 범위는 인덱스 값이 `0-3`인 노드를 나타냅니다. 시작 범위 값이 생략되면(`:n`) 이 범위를 시작하는 데 `0`이 사용됩니다. 종료 범위 값이 생략되면(`n:`) 가능한 최고 노드 인덱스가 범위를 종료하는 데 사용됩니다. 누적 노드 범위는 모든 노드(`0:n`)를 고려해야 합니다. 노드 범위를 중첩할 수 있습니다(예: `0:10` 및 `4:5`). 이 경우 `4:5` 범위 속성이 `0:10` 속성을 재정의합니다.

유형: 문자열

필수 여부: 아니요

`container`

노드 범위에 대한 컨테이너 세부 정보입니다. 자세한 내용은 컨테이너 속성 섹션을 참조하세요.

유형: ContainerProperties 객체

필수 여부: 아니요

재시도 전략

`retryStrategy`

작업 정의를 등록할 때 해당 작업 정의로 제출하여 실패한 작업에 대해 재시도 전략을 지정할 수도 있습니다. SubmitJob 작업 중에 지정된 모든 재시도 전략은 여기에 정의된 재시도 전략을 재정의합니다. 기본적으로 각 작업은 한 번 재시도됩니다. 시도를 두 번 이상으로 지정하면 작업이 실패할 경우 작업이 재시도됩니다. 실패한 시도의 예로는 작업이 `0`이 아닌 종료 코드를 반환하거나 컨테이너 인스턴스가 종료되는 경우를 들 수 있습니다. 자세한 내용은 작업 자동 재시도 섹션을 참조하세요.

유형: RetryStrategy 객체

필수 여부: 아니요

`attempts`

작업이 RUNNABLE 상태로 이동하는 횟수입니다. 1부터 10까지 시도 횟수를 지정할 수 있습니다. `attempts`가 1보다 크면 작업이 실패할 경우 작업이 RUNNABLE 상태로 될 때까지 작업이 재시도됩니다.

```
"attempts": integer
```

유형: Integer

필수 여부: 아니요

`evaluateOnExit`

작업이 다시 시도되거나 실패하는 조건을 지정하는 최대 5개의 객체 배열입니다. 이 파라미터를 지정하면 `attempts` 파라미터도 지정해야 합니다. `evaluateOnExit`를 지정했지만 일치하는 항목이 없는 경우 작업이 다시 시도됩니다.

```
"evaluateOnExit": [
   {
      "action": "string",
      "onExitCode": "string",
      "onReason": "string",
      "onStatusReason": "string"
   }
]
```

유형: EvaluateOnExit 객체 배열

필수 여부: 아니요

`action`

지정된 모든 조건 (`onStatusReason`, `onReason` 및 `onExitCode`) 이 충족되는 경우 수행할 작업을 지정합니다. 값은 대/소문자를 구분하지 않습니다.

유형: 문자열

필수 항목 여부: 예

유효한 값: RETRY | EXIT

onExitCode

작업에 대해 반환된 ExitCode의 십진수 표현과 일치하는 glob 패턴을 포함합니다. 패턴의 최대 길이는 512자입니다. 숫자만 포함할 수 있습니다. 문자나 특수 문자를 포함할 수 없습니다. 선택적으로 별표 (*) 로 끝날 수 있으므로 문자열의 시작 부분 만 정확히 일치해야합니다.

유형: 문자열

필수 여부: 아니요

onReason

작업에 대해 반환된 Reason와 일치하는 glob 패턴을 포함합니다. 패턴의 최대 길이는 512자입니다. 문자, 숫자, 마침표(.), 콜론(:) 및 공백(공백, 탭)을 포함할 수 있습니다. 선택적으로 별표 (*) 로 끝날 수 있으므로 문자열의 시작 부분 만 정확히 일치해야합니다.

유형: 문자열

필수 여부: 아니요

onStatusReason

작업에 대해 반환된 StatusReason와 일치하는 glob 패턴을 포함합니다. 패턴의 최대 길이는 512자입니다. 문자, 숫자, 마침표(.), 콜론(:) 및 공백(공백, 탭)을 포함할 수 있습니다. 선택적으로 별표 (*) 로 끝날 수 있으므로 문자열의 시작 부분 만 정확히 일치해야합니다.

유형: 문자열

필수 여부: 아니요

예약 우선 순위

schedulingPriority

이 작업 정의와 함께 제출된 작업의 예약 우선 순위입니다. 이는 공정 공유 정책이 있는 작업 대기열의 작업에만 영향을 줍니다. 예약 우선순위가 높은 작업이 예약 우선순위가 낮은 작업보다 먼저 예약됩니다.

지원되는 최솟값은 0이고, 지원되는 최댓값은 9999입니다.

유형: Integer

필수 여부: 아니요

태그

`tags`

작업 정의와 연결할 키-값 쌍 태그. 자세한 내용은 AWS Batch 리소스 태깅 섹션을 참조하세요.

유형: 문자열 대 문자열 맵

필수 항목 여부: 아니요

Timeout

`timeout`

작업이 더 오래 실행되면 AWS Batch(이)가 작업을 종료하도록 작업의 제한 시간을 구성할 수 있습니다. 자세한 내용은 작업 제한 시간 섹션을 참조하세요. 제한 시간으로 인해 작업이 종료되면 재시도되지 않습니다. SubmitJob 작업 중에 지정된 모든 제한 시간 구성은 여기에 정의된 제한 시간 구성을 재정의합니다. 자세한 내용은 작업 제한 시간 섹션을 참조하세요.

유형: JobTimeout 객체

필수 여부: 아니요

`attemptDurationSeconds`

AWS Batch가 완료되지 않은 작업을 종료할 때까지의 기간(초)(작업 시도의 `startedAt` 타임스탬프에서 측정됨)입니다. 제한 시간의 최소 값은 60초입니다.

배열 작업의 경우 상위 배열 작업이 아닌 하위 작업에 제한 시간이 적용됩니다.

다중 노드 병렬(MNP) 작업의 경우 개별 노드가 아닌 전체 작업에 제한 시간이 적용됩니다.

유형: Integer

필수 여부: 아니요

EcsProperties를 사용하여 작업 정의 생성

를 사용하는 AWS Batch 작업 정의를 사용하면 별도의 컨테이너에서 하드웨어, 센서, 3D 환경 및 기타 시뮬레이션을 모델링`EcsProperties`할 수 있습니다. 이 기능을 사용하여 워크로드 구성 요소를

논리적으로 구성하고 기본 애플리케이션과 분리할 수 있습니다. 이 기능은 Amazon Elastic Container Service(Amazon ECS), Amazon Elastic Kubernetes Service(Amazon EKS) 및 AWS Batch 에서와 함께 사용할 수 있습니다 AWS Fargate.

`ContainerProperties` 및 `EcsProperties` 작업 정의

사용 사례에 따라 `ContainerProperties` 또는 `EcsProperties` 작업 정의를 사용하도록 선택할 수 있습니다. 상위 수준에서를 사용하여 AWS Batch 작업을 실행하는 `EcsProperties` 것은를 사용하여 작업을 실행하는 것과 유사합니다`ContainerProperties`.

`ContainerProperties`를 사용하는 레거시 작업 정의 구조는 계속 지원됩니다. 현재 이 구조를 사용하는 워크플로가 있는 경우 워크플로를 계속 실행할 수 있습니다.

주요 차이점은 작업 정의에 `EcsProperties` 기반 정의를 수용할 새 객체가 추가되었다는 것입니다.

예를 들어 Amazon ECS 및 Fargate에서 `ContainerProperties`를 사용하는 작업 정의의 구조는 다음과 같습니다.

```
{
   "containerProperties": {
     ...
     "image": "my_ecr_image1",
     ...
  },
...
}
```

Amazon ECS 및 Fargate에서 `EcsProperties`를 사용하는 작업 정의의 구조는 다음과 같습니다.

```
{
  "ecsProperties": {
    "taskProperties": [{
      "containers": [
        {
          ...
          "image": "my_ecr_image1",
          ...
        },
        {
          ...
          "image": "my_ecr_image2",
          ...
```

```
        },
```

AWS Batch APIs에 대한 일반 변경 사항

다음은 ContainerProperties를 사용할 때와 EcsProperties API 데이터 유형을 사용할 때의 몇 가지 주요 차이점에 대한 간략한 설명입니다.

- ContainerProperties 내에서 사용되는 많은 파라미터가 TaskContainerProperties 내에 표시됩니다. 몇 가지 예는 command, image, privileged, secrets 및 users입니다. 모두 TaskContainerProperties에서 찾을 수 있습니다.
- 일부 TaskContainerProperties 파라미터에는 레거시 구조에 기능적 등가물이 없습니다. 몇 가지 예는 dependsOn, essential, name, ipcMode 및 pidMode입니다. 자세한 내용은 EcsTaskDetails 및 TaskContainerProperties를 참조하세요.

  또한 일부 ContainerProperties 파라미터에는 EcsProperties 구조에 등가물 또는 애플리케이션이 없습니다. taskProperties에서 새 객체가 최대 10개의 요소를 수용할 수 있도록 container가 containers로 대체되었습니다. 자세한 내용은 RegisterJobDefinition:containerProperties 및 EcsTaskProperties:containers를 참조하세요.
- taskRoleArn은 기능면에서 jobRoleArn과 동일합니다. 자세한 내용은 EcsTaskProperties:taskRoleArn 및 ContainerProperties:jobRoleArn을 참조하세요.
- EcsProperties 구조에 컨테이너를 1개에서 10개까지 포함할 수 있습니다. 자세한 내용은 EcsTaskProperties:containers를 참조하세요.
- taskProperties 및 instanceTypes 객체는 배열이지만 현재는 하나의 요소만 수용합니다. 예를 들어 EcsProperties:taskProperties 및 NodeRangeProperty:instanceTypes가 있습니다.

Amazon ECS용 다중 컨테이너 작업 정의

Amazon ECS의 다중 컨테이너 구조를 수용하기 위해 일부 API 데이터 유형이 다릅니다. 예:

- ecsProperties는 단일 컨테이너 정의에서 containerProperties와 동일한 수준입니다. 자세한 내용은AWS Batch API 참조 안내서의 EcsProperties를 참조하세요.
- taskProperties에는 Amazon ECS 작업에 대해 정의된 속성이 포함되어 있습니다. 자세한 내용은AWS Batch API 참조 안내서의 EcsProperties를 참조하세요.
- containers는 단일 컨테이너 정의에서 containerProperties와 유사한 정보를 포함합니다. 주요 차이점은 containers를 사용하면 최대 10개의 컨테이너를 정의할 수 있다는 점입니다. 자세한 내용은 AWS Batch API 참조 안내서의 ECSTaskProperties:containers를 참조하세요.

- essential 파라미터는 컨테이너가 작업에 미치는 영향을 나타냅니다. 작업이 진행되려면 모든 필수 컨테이너가 성공적으로 완료(0으로 종료)되어야 합니다. 필수로 표시된 컨테이너가 실패(0이 아닌 상태로 종료)하면 작업이 실패합니다.

  기본값은 true이며 적어도 하나의 컨테이너가 essential로 표시되어야 합니다. 자세한 내용은 AWS Batch API 참조 안내서의 essential 섹션을 참조하세요.
- dependsOn 파라미터를 사용하여 컨테이너 종속성 목록을 정의할 수 있습니다. 자세한 내용은AWS Batch API 참조 안내서의 dependsOn 섹션을 참조하세요.

> **Note**
>
> dependsOn 목록의 복잡성과 관련 컨테이너 런타임은 작업의 시작 시간에 영향을 미칠 수 있습니다. 종속성을 실행하는 데 오랜 시간이 걸리는 경우 작업이 완료될 때까지 STARTING 상태가 유지됩니다.

ecsProperties 및 구조에 대한 자세한 내용은 ecsProperties에 대한 RegisterJobDefinition 요청 구문을 참조하세요.

Amazon EKS에 대한 다중 컨테이너 작업 정의

Amazon EKS의 다중 컨테이너 구조를 수용하기 위해 일부 API 데이터 유형이 다릅니다. 예:

- name은 컨테이너의 고유 식별자입니다. 이 객체는 단일 컨테이너에는 필요하지 않으며 포드에 여러 컨테이너를 정의할 때 필요합니다. name이 단일 컨테이너에 정의되지 않은 경우 기본 이름인 default가 적용됩니다.
- initContainers는 eksPodProperties 데이터 유형 내에 정의됩니다. 이들은 애플리케이션 컨테이너 전에 실행되며, 항상 완료될 때까지 실행되고, 다음 컨테이너가 시작되기 전에 성공적으로 완료되어야 합니다.

  이러한 컨테이너는 Amazon EKS Connector 에이전트에 등록되며 Amazon Elastic Kubernetes Service 백엔드 데이터 스토어에 등록 정보를 유지합니다. initContainers 객체는 최대 10개의 요소를 수용할 수 있습니다. 자세한 내용은 Kubernetes 설명서의 Init Containers를 참조하세요.

> **Note**
>
> `initContainers` 객체는 작업의 시작 시간에 영향을 미칠 수 있습니다. `initContainers`를 실행하는 데 오랜 시간이 걸리는 경우 작업이 완료될 때까지 `STARTING` 상태가 유지됩니다.

- `shareProcessNamespace`는 포드의 컨테이너가 동일한 프로세스 네임스페이스를 공유할 수 있는지 여부를 나타냅니다. 기본값은 `false`입니다. 이 값을 `true`로 설정하여 컨테이너가 동일한 포드에 위치한 다른 컨테이너의 프로세스를 보고 신호를 보낼 수 있게 합니다.
- 모든 컨테이너가 중요합니다. 작업이 성공하려면 모든 컨테이너가 성공적으로 완료(0으로 종료)되어야 합니다. 컨테이너 하나가 실패하면(0이 아닌 값으로 종료) 작업이 실패합니다.

`eksProperties` 및 구조에 대한 자세한 내용은 eksProperties에 대한 RegisterJobDefinition 요청 구문을 참조하세요.

참조: EcsProperties를 사용하는 AWS Batch 작업 시나리오

`EcsProperties`를 사용하는 AWS Batch 작업 정의를 필요에 따라 구성하는 방법을 설명하기 위해 이 주제에서는 다음 `RegisterJobDefinition` 페이로드를 제공합니다. 이러한 예제를 파일로 복사하고 필요에 맞게 사용자 지정한 다음 AWS Command Line Interface(AWS CLI)를 사용하여 `RegisterJobDefinition`을 호출할 수 있습니다.

Amazon EC2의 Amazon ECS에 대한 AWS Batch 작업

다음은 Amazon Elastic Compute Cloud의 Amazon Elastic Container Service에 대한 AWS Batch 작업의 예입니다.

```
{
    "jobDefinitionName": "multicontainer-ecs-ec2",
    "type": "container",
    "ecsProperties": {
        "taskProperties": [
          {
            "containers": [
              {
                "name": "c1",
                "essential": false,
                "command": [
                  "echo",
```

```
            "hello world"
          ],
          "image": "public.ecr.aws/amazonlinux/amazonlinux:latest",
          "resourceRequirements": [
            {
              "type": "VCPU",
              "value": "2"
            },
            {
              "type": "MEMORY",
              "value": "4096"
            }
          ]
        },
        {
          "name": "c2",
          "essential": false,
          "command": [
            "echo",
            "hello world"
          ],
          "image": "public.ecr.aws/amazonlinux/amazonlinux:latest",
          "resourceRequirements": [
            {
              "type": "VCPU",
              "value": "2"
            },
            {
              "type": "MEMORY",
              "value": "4096"
            }
          ]
        },
        {
          "name": "c3",
          "essential": true,
          "command": [
            "echo",
            "hello world"
          ],
          "image": "public.ecr.aws/amazonlinux/amazonlinux:latest",
          "firelensConfiguration": {
            "type": "fluentbit",
            "options": {
```

```
                    "enable-ecs-log-metadata": "true"
                  }
                 },
                "resourceRequirements": [
                  {
                    "type": "VCPU",
                    "value": "6"
                  },
                  {
                    "type": "MEMORY",
                    "value": "12288"
                  }
                ]
              }
            ]
          }
        ]
  }
}
```

Fargate의 Amazon ECS에 대한 AWS Batch 작업

다음은 AWS Fargate의 Amazon Elastic Container Service에 대한 AWS Batch 작업의 예입니다.

```
{
    "jobDefinitionName": "multicontainer-ecs-fargate",
    "type": "container",
    "platformCapabilities": [
        "FARGATE"
    ],
    "ecsProperties": {
        "taskProperties": [
          {
            "containers": [
              {
                "name": "c1",
                "command": [
                  "echo",
                  "hello world"
                ],
                "image": "public.ecr.aws/amazonlinux/amazonlinux:latest",
                "resourceRequirements": [
                  {
```

```
                    "type": "VCPU",
                    "value": "2"
                  },
                  {
                    "type": "MEMORY",
                    "value": "4096"
                  }
                ]
              },
              {
                "name": "c2",
                "essential": true,
                "command": [
                  "echo",
                  "hello world"
                ],
                "image": "public.ecr.aws/amazonlinux/amazonlinux:latest",
                "resourceRequirements": [
                  {
                    "type": "VCPU",
                    "value": "6"
                  },
                  {
                    "type": "MEMORY",
                    "value": "12288"
                  }
                ]
              }
            ],
            "executionRoleArn": "arn:aws:iam::1112223333:role/ecsTaskExecutionRole"
          }
        ]
  }
}
```

Amazon EKS에 대한 AWS Batch 작업

다음은 Amazon Elastic Kubernetes Service에 대한 AWS Batch 작업의 예입니다.

```
{
  "jobDefinitionName": "multicontainer-eks",
  "type": "container",
  "eksProperties": {
```

```
    "podProperties": {
      "shareProcessNamespace": true,
      "initContainers": [
        {
          "name": "init-container",
          "image": "public.ecr.aws/amazonlinux/amazonlinux:2",
          "command": [
            "echo"
          ],
          "args": [
            "hello world"
          ],
          "resources": {
            "requests": {
              "cpu": "1",
              "memory": "512Mi"
            }
          }
        },
        {
          "name": "init-container-2",
          "image": "public.ecr.aws/amazonlinux/amazonlinux:2",
          "command": [
            "echo",
            "my second init container"
          ],
          "resources": {
            "requests": {
              "cpu": "1",
              "memory": "512Mi"
            }
          }
        }
      ],
      "containers": [
        {
          "name": "c1",
          "image": "public.ecr.aws/amazonlinux/amazonlinux:2",
          "command": [
            "echo world"
         ],
          "resources": {
            "requests": {
              "cpu": "1",
```

```
              "memory": "512Mi"
            }
          }
        },
        {
          "name": "sleep-container",
          "image": "public.ecr.aws/amazonlinux/amazonlinux:2",
          "command": [
            "sleep",
            "20"
          ],
          "resources": {
            "requests": {
              "cpu": "1",
              "memory": "512Mi"
            }
          }
        }
      ]
    }
  }
}
```

노드당 여러 컨테이너가 있는 MNP AWS Batch 작업

다음은 노드당 여러 컨테이너가 있는 다중 노드 병렬(MNP) AWS Batch 작업의 예입니다.

```
{
  "jobDefinitionName": "multicontainer-mnp",
  "type": "multinode",
  "nodeProperties": {
    "numNodes": 6,
    "mainNode": 0,
    "nodeRangeProperties": [
      {
        "targetNodes": "0:5",
        "ecsProperties": {
          "taskProperties": [
            {
              "containers": [
                {
                  "name": "range05-c1",
                  "command": [
```

```
                "echo",
                "hello world"
              ],
              "image": "public.ecr.aws/amazonlinux/amazonlinux:latest",
              "resourceRequirements": [
                {
                  "type": "VCPU",
                  "value": "2"
                },
                {
                  "type": "MEMORY",
                  "value": "4096"
                }
              ]
            },
            {
              "name": "range05-c2",
              "command": [
                "echo",
                "hello world"
              ],
              "image": "public.ecr.aws/amazonlinux/amazonlinux:latest",
              "resourceRequirements": [
                {
                  "type": "VCPU",
                  "value": "2"
                },
                {
                  "type": "MEMORY",
                  "value": "4096"
                }
              ]
            }
          ]
        }
      ]
    }
  }
}
```

awslogs 로그 드라이버 사용

기본적으로는 awslogs 로그 드라이버가 CloudWatch Logs로 로그 정보를 전송할 수 있도록 AWS Batch 합니다. 이 기능을 사용하면 사용자는 편리한 위치에서 자신의 컨테이너에서 다양한 로그를 볼 수 있으며, 컨테이너 로그가 컨테이너 인스턴스에서 디스크 공간을 차지하는 것이 방지됩니다. 이 주제는 작업 정의에서 awslogs 로그 드라이버의 구성을 도와줍니다.

> **Note**
>
> AWS Batch 콘솔에서 작업 정의를 생성할 때 로깅 구성 섹션에서 awslogs 로그 드라이버를 구성할 수 있습니다.

> **Note**
>
> 작업 컨테이너가 로깅하는 정보 유형은 대부분 ENTRYPOINT 명령에 따라 결정됩니다. 기본적으로 수집되는 로그는 컨테이너를 로컬에서 실행했을 때 일반적으로 대화식 터미널에 표시되는 명령 출력으로 STDOUT 및 STDERR I/O 스트림을 나타냅니다. awslogs 로그 드라이버는 이러한 로그를 Docker에서 CloudWatch Logs로 전달하는 역할만 합니다. 다른 파일 데이터 또는 스트림을 수집할 수 있는 대체 방법을 포함해 Docker 로그가 처리되는 방식에 대한 자세한 정보는 Docker 설명서에서 컨테이너 또는 서비스 로그 보기 섹션을 참조하세요.

컨테이너 인스턴스에서 CloudWatch 로그로 시스템 로그를 전송하려면 에서 CloudWatch Logs 사용 AWS Batch 섹션을 참조하세요. CloudWatch Logs에 대한 자세한 정보는 Amazon CloudWatch Logs 사용 설명서의 로그 파일 모니터링 및 CloudWatch Logs 할당량을 참조하세요.

AWS Batch JobDefiniton 데이터 형식의 awslogs 로그 드라이버 옵션

awslogs 로그 드라이버는 AWS Batch 작업 정의에서 다음 옵션을 지원합니다. 자세한 내용은 Docker 문서의 CloudWatch 로그 로깅 드라이버를 참조하세요.

awslogs-region

필수 여부: 아니요

awslogs 로그 드라이버가 Docker 로그를 전송할 리전을 지정합니다. 기본적으로 사용되는 리전은 작업에 사용되는 리전과 동일합니다. 사용자는 여러 리전 작업의 모든 로그를 CloudWatch 로그의

단일 리전으로 전송할 수 있습니다. 이렇게 하면 한 위치에서 모든 작업을 볼 수 있습니다. 또는, 리전별로 구분하여 세분화할 수 있습니다. 그러나 이 옵션을 선택할 경우, 지정한 로그 그룹이 지정한 리전에 위치하는지 확인해야 하세요.

`awslogs-group`

필수 항목 여부: 선택 사항

`awslogs-group` 옵션을 사용하면 사용자는 `awslogs` 로그 드라이버가 로그 스트림을 전송할 로그 그룹을 지정할 수 있습니다. 지정되지 않은 경우 `aws/batch/job`이 사용됩니다.

`awslogs-stream-prefix`

필수 항목 여부: 선택 사항

`awslogs-stream-prefix` 옵션을 사용하면 로그 스트림을 지정된 접두사 및 컨테이너가 속한 AWS Batch 작업의 Amazon ECS 작업 ID와 연결할 수 있습니다. 이 옵션을 사용하여 접두사를 지정하는 경우 로그 스트림은 다음 형식을 취합니다.

```
prefix-name/default/ecs-task-id
```

`awslogs-datetime-format`

필수 여부: 아니요

이 옵션은 Python `strftime` 형식의 여러 줄 시작 패턴을 정의합니다. 로그 메시지는 패턴과 일치하는 하나의 라인과 패턴과 일치하지 않는 나머지 라인으로 이루어져 있습니다. 따라서 일치하는 줄은 로그 메시지 간의 구분 기호입니다.

이 형식을 사용하는 사용 사례의 한 예는 스택 덤프와 같은 출력을 구문 분석하는 것이며, 그렇지 않으면 여러 항목에 기록될 수 있습니다. 올바른 패턴을 통해 단일 항목으로 캡처할 수 있습니다.

자세한 내용은 awslogs-datetime-format을 참조하세요.

`awslogs-datetime-format`과 `awslogs-multiline-pattern`이 모두 구성된 경우, 이 옵션은 항상 우선순위를 갖습니다.

> **Note**
>
> 여러 줄 로깅은 모든 로그 메시지의 정규식 구문 분석 및 일치 태스크를 수행합니다. 이는 로깅 성능에 부정적인 영향을 줄 수 있습니다.

`awslogs-multiline-pattern`

필수 여부: 아니요

이 옵션은 정규식을 사용하여 여러 줄 시작 패턴을 정의합니다. 로그 메시지는 패턴과 일치하는 하나의 라인과 패턴과 일치하지 않는 나머지 라인으로 이루어져 있습니다. 따라서 일치하는 라인은 로그 메시지 간의 구분 기호입니다.

자세한 내용은 Docker 설명서의 awslogs-multiline-pattern을 참조하세요.

이 옵션은 `awslogs-datetime-format`을 구성하는 경우에 무시됩니다.

> **Note**
>
> 여러 줄 로깅은 모든 로그 메시지의 정규식 구문 분석 및 일치 태스크를 수행합니다. 이는 로깅 성능에 부정적인 영향을 줄 수 있습니다.

`awslogs-create-group`

필수 여부: 아니요

로그 그룹을 자동으로 생성할지를 지정합니다. 이 옵션이 지정되지 않은 경우 기본적으로 `false`로 설정됩니다.

> **Warning**
>
> 이 옵션은 권장되지 않습니다. 각 작업에서 로그 그룹을 생성하려고 시도하면 작업이 실패할 가능성이 높아지므로 CloudWatch Logs CreateLoggroup API 작업을 사용하여 로그 그룹을 미리 생성하는 것이 좋습니다.

> **Note**
>
> `awslogs-create-group`을 사용하기 전에 실행 역할에 대한 IAM 정책이 `logs:CreateLogGroup` 권한을 포함해야 합니다

작업 정의에서 로그 구성 지정

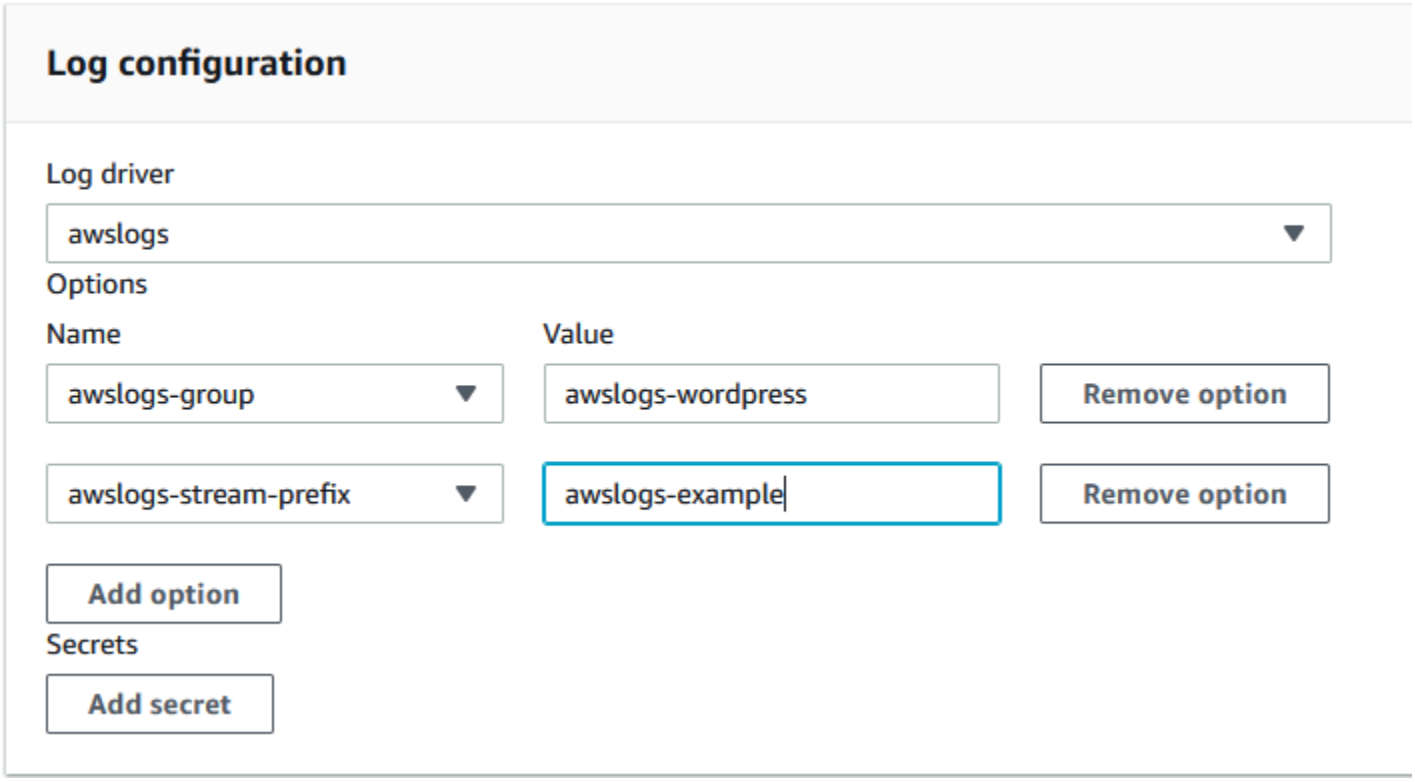
기본적으로는 awslogs 로그 드라이버를 AWS Batch 활성화합니다. 이 섹션에서는 작업의 awslogs 로그 구성을 사용자 지정하는 방법을 설명합니다. 자세한 내용은 단일 노드 작업 정의 생성 단원을 참조하십시오.

다음 로그 구성 JSON 스니펫에는 각 작업에 지정된 logConfiguration 개체가 있습니다. 하나는 awslogs-wordpress라는 로그 그룹에 로그를 보내는 WordPress 작업용이고 다른 하나는 awslogs-mysql이라는 로그 그룹에 로그를 보내는 MySQL 컨테이너용입니다. 두 컨테이너 모두 awslogs-example 로그 스트림 접두사를 사용합니다.

```
"logConfiguration": {
    "logDriver": "awslogs",
    "options": {
        "awslogs-group": "awslogs-wordpress",
        "awslogs-stream-prefix": "awslogs-example"
    }
}
```

```
"logConfiguration": {
    "logDriver": "awslogs",
    "options": {
        "awslogs-group": "awslogs-mysql",
        "awslogs-stream-prefix": "awslogs-example"
    }
}
```

AWS Batch 콘솔에서 wordpress 작업 정의에 대한 로그 구성은 다음 이미지와 같이 지정됩니다.



작업 정의 로그 구성에 awslogs 로그 드라이버로 태스크 정의를 등록하면 사용자는 CloudWatch 로그로 로그 전송을 시작할 수 있는 작업 정의를 가진 작업을 제출할 수 있습니다. 자세한 내용은 자습서: 작업 제출 단원을 참조하십시오.

민감한 데이터 지정

를 사용하면 민감한 데이터를 보안 암호 또는 AWS Systems Manager 파라미터 스토어 파라미터에 AWS Secrets Manager 저장한 다음 작업 정의에서 참조하여 작업에 민감한 데이터를 주입할 AWS Batch수 있습니다.

암호는 다음과 같은 방법으로 작업에 노출될 수 있습니다.

- 환경 변수로 민감한 데이터를 컨테이너에 삽입하려면 secrets 작업 정의 파라미터를 사용하세요.
- 작업의 로그 구성에서 중요한 정보를 참조하려면 secretOptions 작업 정의 파라미터를 사용하세요.

주제

- Secrets Manager를 사용하여 민감한 데이터 지정
- Systems Manager Parameter Store를 사용하여 민감한 데이터 지정

Secrets Manager를 사용하여 민감한 데이터 지정

를 사용하면 민감한 데이터를 AWS Secrets Manager 보안 암호에 저장한 다음 작업 정의에서 참조하여 작업에 민감한 데이터를 주입할 AWS Batch수 있습니다. Secrets Manager 암호에 저장된 민감한 데이터는 환경 변수 또는 로그 구성의 일부로 작업에 노출될 수 있습니다.

비밀을 환경 변수로 주입하는 경우 주입할 비밀의 JSON 키 또는 버전을 지정할 수 있습니다. 이 프로세스는 작업에 노출되는 중요한 데이터를 제어하는 데 도움이 됩니다. 보안 버전 관리에 대한 자세한 정보는 AWS Secrets Manager 사용 설명서의 AWS Secrets Manager의 주요 개념 및 용어를 참조하세요.

Secrets Manager를 사용하여 민감한 데이터 지정 시 고려할 사항

Secrets Manager를 사용하여 작업에 대한 민감한 데이터를 지정할 때 다음 사항을 고려해야 합니다.

- 보안 암호의 특정 JSON 키 또는 버전을 사용하여 보안 암호를 주입하려면 컴퓨팅 환경의 컨테이너 인스턴스에 Amazon ECS 컨테이너 에이전트 버전 1.37.0 이상이 있어야 합니다. 그러나 최신 컨테이너 에이전트 버전을 사용하는 것이 좋습니다. 에이전트 버전을 확인하고 최신 버전으로 업데이트하는 방법에 대한 자세한 내용은 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS 컨테이너 에이전트 업데이트를 참조하세요.

  보안 정보의 전체 내용을 환경 변수로 주입하거나 로그 구성에 보안 정보를 주입하려면 컨테이너 인스턴스에 컨테이너 에이전트 버전 1.23.0 이상이 있어야 합니다.
- CreateSecret API의 `SecretString` 파라미터로 생성된 암호이며 텍스트 데이터를 저장하는 암호만이 지원됩니다. CreateSecret API의 `SecretBinary` 파라미터로 생성된 암호이며 바이너리 데이터를 저장하는 암호는 지원되지 않습니다.
- Secrets Manager 암호를 참조하여 작업에 대한 민감한 데이터를 검색하는 작업 정의를 사용할 때, 인터페이스 VPC 엔드포인트도 사용하는 경우 Secrets Manager에 대한 인터페이스 VPC 엔드포인트를 생성해야 합니다. 자세한 내용은 AWS Secrets Manager 사용 설명서의 VPC 엔드포인트와 함께 Secrets Manager 사용을 참조하세요.
- 작업이 처음 시작될 때 해당 작업에 중요한 정보가 주입됩니다. 암호가 이후에 업데이트되거나 교체되면 작업이 업데이트된 값을 자동으로 받지 않습니다. 서비스가 업데이트된 암호 값으로 새 작업을 강제로 시작하도록 하려면 새 작업을 시작해야 합니다.

AWS Batch 보안 암호에 필요한 IAM 권한

이 기능을 사용하려면 실행 역할이 있어야 하며 작업 정의에서 해당 역할을 참조해야 합니다. 이렇게 하면 컨테이너 에이전트가 필요한 Secrets Manager 리소스를 가져올 수 있습니다. 자세한 내용은 AWS Batch IAM 실행 역할 단원을 참조하십시오.

생성하는 Secretes Manger 암호에 액세스 권한을 부여하려면 다음 권한을 인라인 정책으로 실행 역할에 수동으로 추가하세요. 자세한 내용은 IAM 사용 설명서의 IAM 정책 추가 및 제거를 참조하세요.

- `secretsmanager:GetSecretValue`–Secrets Manager 암호를 참조하는 경우에 필요합니다.
- `kms:Decrypt`암호가 사용자 지정 KMS 키를 사용하고 기본 키를 사용하지 않는 경우에 필요합니다. 사용자 지정 키의 ARN을 리소스로 추가해야 합니다.

다음 예제에서는 인라인 정책이 필수 권한을 추가합니다.

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "secretsmanager:GetSecretValue",
                "kms:Decrypt"
            ],
            "Resource": [
                "arn:aws:secretsmanager:us-
east-2:777777777777:secret:<secret_name>",
                "arn:aws:kms:us-east-2:777777777777:key/<key_id>"
            ]
        }
    ]
}
```

민감한 데이터를 환경 변수로 삽입

작업 정의 내에서 다음을 지정할 수 있습니다.

- 작업에 설정할 환경 변수의 이름을 포함하는 `secrets` 객체
- Secrets Manager 암호의 Amazon 리소스 이름(ARN)
- 작업에 제공할 중요한 데이터를 포함하는 추가 파라미터

다음 예제에서는 Secrets Manager 암호에 대해 지정해야 하는 전체 구문을 보여 줍니다.

```
arn:aws:secretsmanager:region:aws_account_id:secret:secret-name:json-key:version-stage:version-id
```

다음 섹션에서는 추가 파라미터에 대해 설명합니다. 이 파라미터는 선택 사항입니다. 그러나 사용하지 않는 경우 기본값을 사용하려면 콜론(`:`)을 포함시켜야 합니다. 추가 컨텍스트에 대한 예제가 아래에 나와 있습니다.

`json-key`

환경 변수 값으로 설정할 값과 함께 키-값 쌍의 키 이름을 지정합니다. JSON 형식의 값만 지원됩니다. JSON 키를 지정하지 않으면 암호의 전체 내용이 사용됩니다.

`version-stage`

사용할 암호 버전의 스테이징 레이블을 지정합니다. 버전 스테이징 레이블이 지정된 경우 버전 ID를 지정할 수 없습니다. 버전 단계가 지정되지 않은 경우 기본 동작은 `AWSCURRENT` 스테이징 레이블을 사용하여 암호를 검색하는 것입니다.

스테이징 레이블은 암호가 업데이트되거나 교체되는 경우 암호의 여러 버전을 추적하는 데 사용됩니다. 암호의 각 버전에는 하나 이상의 스테이징 레이블과 ID가 있습니다. 자세한 내용은 AWS Secrets Manager 사용 설명서의 AWS Secrets Manager에 대한 주요 용어 및 개념을 참조하세요.

`version-id`

사용하고자 하는 암호 버전의 고유 식별자를 지정합니다. 버전 ID가 지정된 경우 버전 스테이징 레이블을 지정할 수 없습니다. 버전 ID가 지정되지 않은 경우 기본 동작은 `AWSCURRENT` 스테이징 레이블을 사용하여 암호를 검색하는 것입니다.

버전 ID는 암호가 업데이트되거나 교체되는 경우 암호의 여러 버전을 추적하는 데 사용됩니다. 암호의 각 버전에는 ID가 있습니다. 자세한 내용은 AWS Secrets Manager 사용 설명서의 AWS Secrets Manager에 대한 주요 용어 및 개념을 참조하세요.

컨테이너 정의 예제

다음 예제에서는 컨테이너 정의에서 Secretes Manager 암호를 참조할 수 있는 방법을 보여 줍니다.

Example전체 암호 참조

다음은 Secret Manager 암호의 전체 텍스트를 참조할 때 형식을 나타내는 태스크 정의의 조각입니다.

```
{
  "containerProperties": [{
    "secrets": [{
      "name": "environment_variable_name",
      "valueFrom": "arn:aws:secretsmanager:region:aws_account_id:secret:secret_name-
AbCdEf"
    }]
  }]
}
```

Example암호 내에서 특정 키 참조

다음은 보안 암호의 내용을 관련 버전 스테이징 레이블 및 버전 ID와 함께 표시하는 >get-secret-value 명령의 출력 예를 보여 줍니다.

```
{
    "ARN": "arn:aws:secretsmanager:region:aws_account_id:secret:appauthexample-AbCdEf",
    "Name": "appauthexample",
    "VersionId": "871d9eca-18aa-46a9-8785-981dd39ab30c",
    "SecretString": "{\"username1\":\"password1\",\"username2\":\"password2\",
\"username3\":\"password3\"}",
    "VersionStages": [
        "AWSCURRENT"
    ],
    "CreatedDate": 1581968848.921
}
```

ARN 끝에 키 이름을 지정하여 컨테이너 정의에서 이전 출력의 특정 키를 참조합니다.

```
{
  "containerProperties": [{
    "secrets": [{
      "name": "environment_variable_name",
      "valueFrom": "arn:aws:secretsmanager:region:aws_account_id:secret:appauthexample-
AbCdEf:username1::"
    }]
  }]
}
```

Example특정 비밀 버전 참조

다음은 보안 암호의 암호화되지 않은 내용을 모든 버전의 암호에 대한 메타데이터와 함께 표시하는 >describe-secret 명령의 출력 예입니다.

```
{
    "ARN": "arn:aws:secretsmanager:region:aws_account_id:secret:appauthexample-AbCdEf",
    "Name": "appauthexample",
    "Description": "Example of a secret containing application authorization data.",
    "RotationEnabled": false,
    "LastChangedDate": 1581968848.926,
    "LastAccessedDate": 1581897600.0,
    "Tags": [],
    "VersionIdsToStages": {
        "871d9eca-18aa-46a9-8785-981dd39ab30c": [
            "AWSCURRENT"
        ],
        "9d4cb84b-ad69-40c0-a0ab-cead36b967e8": [
            "AWSPREVIOUS"
        ]
    }
}
```

ARN 끝에 키 이름을 지정하여 컨테이너 정의에서 이전 출력의 특정 버전 스테이징 레이블을 참조합니다.

```
{
  "containerProperties": [{
    "secrets": [{
      "name": "environment_variable_name",
      "valueFrom": "arn:aws:secretsmanager:region:aws_account_id:secret:appauthexample-
AbCdEf::AWSPREVIOUS:"
    }]
  }]
}
```

ARN 끝에 키 이름을 지정하여 컨테이너 정의에서 이전 출력의 특정 버전 ID를 참조합니다.

```
{
  "containerProperties": [{
    "secrets": [{
      "name": "environment_variable_name",
```

```
      "valueFrom": "arn:aws:secretsmanager:region:aws_account_id:secret:appauthexample-
AbCdEf::9d4cb84b-ad69-40c0-a0ab-cead36b967e8"
    }]
  }]
}
```

Example암호의 특정 키 및 버전 스테이징 레이블 참조

다음은 암호 내 특정 키와 특정 버전 스테이징 레이블을 모두 참조하는 방법을 보여줍니다.

```
{
  "containerProperties": [{
    "secrets": [{
      "name": "environment_variable_name",
      "valueFrom": "arn:aws:secretsmanager:region:aws_account_id:secret:appauthexample-
AbCdEf:username1:AWSPREVIOUS:"
    }]
  }]
}
```

특정 키 및 버전 ID를 지정하려면 다음 구문을 사용합니다.

```
{
  "containerProperties": [{
    "secrets": [{
      "name": "environment_variable_name",
      "valueFrom": "arn:aws:secretsmanager:region:aws_account_id:secret:appauthexample-
AbCdEf:username1::9d4cb84b-ad69-40c0-a0ab-cead36b967e8"
    }]
  }]
}
```

로그 구성에 민감한 데이터 삽입

작업 정의 내에서 `logConfiguration`을 지정할 때 컨테이너에 설정할 로그 드라이버 옵션의 이름과 컨테이너에 제공할 민감한 데이터가 들어있는 Secret Manager 암호의 전체 ARN을 사용하여 `secretOptions`를 지정할 수 있습니다.

다음은 Secrets Manager 암호를 참조할 때 형식을 나타내는 작업 정의의 조각입니다.

```
{
```

```
  "containerProperties": [{
    "logConfiguration": [{
      "logDriver": "splunk",
      "options": {
        "splunk-url": "https://cloud.splunk.com:8080"
      },
      "secretOptions": [{
        "name": "splunk-token",
        "valueFrom": "arn:aws:secretsmanager:region:aws_account_id:secret:secret_name-
AbCdEf"
      }]
    }]
  }]
}
```

AWS Secrets Manager 보안 암호 생성

Secrets Manager 콘솔을 사용하여 민감한 데이터에 대한 암호를 생성할 수 있습니다. 자세한 정보는 AWS Secrets Manager 사용 설명서의 기본 암호 생성을 참조하세요.

기본 암호를 생성하는 방법

Secrets Manager를 사용하여 민감한 데이터에 대한 암호를 생성합니다.

1. https://console.aws.amazon.com/secretsmanager/에서 Secrets Manager 콘솔을 엽니다.
2. 새 비밀 저장을 선택합니다.
3. 암호 유형 선택에서 다른 유형의 암호를 선택합니다.
4. 사용자 지정 암호의 세부 정보를 키 및 값 쌍으로 지정합니다. 예를 들어 `UserName`이라는 키를 지정한 다음 그 값으로 적절한 사용자 이름을 입력할 수 있습니다. `Password`라는 이름으로 두 번째 키를 추가하고 그 값으로 암호 텍스트를 추가합니다. 또한 데이터베이스 이름, 서버 주소, TCP 포트 등에 해당하는 항목을 추가할 수도 있습니다. 필요한 정보를 저장하는 데 필요한 만큼 많은 쌍을 추가할 수 있습니다.

   또는 일반 텍스트 탭을 선택하고 원하는 방식으로 암호 값을 입력할 수 있습니다.
5. 보안 AWS KMS 암호의 보호된 텍스트를 암호화하는 데 사용할 암호화 키를 선택합니다. 암호화 키를 선택하지 않으면 Secrets Manager에서는 계정에 대한 기본 키가 있는지 확인하고 있는 경우 해당 키를 사용합니다. 기본 키가 없는 경우 Secrets Manager에서는 자동으로 하나를 생성합니다. 또한 새 키 추가를 선택하여 이 암호에 대한 사용자 지정 KMS를 생성할 수 있습니다. KMS 키를 생성하려면 계정에 KMS 키를 생성할 권한이 있어야 합니다.

6. 다음을 선택합니다.
7. 암호 이름으로 **production/MyAwesomeAppSecret** 또는 **development/TestSecret** 같은 선택 경로와 이름을 입력하고, 다음을 선택합니다. 필요한 경우 설명을 추가하면 나중에 이 암호의 용도를 기억하는 데 도움이 됩니다.

   암호 이름은 ASCII 문자, 숫자 또는 다음 문자 중 하나가 되어야 합니다. /_+=.@-
8. (선택 사항) 이때, 암호에 대한 교체를 구성할 수 있습니다. 이 절차에서는 자동 회전 비활성화 상태로 두고 다음을 선택합니다.

   새 보안 암호 또는 기존 보안 암호에 대한 교체를 구성하는 방법에 대한 자세한 내용은 AWS Secrets Manager 보안 암호 교체를 참조하세요.
9. 설정을 검토한 다음 암호 저장을 선택하여 Secrets Manager에 새 암호로 입력한 모든 항목을 저장합니다.

Systems Manager Parameter Store를 사용하여 민감한 데이터 지정

를 사용하면 Parameter Store 파라미터에 AWS Systems Manager 민감한 데이터를 저장한 다음 컨테이너 정의에서 참조하여 컨테이너에 민감한 데이터를 주입할 AWS Batch수 있습니다.

주제

- Systems Manager Parameter Store를 사용하여 민감한 데이터 지정 시 고려할 사항
- AWS Batch 보안 암호에 필요한 IAM 권한
- 민감한 데이터를 환경 변수로 삽입
- 로그 구성에 민감한 데이터 삽입
- AWS Systems Manager 파라미터 스토어 파라미터 생성

Systems Manager Parameter Store를 사용하여 민감한 데이터 지정 시 고려할 사항

Systems Manager 파라미터 스토어 파라미터를 사용하여 컨테이너에 민감한 데이터를 지정할 때 다음 사항을 고려해야 합니다.

- 이 기능을 사용하려면 컨테이너 인스턴스에 컨테이너 에이전트의 버전 1.23.0 이상이 필요합니다. 그러나 최신 컨테이너 에이전트 버전을 사용하는 것이 좋습니다. 에이전트 버전을 확인하고 최신 버전으로 업데이트하는 방법에 대한 자세한 내용은 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS 컨테이너 에이전트 업데이트를 참조하세요.

- 컨테이너가 처음 시작될 때 작업의 해당 컨테이너에 민감한 데이터가 주입됩니다. 암호 또는 파라미터 스토어 파라미터가 이후에 업데이트되거나 교체되면 컨테이너가 업데이트된 값을 자동으로 받지 않습니다. 업데이트된 암호로 새 작업을 강제로 시작하려면 새 작업을 시작해야 합니다.

AWS Batch 보안 암호에 필요한 IAM 권한

이 기능을 사용하려면 실행 역할이 있어야 하며 작업 정의에서 해당 역할을 참조해야 합니다. 이렇게 하면 Amazon ECS 컨테이너 에이전트가 필요한 AWS Systems Manager 리소스를 가져올 수 있습니다. 자세한 내용은 AWS Batch IAM 실행 역할 단원을 참조하십시오.

생성한 AWS Systems Manager 파라미터 스토어 파라미터에 대한 액세스를 제공하려면 다음 권한을 실행 역할에 인라인 정책으로 수동으로 추가합니다. 자세한 내용은 IAM 사용 설명서의 IAM 정책 추가 및 제거를 참조하세요.

- `ssm:GetParameters` – 태스크 정의에서 Systems Manager 파라미터 스토어 파라미터를 참조하는 경우에 필요합니다.
- `secretsmanager:GetSecretValue` – Secrets Manager 암호를 직접 참조하는 경우 또는 Systems Manager 파라미터 스토어 파라미터가 태스크 정의에서 Secrets Manager 암호를 참조하는 경우에 필요합니다.
- `kms:Decrypt` – 암호가 사용자 지정 KMS 키를 사용하고 기본 키를 사용하지 않는 경우에 필요합니다. 사용자 지정 키의 ARN을 리소스로 추가해야 합니다.

다음 예제에서는 인라인 정책이 필수 권한을 추가합니다.

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "ssm:GetParameters",
                "secretsmanager:GetSecretValue",
                "kms:Decrypt"
            ],
            "Resource": [
                "arn:aws:ssm:us-east-2:999999999999:parameter/<parameter_name>",
```

```
                "arn:aws:secretsmanager:us-
east-2:999999999999:secret:<secret_name>",
                "arn:aws:kms:us-east-2:999999999999:key/<key_id>"
            ]
        }
    ]
}
```

민감한 데이터를 환경 변수로 삽입

컨테이너 정의 내에서 컨테이너에 설정할 환경 변수의 이름으로 `secrets`(을)를 지정하여 컨테이너에 제공할 민감한 데이터가 들어있는 Systems Manager 파라미터 스토어 파라미터의 전체 ARN을 지정합니다.

다음은 Systems Manager 파라미터 스토어 파라미터를 참조할 때 형식을 나타내는 태스크 정의의 조각입니다. Systems Manager 파라미터 스토어 파라미터가 현재 실행 중인 태스크와 동일한 리전에 있는 경우, 파라미터의 전체 ARN 또는 이름을 사용할 수 있습니다. 파라미터가 다른 리전에 있다면 전체 ARN을 지정해야 합니다.

```
{
  "containerProperties": [{
    "secrets": [{
      "name": "environment_variable_name",
      "valueFrom": "arn:aws:ssm:region:aws_account_id:parameter/parameter_name"
    }]
  }]
}
```

로그 구성에 민감한 데이터 삽입

컨테이너 정의 내에서, `logConfiguration`을 정의할 때 컨테이너에 설정할 로그 드라이버 옵션의 이름과 컨테이너에 제공할 민감한 데이터가 들어있는 Systems Manager 파라미터 스토어 파라미터의 전체 ARN을 사용하여 `secretOptions`를 지정할 수 있습니다.

> **Important**
>
> Systems Manager 파라미터 스토어 파라미터가 현재 실행 중인 태스크와 동일한 리전에 있는 경우, 파라미터의 전체 ARN 또는 이름을 사용할 수 있습니다. 파라미터가 다른 리전에 있다면 전체 ARN을 지정해야 합니다.

다음은 Systems Manager 파라미터 스토어 파라미터를 참조할 때 형식을 나타내는 태스크 정의의 조각입니다.

```
{
  "containerProperties": [{
    "logConfiguration": [{
      "logDriver": "fluentd",
      "options": {
        "tag": "fluentd demo"
      },
      "secretOptions": [{
        "name": "fluentd-address",
        "valueFrom": "arn:aws:ssm:region:aws_account_id:parameter/parameter_name"
      }]
    }]
  }]
}
```

AWS Systems Manager 파라미터 스토어 파라미터 생성

AWS Systems Manager 콘솔을 사용하여 민감한 데이터에 대한 Systems Manager 파라미터 스토어 파라미터를 생성할 수 있습니다. 자세한 정보는 AWS Systems Manager 사용 설명서의 연습: 명령에 파라미터 생성 및 사용(콘솔)을 참조하세요.

파라미터 스토어 파라미터를 생성하려면

1. https://console.aws.amazon.com/systems-manager/ AWS Systems Manager 콘솔을 엽니다.
2. 탐색 창에서 파라미터 스토어(Parameter Store), 파라미터 생성(Create parameter)을 차례대로 선택합니다.
3. 이름(Name)에서 계층 구조와 파라미터 이름을 입력합니다. 예를 들어 `test/database_password`를 입력합니다.
4. 설명(Description)에 선택적 설명을 입력합니다.
5. 유형에서 String, StringList 또는 SecureString을 선택합니다.

> **Note**
>
> - SecureString을 선택한 경우, KMS 키 ID 필드가 표시됩니다. KMS 키 ID, KMS 키 ARN, 별칭 이름 또는 별칭 ARN을 제공하지 않으면 시스템에서 `alias/aws/ssm`을 사용합니다. Systems Manager의 기본 KMS 키입니다. 이 키의 사용을 방지하려면 사용자 지정

키를 선택합니다. 자세한 정보는 AWS Systems Manager 사용 설명서의 보안 문자열 파라미터 사용을 참조하세요.

- 사용자 지정 KMS 키 별칭 이름 또는 별칭 ARN과 함께 `key-id` 파라미터를 사용하여 콘솔에 보안 문자열 파라미터를 생성할 때에는 별칭 앞에 접두사 `alias/`를 지정해야 합니다. ARN 예제는 다음과 같습니다.

```
arn:aws:kms:us-east-2:123456789012:alias/MyAliasName
```

별칭 이름 예제는 다음과 같습니다.

```
alias/MyAliasName
```

6. 값(Value)에 값을 입력합니다. 예를 들어 `MyFirstParameter`입니다. SecureString을 선택한 경우, 입력하는 값이 정확하게 마스킹됩니다.
7. 파라미터 생성(Create parameter)을 선택합니다.

작업에 대한 프라이빗 레지스트리 인증

를 사용한 작업에 대한 프라이빗 레지스트리 인증을 AWS Secrets Manager 사용하면 자격 증명을 안전하게 저장한 다음 작업 정의에서 참조할 수 있습니다. 이를 통해 작업 정의에서 인증이 AWS 필요한 외부의 프라이빗 레지스트리에 있는 컨테이너 이미지를 참조할 수 있습니다. 이 기능은 Amazon EC2 인스턴스 및 Fargate에서 호스팅되는 작업에서 지원됩니다.

> **Important**
>
> 작업 정의가 Amazon ECR에 저장된 이미지를 참조하는 경우 이 주제가 적용되지 않습니다. 자세한 정보는 Amazon Elastic Container Registry 사용 설명서의 Amazon ECS에서 Amazon ECR 이미지 사용을 참조하세요.

Amazon EC2 인스턴스에서 호스팅되는 작업의 경우 이 기능을 사용하려면 버전 `1.19.0` 이상의 컨테이너 에이전트가 있어야 합니다. 그러나 최신 컨테이너 에이전트 버전을 사용하는 것이 좋습니다. 에이전트 버전을 확인하고 최신 버전으로 업데이트하는 방법에 대한 자세한 내용은 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS 컨테이너 에이전트 업데이트를 참조하세요.

Fargate에서 호스팅되는 작업의 경우 이 기능을 사용하려면 플랫폼 버전 `1.2.0` 이상이 필요합니다. 자세한 정보는 Amazon Elastic Container Service 개발자 안내서의 AWS Fargate Linux 플랫폼 버전을 참조하세요.

컨테이너 정의에서 자신이 생성한 암호의 세부 정보와 함께 `repositoryCredentials` 객체를 지정합니다. 참조하는 보안 암호는 이를 사용하는 작업과 다른 AWS 리전 계정 또는 다른 계정에서 가져올 수 있습니다.

> **Note**
>
> AWS Batch API AWS CLI또는 AWS SDK를 사용할 때 시작 중인 작업 AWS 리전 과 동일한 에 보안 암호가 있는 경우 보안 암호의 전체 ARN 또는 이름을 사용할 수 있습니다. 암호가 다른 계정에 있는 경우 암호의 전체 ARN을 지정해야 합니다. 를 사용할 때는 보안 암호 AWS Management Console의 전체 ARN을 항상 지정해야 합니다.

다음은 필요한 파라미터를 나타낸 작업 정의의 조각입니다.

```
"containerProperties": [
  {
    "image": "private-repo/private-image",
    "repositoryCredentials": {
      "credentialsParameter":
 "arn:aws:secretsmanager:region:123456789012:secret:secret_name"
    }
  }
]
```

프라이빗 레지스트리 인증에 대한 필수 IAM 권한

실행 역할은 이 기능을 사용해야 합니다. 컨테이너 에이전트는 이 기능을 통해 컨테이너 이미지를 가져올 수 있습니다. 자세한 내용은 AWS Batch IAM 실행 역할 단원을 참조하십시오.

생성하는 암호에 액세스 권한을 부여하려면 다음 권한을 인라인 정책으로 실행 역할에 추가하세요. 자세한 정보는 IAM 정책 추가 및 제거 섹션을 참조하세요.

- `secretsmanager:GetSecretValue`
- `kms:Decrypt`–사용자 키가 기본 KMS 키가 아닌 사용자 지정 KMS 키를 사용하는 경우에만 필요합니다. 사용자 지정 키의 Amazon 리소스 이름(ARN)을 리소스로 추가해야 합니다.

다음 예제에서는 인라인 정책이 권한을 추가합니다.

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "kms:Decrypt",
                "secretsmanager:GetSecretValue"
            ],
            "Resource": [
                "arn:aws:secretsmanager:us-
east-1:123456789012:secret:secret_name",
                "arn:aws:kms:us-east-1:123456789012:key/key_id"
            ]
        }
    ]
}
```

자습서: 프라이빗 레지스트리 인증을 위한 보안 암호 생성

다음 단계를 완료하여를 사용하여 프라이빗 레지스트리 자격 증명에 대한 보안 암호를 생성합니다 AWS Secrets Manager.

기본 보안 암호 생성

1. https://console.aws.amazon.com/secretsmanager/ AWS Secrets Manager 콘솔을 엽니다.
2. 새 비밀 저장을 선택합니다.
3. 암호 유형 선택에서 다른 유형의 암호를 선택합니다.
4. 일반 텍스트(Plaintext)를 선택하고 다음 형식과 같이 프라이빗 레지스트리 자격 증명을 입력합니다.

```
{
  "username" : "privateRegistryUsername",
  "password" : "privateRegistryPassword"
}
```

5. 다음(Next)을 선택합니다.
6. 암호 이름(Secret name)으로 **`production/MyAwesomeAppSecret`** 또는 **`development/TestSecret`** 같은 선택 경로와 이름을 입력하고, 다음(Next)을 선택합니다. 필요한 경우 설명을 추가하면 나중에 이 암호의 용도를 기억하는 데 도움이 됩니다.

   암호 이름은 ASCII 문자, 숫자 또는 다음 문자 중 하나가 되어야 합니다. `/_+=.@-`
7. (선택 사항) 이때, 암호에 대한 교체를 구성할 수 있습니다. 이 절차에서는 자동 회전 비활성화 상태로 두고 다음을 선택합니다.

   새 보안 암호 또는 기존 보안 암호에 대한 교체를 구성하는 방법에 대한 지침은 AWS Secrets Manager 보안 암호 교체를 참조하세요.
8. 설정을 검토한 다음 암호 저장을 선택하여 Secrets Manager에 새 암호로 입력한 모든 항목을 저장합니다.

작업 정의를 등록하고 프라이빗 레지스트리에서 프라이빗 레지스트리 인증을 켭니다. 그런 다음 Secrets Manager ARN 또는 이름에 보안 암호의 Amazon 리소스 이름(ARN)을 입력합니다. 자세한 내용은 프라이빗 레지스트리 인증에 대한 필수 IAM 권한 단원을 참조하십시오.

Amazon EFS 볼륨

Amazon Elastic File System(Amazon EFS)은 AWS Batch 작업에 간단하고 규모 조정이 가능한 파일 스토리지를 제공합니다. Amazon EFS를 사용하면 스토리지 용량이 탄력적입니다. 파일을 추가 및 제거하면 스토리지 용량의 규모가 자동으로 조정됩니다. 애플리케이션에서 스토리지가 필요할 때 필요한 만큼 확보할 수 있습니다.

AWS Batch에서 Amazon EFS 파일 시스템을 사용하여 컨테이너 인스턴스 집합 간에 파일 시스템 데이터를 내보낼 수 있습니다. 이렇게 하면 작업이 동일한 영구 스토리지에 액세스할 수 있습니다. 하지만 Docker 대몬(daemon)을 시작하기 전에 Amazon EFS 파일 시스템을 마운트하도록 컨테이너 인스턴스 AMI를 구성해야 합니다. 또한 작업 정의는 컨테이너 인스턴스에서 볼륨 마운트를 참조하여 파일 시스템을 사용해야 합니다. 다음 섹션은 AWS Batch에서 Amazon EFS를 사용하기 시작하는 데 도움이 됩니다.

Amazon EFS 볼륨 고려 사항

Amazon EFS 볼륨을 사용할 때는 다음 사항을 고려해야 합니다.

- EC2 리소스를 사용하는 작업의 경우, Amazon EFS 파일 시스템 지원이 컨테이너 에이전트 버전이 1.35.0인 Amazon ECS 최적화 AMI 버전 `20191212`의 공개 미리 보기로 추가되었습니다. 그러나

Amazon EFS 파일 시스템 지원은 Amazon EFS 액세스 포인트 및 IAM 권한 부여 기능이 포함된 컨테이너 에이전트 버전이 1.38.0인 Amazon ECS 최적화 AMI 버전 20200319의 정식 출시를 시작했습니다. 이러한 기능을 활용하려면 Amazon ECS 최적화 AMI 버전 20200319 이상을 사용하는 것이 좋습니다. 자세한 내용을 알아보려면 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS 최적화 AMI 버전을 참조하세요.

> **Note**
>
> 자체 AMI를 생성하는 경우 컨테이너 에이전트 1.38.0 이상, `ecs-init` 버전 1.38.0-1 이상을 사용하고 Amazon EC2 인스턴스에서 다음 명령을 실행해야 합니다. 이는 모두 Amazon ECS 볼륨 플러그인을 활성화하기 위한 것입니다. 이 명령은 기본 이미지로 Amazon Linux 2 또는 Amazon Linux를 사용하는지에 따라 달라집니다.
>
> Amazon Linux 2
>
> ```
> $ yum install amazon-efs-utils
> systemctl enable --now amazon-ecs-volume-plugin
> ```
>
> Amazon Linux
>
> ```
> $ yum install amazon-efs-utils
> sudo shutdown -r now
> ```

- 버전 1.4.0 이상의 플랫폼에는 Fargate 리소스를 사용하는 작업에 Amazon EFS 파일 시스템 지원이 추가되었습니다. 자세한 내용을 알아보려면 Amazon Elastic Container Service 개발자 안내서의 AWS Fargate 플랫폼 버전을 참조하세요.
- Fargate 리소스를 사용하는 작업에 Amazon EFS 볼륨 지정 시 Fargate는 Amazon EFS 볼륨을 관리하는 감독자 컨테이너를 생성합니다. 감독자 컨테이너는 소량의 작업 메모리를 사용합니다. 감독자 컨테이너는 태스크 메타데이터 버전 4 엔드포인트를 쿼리할 때 표시됩니다. 자세한 정보는 AWS Fargate에 대한 Amazon Elastic Container Service 사용 설명서의 태스크 메타데이터 엔드포인트 버전 4를 참조하세요.

Amazon EFS 액세스 포인트 사용

Amazon EFS 액세스 포인트는 EFS 파일 시스템에 대한 애플리케이션별 진입점으로, 공유 데이터 세트에 대한 애플리케이션 액세스를 관리하는 것을 지원합니다. Amazon EFS 액세스 포인트와 액세스

제어 방법에 대한 자세한 정보는 Amazon Elastic File System 사용 설명서의 Amazon EFS 액세스 포인트 사용하기를 참조하세요.

액세스 포인트는 액세스 포인트를 통해 이루어지는 모든 파일 시스템 요청에 대해 사용자의 POSIX 그룹을 포함한 사용자 자격 증명을 적용할 수 있습니다. 또한 클라이언트가 지정된 디렉터리 또는 하위 디렉터리의 데이터에만 액세스할 수 있도록 파일 시스템에 대해 다른 루트 디렉터리를 적용할 수 있습니다.

> **Note**
>
> EFS 액세스 포인트를 생성할 때 파일 시스템에서 루트 디렉터리 역할을 하는 경로를 지정합니다. AWS Batch 작업 정의의 액세스 포인트 ID로 EFS 파일 시스템을 참조할 때 루트 디렉터리는 생략하거나 /로 설정할 수 있습니다. 그러면 EFS 액세스 포인트에 설정된 경로가 적용됩니다.

AWS Batch 작업 역할을 사용하여 특정 애플리케이션에서 특정 액세스 포인트를 사용하도록 적용할 수 있습니다. IAM 정책을 액세스 포인트와 결합하면 애플리케이션의 특정 데이터 세트에 안전하게 액세스할 수 있습니다. 이 기능은 태스크 기능에 Amazon ECS IAM 역할을 사용합니다. 자세한 내용은 Amazon Elastic Container Service 개발자 안내서의 태스크에 대한 IAM 역할을 참조하세요.

작업 정의에서 Amazon EFS 파일 시스템 지정

컨테이너에 Amazon EFS 파일 시스템 볼륨을 사용하려면 작업 정의에 볼륨 및 마운트 지점 구성을 지정해야 합니다. 다음 작업 정의 JSON 코드 조각은 컨테이너에 사용할 `volumes` 및 `mountPoints` 객체의 구문을 나타냅니다.

```
{
    "containerProperties": [
        {
            "image": "amazonlinux:2",
            "command": [
                "ls",
                "-la",
                "/mount/efs"
            ],
            "mountPoints": [
                {
                    "sourceVolume": "myEfsVolume",
                    "containerPath": "/mount/efs",
```

```
                    "readOnly": true
                }
            ],
            "volumes": [
                {
                    "name": "myEfsVolume",
                    "efsVolumeConfiguration": {
                        "fileSystemId": "fs-12345678",
                        "rootDirectory": "/path/to/my/data",
                        "transitEncryption": "ENABLED",
                        "transitEncryptionPort": integer,
                        "authorizationConfig": {
                            "accessPointId": "fsap-1234567890abcdef1",
                            "iam": "ENABLED"
                        }
                    }
                }
            ]
        }
    ]
}
```

`efsVolumeConfiguration`

유형: 객체

필수 여부: 아니요

이 파라미터는 Amazon EFS 볼륨을 사용할 때 지정됩니다.

`fileSystemId`

유형: 문자열

필수 항목 여부: 예

사용할 Amazon EFS 파일 시스템 ID입니다.

`rootDirectory`

유형: 문자열

필수 여부: 아니요

호스트 내의 루트 디렉터리로 탑재할 Amazon EFS 파일 시스템 내 디렉터리입니다. 이 파라미터가 생략되면 Amazon EFS 볼륨의 루트가 사용됩니다. /를 지정하면 이 파라미터를 생략하는 것과 동일한 효과가 있습니다. 이름의 최대 길이는 4.096자입니다.

> **Important**
>
> authorizationConfig에서 EFS 액세스 포인트를 지정하는 경우 루트 디렉터리 파라미터를 생략하거나 /로 설정해야 합니다. 그러면 EFS 액세스 포인트에 설정된 경로가 적용됩니다.

`transitEncryption`

유형: 문자열

유효한 값: ENABLED | DISABLED

필수 여부: 아니요

AWS Batch 호스트와 Amazon EFS 서버 간 전송 중 Amazon EFS 데이터에 대한 암호화를 활성화할지 여부를 결정합니다. Amazon EFS IAM 권한 부여를 사용하는 경우 전송 중 데이터 암호화를 활성화해야 합니다. 이 파라미터가 누락되면 DISABLED의 기본값이 사용됩니다. 자세한 내용을 알아보려면 Amazon Elastic File System 사용 설명서의 전송 중 데이터 암호화를 참조하세요.

`transitEncryptionPort`

유형: Integer

필수 여부: 아니요

AWS Batch 호스트와 Amazon EFS 서버 간에 암호화된 데이터를 전송할 때 사용할 포트입니다. 전송 중 데이터 암호화 포트를 지정하지 않으면 Amazon EFS 탑재 헬퍼가 사용하는 포트 선택 전략이 사용됩니다. 이 값은 0~65,535여야 합니다. 자세한 정보는 Amazon Elastic File System 사용 설명서의 EFS 탑재 헬퍼를 참조하세요.

`authorizationConfig`

유형: 객체

필수 여부: 아니요

Amazon EFS 파일 시스템에 대한 권한 부여 구성 세부 정보입니다.

`accessPointId`

유형: 문자열

필수 여부: 아니요

사용할 액세스 포인트 ID입니다. 액세스 포인트를 지정하는 경우 `efsVolumeConfiguration`의 루트 디렉터리 값을 생략하거나 `/`로 설정해야 합니다. 그러면 EFS 액세스 포인트에 설정된 경로가 적용됩니다. 액세스 포인트를 사용하는 경우 `EFSVolumeConfiguration`에서 전송 중 데이터 암호화를 활성화해야 합니다. 자세한 정보는 Amazon Elastic File System 사용 설명서의 Amazon EFS 액세스 포인트 태스크를 참조하세요.

`iam`

유형: 문자열

유효한 값: `ENABLED` | `DISABLED`

필수 여부: 아니요

Amazon EFS 파일 시스템을 탑재할 때 작업 정의에 정의된 AWS Batch 작업 IAM 역할을 사용할지 여부를 결정합니다. 활성화된 경우 `EFSVolumeConfiguration`에서 전송 중 데이터 암호화를 활성화해야 합니다. 이 파라미터가 누락되면 `DISABLED`의 기본값이 사용됩니다. IAM 실행 역할에 대한 자세한 정보는 AWS Batch IAM 실행 역할 섹션을 참조하세요.

Amazon S3 파일 볼륨

S3 파일은 Amazon Simple Storage Service(Amazon S3) 버킷에 저장된 데이터에 대한 직접 파일 시스템 액세스를 제공합니다. 를 사용하면 작업 정의에서 S3 파일 볼륨을 정의하여 컨테이너가 표준 파일 작업을 사용하여 Amazon S3 데이터를 읽고 쓸 수 있도록 AWS Batch할 수 있습니다.

S3 파일 볼륨을 사용하려면 컴퓨팅 환경과 AWS Batch 동일한 VPC에 구성된 S3 파일 시스템 및 탑재 대상이 필요합니다. 버킷 구성, IAM 역할, 파일 시스템 생성 및 탑재 대상을 포함한 전체 설정 지침은 Amazon S3 사용 설명서의 S3 파일 사전 조건 및 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS용 S3 파일 구성을 참조하세요. Amazon S3

Amazon S3 파일 볼륨 고려 사항

S3 파일 볼륨을 사용할 때는 다음 사항을 고려하세요.

- **Important**

  S3 Files는 현재 Amazon EC2 시작 유형에서 지원되지 않습니다. 작업 정의에서 S3 파일 시스템을 구성하고 Amazon EC2 시작 유형에서 실행하려고 하면 시작 시 작업이 실패합니다. Amazon EC2 시작 유형 지원은 향후 릴리스에 계획되어 있습니다.

- 전송 암호화는 S3 파일 볼륨에 대해 항상 활성화됩니다. 선택적으로 `transitEncryptionPort` 파라미터를 사용하여 포트를 지정할 수 있습니다. 기본 포트는 `2049`입니다.
- 작업 역할(Amazon ECS 작업 역할에 해당)에는 파일 시스템에 대한 `s3files:ClientMount` 및 `s3files:ClientWrite` 권한이 있어야 합니다. Amazon S3에서 직접 읽기의 경우 역할에는 버킷에 대한 `s3:GetObjects3:GetObjectVersion`, 및 `s3:ListBucket` 권한도 필요합니다.
- S3 파일 시스템 탑재 대상은 동일한 VPC에 있어야 하며 컴퓨팅 환경의 서브넷에서 연결할 수 있어야 합니다 AWS Batch . 탑재 대상 보안 그룹은 컴퓨팅 환경 보안 그룹의 인바운드 NFS 트래픽(TCP 포트 2049)을 허용해야 합니다.

Amazon S3 파일 액세스 포인트 사용

S3 파일 액세스 포인트는 모든 파일 시스템 요청에 대해 POSIX 사용자 ID 및 루트 디렉터리를 적용하는 파일 시스템의 애플리케이션별 진입점입니다. 각 작업이 공유 파일 시스템 내의 자체 디렉터리에만 액세스할 수 있도록 액세스 포인트를 사용하여 테넌트를 격리할 수 있습니다.

> **Note**
>
> `accessPointArn` 파라미터를 사용하여 액세스 포인트를 지정할 때를 생략하거나 로 설정해야 `rootDirectory` 합니다`/`. 액세스 포인트는 자체 루트 디렉터리 경로를 적용합니다.

액세스 포인트 생성 및 관리에 대한 자세한 내용은 Amazon S3 사용 설명서의 S3 파일 시스템의 액세스 포인트 생성을 참조하세요. Amazon S3 파일 시스템 정책을 사용하여 액세스 포인트 격리를 적용하는 방법에 대한 자세한 내용은 Amazon S3 사용 설명서의 IAM에서 S3 파일이 작동하는 방법을 참조하세요. Amazon S3

작업 정의에서 Amazon S3 파일 시스템 지정

컨테이너에 S3 파일 볼륨을 사용하려면 작업 정의에서 볼륨 및 탑재 지점 구성을 지정해야 합니다. 다음 작업 정의 JSON 코드 조각은 컨테이너에 사용할 `volumes` 및 `mountPoints` 객체의 구문을 나타냅니다.

```
{
    "ecsProperties": {
        "taskProperties": [
            {
                ...,
                "taskRoleArn": "arn:aws:iam::<account>:role/<job-role-name>",
                "containers": [
                    {
                        ...,
                        "mountPoints": [
                            {
                                "sourceVolume": "myS3FilesVolume",
                                "containerPath": "/mnt/s3data",
                                "readOnly": false
                            }
                        ]
                    }
                ],
                "volumes": [
                    {
                        "name": "myS3FilesVolume",
                        "s3filesVolumeConfiguration": {
                            "fileSystemArn": "arn:aws:s3files:<region>:<account>:file-
system/<fs-id>",
                            "rootDirectory": "/keypath/in/s3"
                        }
                    }
                ]
            }
        ]
    }
}
```

`s3filesVolumeConfiguration`

유형: 객체

필수 여부: 아니요

이 파라미터는 S3 파일 볼륨을 사용할 때 지정됩니다.

`fileSystemArn`

유형: 문자열

필수 항목 여부: 예

사용할 S3 파일 시스템의 전체 ARN입니다.

`rootDirectory`

유형: 문자열

필수 항목 여부: 아니요

호스트 내의 루트 디렉터리로 탑재할 S3 파일 시스템 내의 디렉터리입니다. 이 파라미터를 생략하면 파일 시스템의 루트가 사용됩니다. `/`를 지정하면 이 파라미터를 생략하는 것과 동일한 효과가 있습니다. 이름의 최대 길이는 4,096자입니다.

> **Important**
>
> 에 S3 파일 액세스 포인트가 지정된 경우 `accessPointArn`루트 디렉터리 파라미터를 생략하거나 로 설정해야 합니다`/`. 이렇게 하면 액세스 포인트에 설정된 경로가 적용됩니다.

`transitEncryptionPort`

유형: 정수

필수 항목 여부: 아니요

AWS Batch 호스트와 S3 파일 서버 간에 암호화된 데이터를 전송할 때 사용할 포트입니다. 전송 암호화 포트를 지정하지 않으면의 기본값`2049`이 사용됩니다. 이 값은 0~65,535여야 합니다. 전송 암호화는 S3 파일 볼륨에 대해 항상 활성화됩니다.

`accessPointArn`

유형: 문자열

필수 항목 여부: 아니요

사용할 S3 파일 액세스 포인트의 ARN입니다. 액세스 포인트를 지정하는 경우 `s3filesVolumeConfiguration`의 루트 디렉터리 값을 생략하거나 `/`로 설정해야 합니다. 이렇게 하면 액세스 포인트에 설정된 경로가 적용됩니다. 액세스 포인트는 POSIX 사용자 자격 증명을 적용하고 파일 시스템 내의 특정 디렉터리에 대한 액세스를 제한할 수 있습니다. 자세한 내

용은 Amazon S3 사용 설명서의 S3 파일 시스템의 액세스 포인트 생성을 참조하세요. Amazon S3

AWS Batch 및 Amazon EKS에서 S3 파일 볼륨 사용

Amazon EKS 리소스를 사용하는 작업의 경우는 EKS 작업 정의 볼륨 구성`persistentVolumeClaim`의를 통해 S3 파일 볼륨을 AWS Batch 지원합니다. 작업 정의에서 참조하기 전에 Amazon EKS 클러스터에서 영구 볼륨 및 영구 볼륨 클레임을 미리 생성해야 합니다.

다음 작업 정의 코드 조각은 S3 Files 영구 볼륨 클레임을 참조하는 방법을 보여줍니다.

```
{
    "eksProperties": {
        "podProperties": {
            ...,
            "containers": [
                {
                    ...,
                    "volumeMounts": [
                        {
                            "name": "s3files-vol",
                            "mountPath": "/mnt/s3data"
                        }
                    ]
                }
            ],
            "volumes": [
                {
                    "name": "s3files-vol",
                    "persistentVolumeClaim": {
                        "claimName": "<s3files-pvc-name>"
                    }
                }
            ]
        }
    }
}
```

Amazon EKS를 사용하여 S3 파일을 설정하는 방법에 대한 자세한 내용은 Amazon S3 사용 설명서의 Amazon EKS에 S3 파일 시스템 탑재를 참조하세요. Amazon S3 전체 볼륨 파라미터 참조는 AWS Batch API 참조의 EksVolume을 참조하세요.

작업 정의 예

아래 주제의 작업 정의 예제는 환경 변수, 파라미터 대체, 볼륨 마운트 등 공통 패턴을 사용하는 방법에 대해서 설명하고 있습니다.

내용

환경 변수

다음은 환경 변수를 사용하여 파일 형식과 Amazon S3 URL을 지정하는 작업 정의 예제입니다. 여기에서 소개하는 예제는 컴퓨팅 블로그 포스트인 Creating a Simple "Fetch & Run" AWS Batch Job에서 가져왔습니다. 이 블로그 포스트에 설명되어 있는 `fetch_and_run.sh` 스크립트는 환경 변수를 사용하여 `myjob.sh` 스크립트를 S3에서 다운로드한 후 파일 형식을 선언합니다.

이번 예제에서는 명령 및 환경 변수가 작업 정의로 하드 코딩되어 있지만 명령 및 환경 변수 재정의를 지정하면 더욱 다양한 목적에 맞게 작업 정의를 생성할 수 있습니다.

```
{
    "jobDefinitionName": "fetch_and_run",
    "type": "container",
    "containerProperties": {
        "image": "123456789012.dkr.ecr.us-east-1.amazonaws.com/fetch_and_run",
        "resourceRequirements": [
            {
                "type": "MEMORY",
                "value": "2000"
            },
            {
                "type": "VCPU",
                "value": "2"
            }
        ],
        "command": [
            "myjob.sh",
            "60"
```

```
        ],
        "jobRoleArn": "arn:aws:iam::123456789012:role/AWSBatchS3ReadOnly",
        "environment": [
            {
                "name": "BATCH_FILE_S3_URL",
                "value": "s3://amzn-s3-demo-source-bucket/myjob.sh"
            },
            {
                "name": "BATCH_FILE_TYPE",
                "value": "script"
            }
        ],
        "user": "nobody"
    }
}
```

파라미터 대체

다음은 파라미터 대체를 사용하거나, 기본값을 설정하는 방법을 설명하는 작업 정의 예제입니다.

`Ref::` 영역의 `command` 선언은 파라미터 대체를 위한 자리 표시자를 설정하는 데 사용됩니다. 아래 작업 정의로 작업을 제출할 때는 파라미터 재정의를 지정하여 `inputfile`이나 `outputfile` 같은 값을 작성합니다. 아래에서 `parameters` 섹션은 `codec`에 대한 기본값을 설정하지만, 필요에 따라 해당 파라미터를 재정의할 수 있습니다.

자세한 내용은 Parameters 섹션을 참조하세요.

```
{
    "jobDefinitionName": "ffmpeg_parameters",
    "type": "container",
    "parameters": {"codec": "mp4"},
    "containerProperties": {
        "image": "my_repo/ffmpeg",
        "resourceRequirements": [
            {
                "type": "MEMORY",
                "value": "2000"
            },
            {
                "type": "VCPU",
                "value": "2"
            }
        ],
```

```
        "command": [
            "ffmpeg",
            "-i",
            "Ref::inputfile",
            "-c",
            "Ref::codec",
            "-o",
            "Ref::outputfile"
        ],
        "jobRoleArn": "arn:aws:iam::123456789012:role/ECSTask-S3FullAccess",
        "user": "nobody"
    }
}
```

GPU 기능 테스트

다음 예제의 작업 정의는 GPU 워크로드 AMI 사용에서 설명한 GPU 워크로드 AMI가 제대로 구성되었는지를 테스트합니다. 이 예제 작업 정의에서는 GitHub에서 TensorFlow deep MNIST 분류자 예제를 실행합니다.

```
{
    "containerProperties": {
        "image": "tensorflow/tensorflow:1.8.0-devel-gpu",
        "resourceRequirements": [
            {
                "type": "MEMORY",
                "value": "32000"
            },
            {
                "type": "VCPU",
                "value": "8"
            }
        ],
        "command": [
            "sh",
            "-c",
            "cd /tensorflow/tensorflow/examples/tutorials/mnist; python mnist_deep.py"
        ]
    },
    "type": "container",
    "jobDefinitionName": "tensorflow_mnist_deep"
}
```

위의 JSON 텍스트가 포함된 tensorflow_mnist_deep.json이라는 파일을 생성한 후 다음 명령을 사용하여 AWS Batch 작업 정의를 등록할 수 있습니다.

```
aws batch register-job-definition --cli-input-json file://tensorflow_mnist_deep.json
```

다중 노드 병렬 작업

다음 작업 정의 예에서는 다중 노드 병렬 작업을 보여 줍니다. 자세한 내용은 AWS 컴퓨팅 블로그의 AWS Batch 내 다중 노드 병렬 작업으로 긴밀하게 결합된 분자 역학 워크플로 구축을 참조하세요.

```
{
  "jobDefinitionName": "gromacs-jobdef",
  "jobDefinitionArn": "arn:aws:batch:us-east-2:123456789012:job-definition/gromacs-
jobdef:1",
  "revision": 6,
  "status": "ACTIVE",
  "type": "multinode",
  "parameters": {},
  "nodeProperties": {
    "numNodes": 2,
    "mainNode": 0,
    "nodeRangeProperties": [
      {
        "targetNodes": "0:1",
        "container": {
          "image": "123456789012.dkr.ecr.us-east-2.amazonaws.com/gromacs_mpi:latest",
          "resourceRequirements": [
              {
                  "type": "MEMORY",
                  "value": "24000"
              },
              {
                  "type": "VCPU",
                  "value": "8"
              }
          ],
          "command": [],
          "jobRoleArn": "arn:aws:iam::123456789012:role/ecsTaskExecutionRole",
          "ulimits": [],
          "instanceType": "p3.2xlarge"
        }
      }
```

```
      ]
    }
}
```

작업

작업은에서 시작하는 작업 단위입니다 AWS Batch. 작업은 ECS 클러스터의 Amazon ECS 컨테이너 인스턴스에서 실행 중인 컨테이너화된 애플리케이션으로 실행될 수 있습니다.

컨테이너화된 작업은 컨테이너 이미지, 명령 및 파라미터를 참조할 수 있습니다. 자세한 내용은 JobDefinition을 참조하세요.

많은 수의 독립적인 단순 작업을 제출할 수 있습니다.

주제

- 자습서: 작업 제출
- 의 서비스 작업 AWS Batch
- 작업 상태
- AWS Batch 작업 환경 변수
- 작업 자동 재시도
- 작업 종속성
- 작업 제한 시간
- Amazon EKS 작업
- 다중 노드 병렬 작업
- Amazon EKS의 다중 노드 병렬 작업
- 배열 작업
- GPU 작업 실행
- AWS Batch 작업 대기열에서 작업 보기
- 작업 대기열에서 작업 AWS Batch 검색
- AWS Batch 작업에 대한 네트워킹 모드
- CloudWatch Logs에서 AWS Batch 작업 로그 보기
- AWS Batch 작업 정보 검토

자습서: 작업 제출

작업 정의를 등록한 후 AWS Batch 작업 대기열에 작업으로 제출할 수 있습니다. 작업 정의에 지정된 많은 파라미터는 실행 시간에 재정의될 수 있습니다.

작업을 제출하는 방법

1. https://console.aws.amazon.com/batch/ AWS Batch 콘솔을 엽니다.
2. 탐색 모음에서 사용할 AWS 리전 를 선택합니다.
3. 탐색 창에서 작업을 선택합니다.
4. 작업 제출을 선택합니다.
5. 이름(Name)에 고유한 작업 정의 이름을 입력합니다. 각 이름의 최대 길이는 128자입니다. 대문자 및 소문자, 숫자, 하이픈(-) 및 밑줄(_)을 포함할 수 있습니다.
6. 작업 정의에서 작업에 대해 이전에 생성한 작업 정의를 선택합니다. 자세한 내용은 단일 노드 작업 정의 생성 단원을 참조하십시오.
7. 작업 대기열에서 기존 작업 대기열을 선택합니다. 자세한 내용은 작업 대기열 생성 단원을 참조하십시오.
8. 작업 종속성에서 작업 종속성 추가를 선택합니다.
   - 작업 ID에는 모든 종속성에 대한 작업 ID를 입력합니다. 그런 다음 작업 종속성 추가를 선택합니다. 작업에는 최대 20개의 종속성이 있을 수 있습니다. 자세한 내용은 작업 종속성 단원을 참조하십시오.
9. (배열 작업만 해당)배열 크기에서 배열 크기를 2~10,000 사이로 지정합니다.
10. (선택 사항) 태그를 확장한 다음 태그 추가를 선택하여 리소스에 태그를 추가합니다. 키와 선택 값을 입력하고 태그 추가를 선택합니다.
11. 다음 페이지를 선택합니다.
12. 작업 재정의 섹션에서:
    a. (선택 사항) 예약 우선 순위에 0에서 100 사이의 예약 우선 순위 값을 입력합니다. 값이 높을수록 우선 순위가 높습니다.
    b. (선택 사항) 작업 시도에 AWS Batch (이)가 작업을 특정 RUNNABLE 상태로 전환하기 위해 시도하는 최대 횟수를 입력합니다. 1~10 사이의 숫자를 입력합니다. 자세한 내용은 작업 자동 재시도 단원을 참조하십시오.
    c. (선택 사항) 실행 제한 시간에 제한 시간 값(초)을 입력합니다. 실행 제한 시간은 완료되지 않은 작업이 종료되기까지의 시간입니다. 시도가 제한 시간을 초과하면 중지되고 상태가 FAILED(으)로 변경됩니다. 자세한 내용은 작업 제한 시간 단원을 참조하십시오. 최솟값은 60초입니다.

> **Important**
>
> Fargate 리소스에서 실행되는 작업이 14일 이상 실행될 것이라고 기대하지 마세요. 14일이 지나면 작업이 종료되어 Fargate 리소스를 더 이상 사용할 수 없게 될 수 있습니다.

d. (선택 사항) 작업 및 작업 정의에서 Amazon ECS 태스크로 태그를 전파하려면 태그 전파를 활성화합니다.

13. 추가 구성을 확장합니다.
14. (선택 사항) 재시도 전략 조건의 경우 종료 시 평가 추가를 선택합니다. 파라미터 값을 하나 이상 입력한 다음 작업을 선택합니다. 각 조건 세트에 대해 작업을 재시도 또는 종료로 설정해야 합니다. 이러한 작업은 다음을 의미합니다.

    - 재시도 - 지정한 작업 시도 횟수에 도달할 때까지 AWS Batch 재시도합니다.
    - 종료 - 작업 재시도를 AWS Batch 중지합니다.

    > **Important**
    >
    > 종료 시 평가 추가를 선택한 경우 하나 이상의 파라미터를 구성하고 작업을 선택하거나 종료 시 평가 제거를 선택합니다.

15. 파라미터에서 파라미터 추가를 선택하여 파라미터 대체 자리 표시자를 추가합니다. 키와 선택 사항으로 값을 입력합니다.
16. 컨테이너 재정의의 섹션에서:

    a. 명령에서 필드에 명령을 JSON 문자열 배열 형식으로 입력합니다.

    이 파라미터는 도커 원격 API(Docker Remote API)의 컨테이너 생성(Create a container) 섹션에 있는 Cmd(와)과 docker run의 COMMAND 파라미터로 매핑됩니다. 도커 CMD 파라미터에 대한 자세한 정보는 https://docs.docker.com/engine/reference/builder/#cmd를 참조하세요.

    > **Note**
    >
    > 이 파라미터는 빈 문자열을 포함할 수 없습니다.

b. vCPU에서 컨테이너에 예약할 vCPU 수를 지정합니다. 이 파라미터는 Docker 원격 API(Docker Remote API)의 컨테이너 생성(Create a container) 섹션에 있는 `CpuShares`(와)과 docker run에 대한 `--cpu-shares` 옵션에 매핑됩니다. 각 vCPU는 1,024개의 CPU 공유와 동일합니다. vCPU를 최소 하나 이상 지정해야 합니다.

c. 메모리에는 컨테이너에 사용할 수 있는 메모리 한도를 입력합니다. 컨테이너가 여기에 지정된 메모리를 초과하려 하면 해당 컨테이너가 중지됩니다. 이 파라미터는 Docker 원격 API(Docker Remote API)의 컨테이너 생성(Create a container) 섹션에 있는 `Memory`(와)과 docker run에 대한 `--memory` 옵션에 매핑됩니다. 한 작업에 대해 메모리를 최소한 4MiB 지정해야 합니다.

> **Note**
>
> 리소스 사용률을 극대화하려면 특정 인스턴스 유형의 작업에 메모리 우선 순위를 지정합니다. 자세한 내용은 컴퓨팅 리소스 메모리 관리 단원을 참조하십시오.

d. (선택 사항)GPU 수에 컨테이너에 예약할 GPU 수를 선택합니다.

e. (선택 사항) 환경 변수의 경우 환경 변수 추가를 선택하여 환경 변수를 이름-값 쌍으로 추가합니다. 이러한 변수는 컨테이너로 전달됩니다.

f. 다음 페이지를 선택합니다.

g. 작업 검토(Job review)에서 구성 단계를 검토하세요. 변경해야 하는 경우 편집을 선택합니다 작업을 마쳤으면 작업 정의 생성을 선택합니다.

의 서비스 작업 AWS Batch

AWS Batch 서비스 작업을 사용하면 AWS Batch 작업 대기열을 통해 AWS 서비스에 요청을 제출할 수 있습니다. 현재는 SageMaker 훈련 작업을 서비스 작업으로 AWS Batch 지원합니다. 가 기본 컨테이너 실행을 AWS Batch 관리하는 컨테이너화된 작업과 달리 서비스 작업은 대상 서비스 AWS (예: SageMaker AI)가 실제 작업 실행을 처리하는 동안가 작업 예약 및 대기열 기능을 AWS Batch 제공할 수 있도록 허용합니다.

AWS Batch SageMaker 훈련 작업용를 사용하면 데이터 과학자가 우선순위가 있는 훈련 작업을 구성 가능한 대기열에 제출할 수 있으므로 리소스를 사용할 수 있게 되는 즉시 워크로드가 개입 없이 실행되도록 할 수 있습니다. 이 기능은 리소스 조정, 우발적인 과다 지출 방지, 예산 제약 조건 충족, 예약 인스턴스 비용 최적화, 팀원 간의 수동 조정 필요 제거와 같은 일반적인 문제를 해결합니다.

서비스 작업은 다음과 같은 몇 가지 주요 측면에서 컨테이너화된 작업과 다릅니다.

- 작업 제출: 서비스 작업은 SubmitServiceJob API를 사용하여 제출해야 합니다. 서비스 작업은 AWS Batch 콘솔을 통해 제출할 수 없습니다.
- 작업 실행: 서비스 작업을 AWS Batch 예약하고 대기열에 추가하지만 대상 AWS 서비스는 실제 작업 워크로드를 실행합니다.
- 리소스 식별자: 서비스 작업은 'job' 대신 'service-job'이 포함된 ARN을 사용하여 컨테이너화된 작업과 구분됩니다.

SageMaker 훈련을 위한 AWS Batch 서비스 작업을 시작하려면 섹션을 참조하세요the section called "SageMaker AI AWS Batch 에서 시작하기".

주제

- 의 서비스 작업 페이로드 AWS Batch
- 에서 서비스 작업 제출 AWS Batch
- AWS Batch 서비스 작업 상태를 SageMaker AI 상태로 매핑
- 의 서비스 작업 재시도 전략 AWS Batch
- AWS Batch 대기열의 서비스 작업 모니터링
- 서비스 작업 종료

의 서비스 작업 페이로드 AWS Batch

SubmitServiceJob을 사용하여 서비스 작업을 제출할 때는 작업을 정의하는 두 가지 주요 파라미터인 `serviceJobType`과 `serviceRequestPayload`를 제공합니다.

- 는 작업을 실행할 AWS 서비스를 `serviceJobType` 지정합니다. SageMaker 훈련 작업의 경우 이 값은 `SAGEMAKER_TRAINING`입니다.
- `serviceRequestPayload`는 일반적으로 대상 서비스로 직접 전송되는 전체 요청을 포함하는 JSON 인코딩 문자열입니다. SageMaker 훈련 작업의 경우 이 페이로드에는 SageMaker AI CreateTrainingJob API와 함께 사용되는 것과 동일한 파라미터가 포함됩니다.

사용 가능한 모든 파라미터의 전체 목록과 해당 설명은 SageMaker AI CreateTrainingJob API 레퍼런스를 참조하세요. `CreateTrainingJob`에서 지원하는 모든 파라미터를 서비스 작업 페이로드에 포함할 수 있습니다.

추가 훈련 작업 구성의 예제는 SageMaker AI 개발자 안내서의 API, CLI 및 SDK를 참조하세요.

PySDK에는 헬퍼 클래스와 유틸리티가 있으므로 서비스 작업 생성에 PySDK를 사용할 것이 권장됩니다. PySDK 사용에 대한 예제는 GitHub의 SageMaker AI 예제를 참조하세요.

서비스 작업 페이로드 예제

다음 예제는 'hello world' 훈련 스크립트를 실행하는 SageMaker 훈련 작업에 대한 간단한 서비스 작업 페이로드를 보여줍니다.

이 페이로드는 `SubmitServiceJob`을 호출할 때 `serviceRequestPayload` 파라미터에 JSON 문자열로 전달됩니다.

```
{
  "TrainingJobName": "my-simple-training-job",
  "RoleArn": "arn:aws:iam::123456789012:role/SageMakerExecutionRole",
  "AlgorithmSpecification": {
    "TrainingInputMode": "File",
    "TrainingImage": "763104351884.dkr.ecr.us-west-2.amazonaws.com/pytorch-
training:2.0.0-cpu-py310",
    "ContainerEntrypoint": [
      "echo",
      "hello world"
    ]
  },
  "ResourceConfig": {
    "InstanceType": "ml.c5.xlarge",
    "InstanceCount": 1,
    "VolumeSizeInGB": 1
  },
  "OutputDataConfig": {
    "S3OutputPath": "s3://your-output-bucket/output"
  },
  "StoppingCondition": {
    "MaxRuntimeInSeconds": 30
  }
}
```

에서 서비스 작업 제출 AWS Batch

서비스 작업을 제출하려면 SubmitServiceJob API를 AWS Batch사용합니다. AWS CLI 또는 SDK를 사용하여 작업을 제출할 수 있습니다.

아직 실행 역할이 없는 경우 먼저 실행 역할을 생성해야 서비스 작업을 제출할 수 있습니다. SageMaker AI 실행 역할을 생성하려면 SageMaker AI 개발자 안내서의 SageMaker AI 실행 역할을 사용하는 방법을 참조하세요.

서비스 작업 제출 워크플로

서비스 작업을 제출하면는 다음 워크플로를 AWS Batch 따릅니다.

1. AWS Batch 는 `SubmitServiceJob` 요청을 수신하고 AWS Batch특정 파라미터를 검증합니다. `serviceRequestPayload`는 검증 없이 전달됩니다.
2. 작업이 `SUBMITTED` 상태로 전환되고 지정된 작업 대기열에 배치됩니다.
3. AWS Batch 는 대기열 앞에 있는 `RUNNABLE` 작업에 대해 서비스 환경에 사용 가능한 용량이 있는지 평가합니다.
4. 용량을 사용할 수 있는 경우 작업이 `SCHEDULED`로 이동하고 작업이 SageMaker AI로 전달됩니다.
5. 용량을 획득하고 SageMaker AI가 서비스 작업 데이터를 다운로드하면 서비스 작업이 초기화를 시작하고 작업이 `STARTING`으로 변경됩니다.
6. SageMaker AI가 작업을 실행하기 시작하면 상태가 `RUNNING`으로 변경됩니다.
7. SageMaker AI가 작업을 실행하는 동안는 진행 상황을 AWS Batch 모니터링하고 서비스 상태를 AWS Batch 작업 상태에 매핑합니다. 서비스 작업 상태가 매핑되는 방식에 대한 자세한 내용은 ??? 섹션을 참조하세요.
8. 서비스 작업이 완료되면 `SUCCEEDED`로 이동하고 출력을 다운로드할 준비가 됩니다.

사전 조건

서비스 작업을 제출하기 전에 다음을 갖추어야 합니다.

- 서비스 환경 - 용량 제한을 정의하는 서비스 환경. 자세한 내용은 에서 서비스 환경 생성 AWS Batch 단원을 참조하십시오.
- SageMaker 작업 대기열 - 작업 예약을 제공하는 SageMaker 작업 대기열. 자세한 내용은 AWS Batch에서 SageMaker 훈련 작업 대기열 생성 단원을 참조하십시오.
- IAM 권한 - AWS Batch 작업 대기열 및 서비스 환경을 생성하고 관리할 수 있는 권한. 자세한 내용은 AWS Batch IAM 정책, 역할 및 권한 단원을 참조하십시오.

서비스 작업 제출

아래 표는 SageMaker Python SDK 또는 AWS CLI를 사용하여 서비스 작업을 제출하는 방법을 보여줍니다.

Submit using the SageMaker Python SDK

SageMaker Python SDK는에 작업 제출을 기본적으로 지원합니다 AWS Batch. 다음 예제에서는 모델 트레이너를 생성하고, 훈련 대기열을 생성하고, 작업을 제출하는 방법을 보여줍니다. 전체 예제는 GitHub의 전체 샘플 노트북을 참조하세요.

훈련 작업 구성을 `ModelTrainer` 정의하는를 생성합니다.

```
from sagemaker.train.model_trainer import ModelTrainer
from sagemaker.train.configs import SourceCode, Compute, StoppingCondition

source_code = SourceCode(command="echo 'Hello World'")

model_trainer = ModelTrainer(
    training_image="123456789012.dkr.ecr.us-east-1.amazonaws.com/pytorch-
training:2.5-gpu-py311",
    source_code=source_code,
    base_job_name="my-training-job",
    compute=Compute(instance_type="ml.g5.xlarge", instance_count=1),
    stopping_condition=StoppingCondition(max_runtime_in_seconds=300),
)
```

작업 대기열을 이름으로 참조하는 `TrainingQueue` 객체를 생성합니다.

```
from sagemaker.train.aws_batch.training_queue import TrainingQueue

queue = TrainingQueue("my-sagemaker-job-queue")
```

를 호출하여 작업을 제출합니다`queue.submit`.

```
job = queue.submit(
    training_job=model_trainer,
    inputs=None,
)
```

Submit using the AWS CLI

다음은 AWS CLI를 사용하여 서비스 작업을 제출하는 방법을 보여줍니다.

```
aws batch submit-service-job \
    --job-name "my-sagemaker-training-job" \
    --job-queue "my-sagemaker-job-queue" \
    --service-job-type "SAGEMAKER_TRAINING" \
    --service-request-payload '{\"TrainingJobName\": \"sagemaker-
training-job-example\", \"AlgorithmSpecification\": {\"TrainingImage\":
 \"123456789012.dkr.ecr.us-east-1.amazonaws.com/pytorch-inference:1.8.0-cpu-
py3\", \"TrainingInputMode\": \"File\", \"ContainerEntrypoint\":  [\"sleep\",
 \"1\"]}, \"RoleArn\":\"arn:aws:iam::123456789012:role/SageMakerExecutionRole\",
 \"OutputDataConfig\": {\"S3OutputPath\": \"s3://example-bucket/model-output/\"},
 \"ResourceConfig\": {\"InstanceType\": \"ml.m5.large\", \"InstanceCount\": 1,
 \"VolumeSizeInGB\": 1}}'
    --client-token "unique-token-12345"
```

`serviceRequestPayload` 파라미터에 대한 자세한 내용은 the section called “서비스 작업 페이로드”을 참조하세요.

AWS Batch 서비스 작업 상태를 SageMaker AI 상태로 매핑

SubmitServiceJob을 사용하여 SageMaker 작업 대기열에 작업을 제출하면는 작업 수명 주기를 AWS Batch 관리하고 AWS Batch 작업 상태를 동등한 SageMaker 훈련 작업 상태로 매핑합니다. SageMaker 훈련 작업과 같은 서비스 작업은 기존 컨테이너 작업과 다른 상태 수명 주기를 따릅니다. 서비스 작업은 대부분의 상태를 컨테이너 작업과 공유하지만 `SCHEDULED` 상태를 사용하며 특히 대상 서비스의 용량 부족 오류 처리 등을 위해 다양한 재시도 동작을 보여줍니다.

다음 표에는 AWS Batch 작업 상태와 해당 SageMaker 상태/SecondaryStatus가 나와 있습니다.

Batch 상태	SageMaker AI 기본 상태	SageMaker AI 보조 상태	설명
`SUBMITTED`	해당 사항 없음	해당 사항 없음	작업이 대기열에 제출되었으며 스케줄러 평가를 기다리는 중입니다.
`RUNNABLE`	해당 사항 없음	해당 사항 없음	작업이 대기열에 있고 예약 준비가 되었습니다. 이 상태의 작업은 작업 대기열에 매핑된 서비스 환경에 충분한 리소스를 사용할

Batch 상태	SageMaker AI 기본 상태	SageMaker AI 보조 상태	설명
			수 있게 되면 바로 시작됩니다. 사용 가능한 리소스가 충분하지 않으면 작업이 이 상태로 무기한 남아 있을 수 있습니다.
SCHEDULED	InProgress	Pending	서비스 작업이 SageMaker AI에 성공적으로 제출되었습니다.
STARTING	InProgress	Downloading	SageMaker 훈련 작업이 데이터 및 이미지를 다운로드하는 중입니다. 훈련 작업 용량이 획득되고 작업 초기화가 시작됩니다.
RUNNING	InProgress	Training	SageMaker 훈련 작업 실행 알고리즘
RUNNING	InProgress	Uploading	SageMaker 훈련 작업이 훈련 완료 후 출력 아티팩트를 업로드하는 중입니다.
SUCCEEDED	Completed	Completed	SageMaker 훈련 작업이 성공적으로 완료되었습니다. 출력 아티팩트의 업로드가 완료되었습니다.
FAILED	Failed	Failed	SageMaker 훈련 작업에서 복구할 수 없는 오류가 발생했습니다.
FAILED	Stopped	Stopped	StopTrainingJob 을 사용하여 SageMaker 훈련 작업을 수동으로 중지했습니다.

의 서비스 작업 재시도 전략 AWS Batch

서비스 작업 재시도 전략을 사용하면 AWS Batch 가 특정 조건에서 실패한 서비스 작업을 자동으로 재시도할 수 있습니다.

서비스 작업은 몇 가지 이유로 여러 번 시도해야 할 수 있습니다.

- 일시적인 서비스 문제: 내부 서비스 오류, 스로틀링 또는 일시적인 중단으로 인해 제출 또는 실행 중에 작업이 실패할 수 있습니다.

- 훈련 초기화 실패: 이미지 가져오기 문제 또는 초기화 오류와 같은 작업 시작 중 문제는 재시도를 통해 해결될 수 있습니다.

적절한 재시도 전략을 구성하면 특히 장기 실행 훈련 워크로드의 경우 작업 성공률을 높이고 수동 개입의 필요성을 줄일 수 있습니다.

> **Note**
>
> 용량 부족으로 인해 SageMaker 훈련 작업이 실패하면는 용량을 사용할 수 있을 때까지 작업을 자동으로 AWS Batch 재시도합니다. 용량 부족으로 인해 실패한 최신 SageMaker 훈련 작업은 유지되며 용량 부족으로 인해 실패한 이전 SageMaker 훈련 작업은 최선의 노력을 다해 삭제됩니다.
> 이러한 용량 기반 재시도는 구성된 재시도를 사용하지 않습니다. 재시도 전략은 주로 작업이 'STARTING'으로 전환된 후 발생하는 알고리즘 오류 또는 서비스 문제와 같은 다른 유형의 실패를 처리합니다.

재시도 전략 구성

서비스 작업 재시도 전략은 단순 재시도 횟수와 조건부 재시도 로직을 모두 지원하는 ServiceJobRetryStrategy를 사용하여 구성됩니다.

재시도 구성

가장 간단한 재시도 전략은 서비스 작업이 실패할 경우 수행해야 하는 재시도 횟수를 지정합니다.

```
{
  "retryStrategy": {
    "attempts": 3
  }
}
```

이 구성을 사용하면 서비스 작업이 실패할 경우 최대 3회까지 재시도할 수 있습니다.

> **Important**
>
> `attempts` 값은 초기 시도를 포함하여 작업을 `RUNNABLE` 상태에 배치할 수 있는 총 횟수를 나타냅니다. 값이 3이면 작업이 처음에 한 번 시도된 다음 실패할 경우 최대 2회 더 시도됩니다.

evaluateOnExit를 사용하여 구성 재시도

`evaluateOnExit` 파라미터를 사용하여 작업을 재시도하거나 실패하도록 허용하는 조건을 지정할 수 있습니다. 이는 다양한 유형의 장애에 서로 다른 처리가 필요한 경우에 유용합니다.

`evaluateOnExit` 배열은 최대 5개의 재시도 전략을 포함할 수 있으며, 각 전략은 상태 사유를 기반으로 작업(`RETRY` 또는 `EXIT`)과 조건을 지정합니다.

```
{
  "retryStrategy": {
    "attempts": 5,
    "evaluateOnExit": [
      {
        "action": "RETRY",
        "onStatusReason": "Received status from SageMaker: InternalServerError*"
      },
      {
        "action": "EXIT",
        "onStatusReason": "Received status from SageMaker: ValidationException*"
      },
      {
        "action": "EXIT",
        "onStatusReason": "*"
      }
    ]
  }
}
```

이 구성은 다음과 같습니다.

- SageMaker AI 내부 서버 오류로 인해 실패한 작업에 대해 재시도 실행
- 검증 예외(다시 시도해도 해결되지 않는 클라이언트 오류)가 발생하는 작업은 즉시 실패
- 다른 모든 장애 유형에 대해 종료 처리하는 포괄적 규칙 포함

상태 사유 패턴 일치

`onStatusReason` 파라미터는 최대 512자의 패턴 일치를 지원합니다. 패턴은 와일드카드(*)를 사용하고 SageMaker AI에서 반환한 상태 사유에 일치시킬 수 있습니다.

서비스 작업의 경우 SageMaker AI의 상태 메시지에는 "SageMaker에서 수신된 상태: " 접두사가 붙어 AWS Batch생성된 메시지와 구분됩니다. 일반적인 패턴은 다음과 같습니다.

- `Received status from SageMaker: InternalServerError*` - 내부 서비스 오류 일치
- `Received status from SageMaker: ValidationException*` - 클라이언트 검증 오류 일치
- `Received status from SageMaker: ResourceLimitExceeded*` - 리소스 제한 오류 일치
- `*CapacityError*` - 용량 관련 실패 일치

> **Tip**
>
> 특정 패턴 일치를 사용하면 다양한 오류 유형을 적절하게 처리할 수 있습니다. 예를 들어 내부 서버 오류는 재시도하지만 작업 파라미터 문제를 나타내는 검증 오류는 즉시 실패 처리할 수 있습니다.

AWS Batch 대기열의 서비스 작업 모니터링

`list-service-jobs` 및 `get-job-queue-snapshot`을 사용하여 SageMaker 훈련 작업 대기열의 작업 상태를 모니터링할 수 있습니다.

대기열에서 실행 중인 작업 보기:

```
aws batch list-service-jobs \
  --job-queue my-sm-training-fifo-jq \
  --job-status RUNNING
```

대기열에서 대기 중인 작업을 봅니다.

```
aws batch list-service-jobs \
  --job-queue my-sm-training-fifo-jq \
  --job-status RUNNABLE
```

SageMaker에 제출되었지만 아직 실행되지 않은 작업 보기:

```
aws batch list-service-jobs \
  --job-queue my-sm-training-fifo-jq \
  --job-status SCHEDULED
```

대기열 앞에 있는 작업의 스냅샷을 가져옵니다.

```
aws batch get-job-queue-snapshot --job-queue my-sm-training-fifo-jq
```

이 명령은 대기열에서 예정된 서비스 작업의 순서를 보여줍니다.

자세한 서비스 작업 정보 가져오기

DescribeServiceJob 작업을 사용하여 현재 상태, 서비스 리소스 식별자 및 자세한 시도 정보를 포함하여 특정 서비스 작업에 대한 포괄적인 정보를 가져옵니다.

특정 작업에 대한 세부 정보를 봅니다.

```
aws batch describe-service-job \
  --job-id a4d6c728-8ee8-4c65-8e2a-9a5e8f4b7c3d
```

이 명령은 다음을 포함하여 작업에 대한 포괄적인 정보를 반환합니다.

- 작업 ARN 및 현재 상태
- 서비스 리소스 식별자(예: SageMaker 훈련 작업 ARN)
- 예약 우선 순위 및 재시도 구성
- 원래 서비스 파라미터를 포함하는 서비스 요청 페이로드
- 시작 및 중지 시간이 포함된 자세한 시도 정보
- 대상 서비스의 상태 메시지

SageMaker 훈련 작업 모니터링

를 통해 SageMaker 훈련 작업을 모니터링할 때 AWS Batch 작업 정보와 기본 SageMaker 훈련 작업 세부 정보에 모두 액세스할 AWS Batch수 있습니다.

작업 세부 정보의 서비스 리소스 식별자에는 SageMaker 훈련 작업 ARN이 포함됩니다.

```
{
  "latestAttempt": {
    "serviceResourceId": {
      "name": "TrainingJobArn",
      "value": "arn:aws:sagemaker:us-east-1:123456789012:training-job/my-training-job"
    }
  }
```

```
}
```

이 ARN을 사용하여 SageMaker에서 추가 세부 정보를 직접 가져올 수 있습니다.

```
aws sagemaker describe-training-job \
  --training-job-name my-training-job
```

AWS Batch 상태와 SageMaker 훈련 작업 상태를 모두 확인하여 작업 진행 상황을 모니터링합니다. AWS Batch 작업 상태에는 전체 작업 수명 주기를 표시하는 반면, SageMaker 훈련 작업 상태는 훈련 프로세스에 대한 서비스별 세부 정보를 제공합니다.

서비스 작업 종료

TerminateServiceJob 작업을 사용하여 실행 중인 서비스 작업을 중지합니다.

특정 서비스 작업을 종료합니다.

```
aws batch terminate-service-job \
  --job-id a4d6c728-8ee8-4c65-8e2a-9a5e8f4b7c3d \
  --reason "Job terminated by user request"
```

서비스 작업을 종료하면가 작업을 AWS Batch 중지하고 대상 서비스에 알립니다. SageMaker 훈련 작업의 경우, SageMaker AI의 훈련 작업도 중지됩니다.

작업 상태

작업을 작업 AWS Batch 대기열에 제출하면 작업이 SUBMITTED 상태로 전환됩니다. 그런 다음 작업은 성공(0 코드와 함께 종료)하거나 실패(0이 아닌 코드와 함께 종료)할 때까지 다음 상태를 통과합니다. AWS Batch 작업은 다음 상태일 수 있습니다.

SUBMITTED

대기열에 제출되었으며 스케줄러에서 아직 평가되지 않은 작업입니다. 스케줄러는 작업을 평가하여 다른 작업이 성공적으로 완료되어야 하는 종속성이 남아 있는지를 확인합니다. 종속성이 있으면 작업이 PENDING 상태로 됩니다. 종속성이 없으면 작업이 RUNNABLE 상태로 됩니다.

PENDING

대기열에 있지만 다른 작업이나 리소스에 대한 종속성으로 인해 아직 실행할 수 없는 작업입니다. 종속성이 해결되면 작업이 RUNNABLE 상태로 됩니다.

> **Note**
>
> 배열 작업 상위는 하위 작업이 로 업데이트될 PENDING 때 로 업데이트되고 하위 작업이 실행되는 동안 PENDING 상태를 RUNNABLE 유지합니다. 이러한 작업을 보려면 모든 하위 작업이 터미널 상태에 도달할 때까지 PENDING 상태를 기준으로 필터링합니다.

RUNNABLE

대기열에 있으며 남아 있는 종속성이 없어서 호스트로 예약된 작업입니다. 이 상태의 작업은 해당 작업 대기열에 매핑된 컴퓨팅 환경 중 하나에서 리소스가 충분해지자마자 시작됩니다. 그러나 사용 가능한 리소스가 충분하지 않으면 작업이 이 상태로 무기한 남아 있을 수 있습니다.

> **Note**
>
> 작업이 STARTING(으)로 진행되지 않으면 문제 해결 섹션의 RUNNABLE 상태에서 정체된 작업을 참조하세요.

STARTING

이러한 작업은 호스트로 일정이 예약되었고 관련 컨테이너 개시 작업이 진행 중입니다. 컨테이너 이미지를 가져와서 컨테이너가 가동 및 실행되면 작업이 RUNNING 상태로 전환됩니다.

이미지 가져오기 지속 시간, Amazon EKS initContainers 완료 지속 시간 및 Amazon ECS containerDependency 해결 지속 시간은 STARTING 상태에서 발생합니다. 작업의 이미지를 가져오는 데 소요되는 시간은 작업이 STARTING 상태에 있는 시간과 동일합니다.

예를 들어, 작업의 이미지를 가져오는 데 3분이 걸리는 경우 작업은 3분 동안 STARTING 상태가 됩니다. initContainers를 완료하는 데 총 10분이 걸리면 Amazon EKS 작업은 10분 동안 STARTING 상태가 됩니다. Amazon ECS 작업에 Amazon ECS containerDependencies 세트가 있는 경우 모든 컨테이너 종속성(런타임)이 해결될 때까지 작업이 STARTING 상태가 됩니다. STARTING 상태는 제한 시간에 포함되지 않습니다. 이 기간은 RUNNING 상태에서 시작됩니다. 자세한 내용은 작업 상태 섹션을 참조하세요.

RUNNING

작업이 컴퓨팅 환경 내의 Amazon ECS 컨테이너 인스턴스에서 컨테이너 작업으로 실행 중입니다. 작업의 컨테이너가 종료되면 프로세스가 종료 코드에 따라 작업의 성공 또는 실패가 결정됩니다. 종료 코드 0(은)는 성공을 나타내고, 0이 아닌 다른 코드는 실패를 나타냅니다. 실패한 시도와 연결

된 작업의 재시도 전략 구성(선택 사항)에 재시도 횟수가 남아 있으면 작업이 다시 RUNNABLE 상태로 됩니다. 자세한 내용은 작업 자동 재시도 단원을 참조하십시오.

> **Note**
>
> RUNNING 작업 로그는 CloudWatch Logs에서 확인할 수 있습니다. 로그 그룹은 /aws/batch/job이며 로그 스트림 이름 형식은 *first200CharsOfJobDefinitionName*/default/*ecs_task_id*입니다. 그러나 이 점은 추후 개선될 것입니다.
> 작업이 RUNNING 상태에 도달한 후에는 DescribeJobs API 작업을 사용하여 프로그래밍 방식으로 해당 로그 스트림 이름을 검색할 수 있습니다. 자세한 내용은 Amazon CloudWatch Logs 사용 설명서의 CloudWatch Logs에 보낸 로그 데이터 보기 섹션을 참조하세요. 기본적으로 이러한 로그는 만료되지 않습니다. 그러나 백업 보존 기간은 수정할 수 있으며, 자세한 내용은 Amazon CloudWatch Logs 사용 설명서의 CloudWatch에서 로그 데이터 보존 기간 변경을 참조하세요.

SUCCEEDED

작업이 종료 코드 0(와)과 함께 성공적으로 완료되었습니다. 작업의 작업 상태는 최소 7일 AWS Batch 동안에 SUCCEEDED 유지됩니다.

> **Note**
>
> SUCCEEDED 작업 로그는 CloudWatch Logs에서 확인할 수 있습니다. 로그 그룹은 /aws/batch/job이며 로그 스트림 이름 형식은 *first200CharsOfJobDefinitionName*/default/*ecs_task_id*입니다. 이 형식은 향후 변경될 수 있습니다.
> 작업이 RUNNING 상태에 도달한 후에는 DescribeJobs API 작업을 사용하여 프로그래밍 방식으로 해당 로그 스트림 이름을 검색할 수 있습니다. 자세한 내용은 Amazon CloudWatch Logs 사용 설명서의 CloudWatch Logs에 보낸 로그 데이터 보기 섹션을 참조하세요. 기본적으로 이러한 로그는 만료되지 않습니다. 그러나 백업 보존 기간은 수정할 수 있으며, 자세한 내용은 Amazon CloudWatch Logs 사용 설명서의 CloudWatch에서 로그 데이터 보존 기간 변경을 참조하세요.

FAILED

작업이 사용 가능한 모든 시도에서 실패했습니다. FAILED 작업의 작업 상태는 AWS Batch 에서 최소 7일 동안 지속됩니다.

> **Note**
>
> FAILED 작업 로그는 CloudWatch Logs에서 확인할 수 있습니다. 로그 그룹은 /aws/batch/job이며 로그 스트림 이름 형식은 *first200CharsOfJobDefinitionName*/default/*ecs_task_id*입니다. 이 형식은 향후 변경될 수 있습니다.
> 작업이 RUNNING 상태에 도달한 후에는 DescribeJobs API 작업을 사용하여 프로그래밍 방식으로 해당 로그 스트림을 검색할 수 있습니다. 자세한 내용은 Amazon CloudWatch Logs 사용 설명서의 CloudWatch Logs에 보낸 로그 데이터 보기 섹션을 참조하세요. 기본적으로 이러한 로그는 만료되지 않습니다. 그러나 백업 보존 기간은 수정할 수 있으며, 자세한 내용은 Amazon CloudWatch Logs 사용 설명서의 CloudWatch에서 로그 데이터 보존 기간 변경을 참조하세요.

AWS Batch 작업 환경 변수

AWS Batch 는 컨테이너 작업에서 특정 환경 변수를 설정합니다. 이러한 환경 변수는 작업 내 컨테이너에 대한 내부 검사를 제공합니다. 애플리케이션 로직에서 이러한 변수 값을 사용할 수 있습니다. 를 AWS Batch 설정하는 모든 변수는 AWS_BATCH_ 접두사로 시작합니다. 이는 보호되는 환경 변수 접두사입니다. 작업 정의 또는 재정의의 자체 변수에는 이 접두사를 사용할 수 없습니다.

다음 환경 변수는 작업 컨테이너에서 사용할 수 있습니다.

AWS_BATCH_CE_NAME

이 변수는 작업이 배치되는 컴퓨팅 환경의 이름으로 설정됩니다.

AWS_BATCH_JOB_ARRAY_INDEX

이 변수는 하위 배열 작업에서만 지정됩니다. 배열 작업 인덱스는 0에서 시작하며, 각 하위 작업은 고유의 인덱스 번호를 받습니다. 예를 들면 하위가 10개인 배열 작업의 인덱스 값은 0~9입니다. 이 인덱스 값을 사용하여 차별화된 배열 작업 하위에 따라 관리할 수 있습니다. 자세한 내용은 배열 작업 인덱스를 사용한 작업 차별화 관리 단원을 참조하십시오.

AWS_BATCH_JOB_ARRAY_SIZE

이 변수는 상위 배열 작업의 크기로 설정됩니다. 상위 배열 작업의 크기는 이 변수의 하위 배열 작업에 전달됩니다.

`AWS_BATCH_JOB_ATTEMPT`

이 변수는 작업 시도 횟수로 지정됩니다. 첫 번째 시도는 번호 1입니다. 자세한 내용은 작업 자동 재시도 단원을 참조하십시오.

`AWS_BATCH_JOB_ID`

이 변수는 AWS Batch 작업 ID로 설정됩니다.

`AWS_BATCH_JOB_KUBERNETES_NODE_UID`

이 변수는 포드가 실행되는 Kubernetes 클러스터에 있는 노드 객체의 Kubernetes UID로 설정됩니다. 이 변수는 Amazon EKS 리소스에서 실행되는 작업에만 설정됩니다. 자세한 내용은 Kubernetes 설명서의 UDI(UIDs)를 참조하세요.

`AWS_BATCH_JOB_MAIN_NODE_INDEX`

이 변수는 다중 노드 병렬 작업에서만 설정됩니다. 이 변수는 작업 기본 노드의 인덱스 번호로 설정됩니다. 애플리케이션 코드는 `AWS_BATCH_JOB_MAIN_NODE_INDEX`(을)를 개별 노드의 `AWS_BATCH_JOB_NODE_INDEX`(와)과 비교하여 이 노드가 기본 노드인지를 확인할 수 있습니다.

`AWS_BATCH_JOB_MAIN_NODE_PRIVATE_IPV4_ADDRESS`

이 변수는 다중 노드 병렬 작업 하위 노드에서만 설정됩니다. 이 변수는 기본 노드에 존재하지 않지만 작업 기본 노드의 프라이빗 IPv4 주소로 설정됩니다. 하위 노드의 애플리케이션 코드는 이 주소를 사용하여 기본 노드와 통신할 수 있습니다.

`AWS_BATCH_JOB_NODE_INDEX`

이 변수는 다중 노드 병렬 작업에서만 설정됩니다. 이 변수는 노드의 노드 인덱스 번호로 설정됩니다. 노드 인덱스는 0에서 시작하며, 각 노드는 고유의 인덱스 번호를 받습니다. 예를 들면 하위가 10개 있는 다중 노드 병렬 작업의 인덱스 값은 0~9입니다.

`AWS_BATCH_JOB_NUM_NODES`

이 변수는 다중 노드 병렬 작업에서만 설정됩니다. 이 변수는 다중 노드 병렬 작업에 대해 요청한 노드 수로 설정됩니다.

`AWS_BATCH_JQ_NAME`

이 변수는 작업이 제출된 대상 작업 대기열의 이름으로 설정됩니다.

작업 자동 재시도

실패한 작업을 자동으로 작업을 재시도하도록 허용하는 재시도 전략을 작업 및 작업 정의에 적용할 수 있습니다. 작업이 실패하는 상황은 다음과 같습니다.

- 컨테이너 작업에서 0이 아닌 종료 코드 반환
- Amazon EC2 인스턴스 실패 또는 종료
- 내부 AWS 서비스 오류 또는 중단

작업이 작업 대기열로 제출되어 RUNNING 상태가 되면 한 번의 시도로 간주됩니다. 기본적으로 각 작업에는 SUCCEEDED 또는 FAILED 작업 상태가 되기 위해 한 번의 시도가 주어집니다. 그러나 작업 정의와 작업 제출 워크플로를 통해 1회부터 10회까지 재시도 전략을 지정할 수 있습니다. evaluateOnExit를 지정하는 경우 최대 5개의 재시도 전략을 포함할 수 있습니다. evaluateOnExit를 지정했지만 일치하는 재시도 전략이 없는 경우 작업이 다시 시도됩니다. 종료와 일치하지 않는 작업의 경우 어떤 이유로든 종료되는 최종 항목을 추가하세요. 예를 들어, 이 evaluateOnExit 객체에는 RETRY 작업이 있는 두 개의 항목과 EXIT 작업이 있는 최종 항목이 있습니다.

```
"evaluateOnExit": [
    {
        "action": "RETRY",
        "onReason": "AGENT"
    },
    {
        "action": "RETRY",
        "onStatusReason": "Task failed to start"
    },
    {
        "action": "EXIT",
        "onReason": "*"
    }
]
```

실행 시간 시 AWS_BATCH_JOB_ATTEMPT 환경 변수가 컨테이너의 해당 작업 시도 횟수에 설정됩니다. 첫 번째 시도에 1(이)가 지정되고 이후 시도는 2, 3, 4 등과 같이 오름차순으로 횟수가 지정됩니다.

예를 들어, 어떤 이유로든 작업 시도가 실패하고 재시도 구성에 지정된 시도 횟수가 AWS_BATCH_JOB_ATTEMPT 수보다 크다고 가정해 보겠습니다. 그러면 작업이 다시 RUNNABLE 상태로 돌아옵니다. 자세한 내용은 작업 상태 단원을 참조하십시오.

> **Note**
>
> 취소되거나 종료된 작업은 재시도되지 않습니다. 잘못된 작업 정의로 인해 실패한 작업도 재시도되지 않습니다.

자세한 내용은 재시도 전략, 단일 노드 작업 정의 생성, 자습서: 작업 제출 및 중지된 태스크 오류 코드를 참조하세요.

작업 종속성

AWS Batch 작업을 제출할 때 작업이 의존하는 작업 IDs를 지정할 수 있습니다. 이렇게 하면 AWS Batch 스케줄러는 지정된 종속성이 성공적으로 완료된 후에만 작업이 실행되도록 합니다. 성공한 이후 종속된 작업은 `PENDING`에서 `RUNNABLE`(으)로, 그런 다음 `STARTING` 및 `RUNNING`(으)로 전환합니다. 어떠한 작업 속성이 실패한 경우 종속된 작업은 자동적으로 `PENDING`에서 `FAILED`(으)로 전환합니다.

예를 들어 Job A는 최대 20개의 다른 작업에 대한 종속성을 나타낼 수 있으며, 이는 실행 전에 성공해야 합니다. 그런 다음 Job A와 최대 19개의 다른 작업에 대해 종속성을 갖는 추가 작업을 제출할 수 있습니다.

배열 작업의 경우 작업 ID를 입력하지 않고 `SEQUENTIAL` 유형 종속성을 지정할 수 있습니다. 그러면 각 하위 배열 작업은 인덱스 0부터 순차적으로 완료됩니다. 작업 ID를 사용하여 `N_TO_N` 유형 종속성을 지정할 수도 있습니다. 이 방법을 사용하면 이 작업의 각 인덱스 하위는 각 종속성의 해당 인덱스 하위가 완료될 때까지 기다린 후에만 시작할 수 있습니다. 자세한 내용은 배열 작업 단원을 참조하십시오.

종속성이 있는 AWS Batch 작업을 제출하려면 섹션을 참조하세요자습서: 작업 제출.

리소스 인식 일정 예약을 사용하면 작업을 실행하는 데 필요한 소모성 리소스를 기반으로 작업을 예약할 수 있습니다. 작업을 실행하는 데 필요한 소모성 리소스를 지정하면 Batch는 작업을 예약할 때 이러한 리소스 종속성을 고려합니다. 필요한 모든 리소스를 사용할 수 있는 작업만 할당하여 컴퓨팅 리소스의 활용도 저하를 줄일 수 있습니다. 리소스 인식 예약은 FIFO 및 공정 공유 예약 정책 모두에 사용할 수 있으며 EKS, ECS 및 Fargate를 포함하여 Batch에서 지원하는 모든 컴퓨팅 플랫폼에서 사용할 수 있습니다. 배열 작업, 다중 노드 병렬(MNP) 작업 및 일반 배치 작업과 함께 사용할 수 있습니다.

작업 제한 시간

작업이 더 오래 실행되면 AWS Batch (이)가 작업을 종료하도록 작업의 제한 시간을 구성할 수 있습니다. 예를 들어, 15분이면 완료되는 작업이 있을 수 있습니다. 경우에 따라 애플리케이션이 루프에서 정체되고 끝없이 실행되면 30분의 제한 시간을 설정하여 정체된 작업을 종료할 수 있습니다.

> **Important**
>
> 기본적으로 AWS Batch 에는 작업 제한 시간이 없습니다. 작업 제한 시간을 정의하지 않으면 컨테이너가 종료될 때까지 작업이 실행됩니다.

작업 정의 또는 작업 제출 시 `attemptDurationSeconds` 파라미터(최소 60초여야 함)를 지정합니다. 작업 시도의 `startedAt` 타임스탬프 이후이 초가 경과하면가 작업을 AWS Batch 종료합니다. 컴퓨팅 리소스에서 작업 컨테이너는 애플리케이션에 정상적으로 종료할 수 있는 기회를 주기 위해 `SIGTERM` 신호를 수신합니다. 컨테이너가 30초 후에도 여전히 실행 중이면 컨테이너를 강제로 종료하기 위해 `SIGKILL` 신호가 전송됩니다.

제한 시간 종료는 최선의 방식으로 처리됩니다. 작업 시도의 시간이 초과될 때 제한 시간이 정확히 종료될 것으로 기대해서는 안됩니다(몇 초 정도 더 오래 걸릴 수 있음). 애플리케이션에 정확한 제한 시간 실행이 필요한 경우 애플리케이션 내에서 이 로직을 구현해야 합니다. 다수의 작업이 동시에 시간 초과되는 경우, 제한 시간 종료는 작업이 일괄적으로 종료되는 FIFO(선입선출) 대기열로 작동합니다.

> **Note**
>
> AWS Batch 작업에 대한 최대 제한 시간 값은 없습니다.

제한 시간을 초과하여 작업이 종료되면 재시도되지 않습니다. 작업 시도가 자체적으로 실패하면 재시도가 활성화된 경우 재시도할 수 있으며 새로운 시도를 위해 제한 시간 카운트다운이 다시 시작됩니다.

> **Important**
>
> Fargate 리소스에서 실행되는 작업은 14일 이상 실행되지 않습니다. 제한 시간이 14일을 초과하면 Fargate 리소스를 더 이상 사용할 수 없으며 작업이 종료됩니다.

배열 작업의 경우 하위 작업은 상위 작업과 제한 시간 구성이 동일합니다.

제한 시간 구성을 사용하여 AWS Batch 작업을 제출하는 방법에 대한 자세한 내용은 섹션을 참조하세요자습서: 작업 제출.

Amazon EKS 작업

작업은에서 가장 작은 작업 단위입니다 AWS Batch. Amazon EKS의 AWS Batch 작업에는 Kubernetes 포드에 대한 one-to-one 매핑이 있습니다. AWS Batch 작업 정의는 AWS Batch 작업에 대한 템플릿입니다. AWS Batch 작업을 제출할 때 작업 정의를 참조하고, 작업 대기열을 대상으로 지정하고, 작업 이름을 제공합니다. Amazon EKS AWS Batch 에서 작업의 작업 정의에서 eksProperties 파라미터는 AWS Batch Amazon EKS에서가 지원하는 파라미터 세트를 정의합니다. SubmitJob 요청에서 eksPropertiesOverride 파라미터를 사용하면 일부 공통 파라미터를 재정의할 수 있습니다. 이렇게 하면 여러 작업에 대한 작업 정의 템플릿을 사용할 수 있습니다. 작업이 Amazon EKS 클러스터로 디스패치되면는 작업을 podspec ()로 AWS Batch 변환합니다`Kind: Pod`. 는 몇 가지 추가 AWS Batch 파라미터를 `podspec` 사용하여 작업이 올바르게 조정되고 예약되도록 합니다.는 레이블과 테인트를 AWS Batch 결합하여 작업이 AWS Batch 관리형 노드에서만 실행되도록 하고 다른 포드는 해당 노드에서 실행되지 않도록 합니다.

> **Important**
>
> - Amazon EKS 작업 정의에서 `hostNetwork` 파라미터가 명시적으로 설정되지 않은 경우의 포드 네트워킹 모드는 AWS Batch 기본적으로 호스트 모드로 설정됩니다. 보다 구체적으로 `hostNetwork=true` 및 `dnsPolicy=ClusterFirstWithHostNet`(와)과 같은 설정이 적용됩니다.
> - AWS Batch 는 포드가 작업을 완료한 직후 작업 포드를 정리합니다. 포드 애플리케이션 로그를 보려면 클러스터의 로깅 서비스를 구성하세요. 자세한 내용은 CloudWatch Logs를 사용하여 Amazon EKS 작업 AWS Batch 에서 모니터링 단원을 참조하십시오.

주제

- 자습서: 실행 중인 작업을 포드 및 노드에 매핑하기
- 자습서: 실행 중인 포드를 해당 작업에 다시 매핑하기

자습서: 실행 중인 작업을 포드 및 노드에 매핑하기

실행 중인 작업의 `podProperties`에는 현재 작업 시도에 대해 설정된 `podName` 및 `nodeName` 파라미터가 있습니다. 이러한 파라미터를 보려면 DescribeJobs API 작업을 사용하세요.

다음은 예제 출력입니다.

```
$ aws batch describe-jobs --job 2d044787-c663-4ce6-a6fe-f2baf7e51b04
{
 "jobs": [
  {
   "status": "RUNNING",
   "jobArn": "arn:aws:batch:us-east-1:123456789012:job/2d044787-c663-4ce6-a6fe-
f2baf7e51b04",
   "jobDefinition": "arn:aws:batch:us-east-1:123456789012:job-definition/
MyJobOnEks_SleepWithRequestsOnly:1",
   "jobQueue": "arn:aws:batch:us-east-1:123456789012:job-queue/My-Eks-JQ1",
   "jobId": "2d044787-c663-4ce6-a6fe-f2baf7e51b04",
   "eksProperties": {
    "podProperties": {
     "nodeName": "ip-192-168-55-175.ec2.internal",
     "containers": [
      {
       "image": "public.ecr.aws/amazonlinux/amazonlinux:2",
       "resources": {
        "requests": {
         "cpu": "1",
         "memory": "1024Mi"
        }
       }
      }
     ],
     "podName": "aws-batch.b0aca953-ba8f-3791-83e2-ed13af39428c"
    }
   }
  }
 ]
}
```

재시도가 활성화된 작업의 경우 완료된 모든 시도 podName 및 nodeName(은)는 DescribeJobs API 작업의 eksAttempts 목록 파라미터에 있습니다. 현재 실행 중인 시도의 podName 및 nodeName(은)는 podProperties 객체에 있습니다.

자습서: 실행 중인 포드를 해당 작업에 다시 매핑하기

포드에는 포드가 속한 컴퓨팅 환경uuid의 jobId 및를 나타내는 레이블이 있습니다.는 작업의 런타임이 작업 정보를 참조할 수 있도록 환경 변수를 AWS Batch 삽입합니다. 자세한 내용은 AWS Batch 작

업 환경 변수 단원을 참조하십시오. 다음 명령을 실행하여 이 작업을 수행할 수 있습니다. 출력값은 다음과 같습니다.

```
$ kubectl describe pod aws-batch.14638eb9-d218-372d-ba5c-1c9ab9c7f2a1 -n my-aws-batch-namespace
Name:         aws-batch.14638eb9-d218-372d-ba5c-1c9ab9c7f2a1
Namespace:    my-aws-batch-namespace
Priority:     0
Node:         ip-192-168-45-88.ec2.internal/192.168.45.88
Start Time:   Wed, 26 Oct 2022 00:30:48 +0000
Labels:       batch.amazonaws.com/compute-environment-uuid=5c19160b-
d450-31c9-8454-86cf5b30548f
              batch.amazonaws.com/job-id=f980f2cf-6309-4c77-a2b2-d83fbba0e9f0
              batch.amazonaws.com/node-uid=a4be5c1d-9881-4524-b967-587789094647
...
Status:       Running
IP:           192.168.45.88
IPs:
  IP:  192.168.45.88
Containers:
  default:
    Image:         public.ecr.aws/amazonlinux/amazonlinux:2
    ...
    Environment:
      AWS_BATCH_JOB_KUBERNETES_NODE_UID:  a4be5c1d-9881-4524-b967-587789094647
      AWS_BATCH_JOB_ID:                   f980f2cf-6309-4c77-a2b2-d83fbba0e9f0
      AWS_BATCH_JQ_NAME:                  My-Eks-JQ1
      AWS_BATCH_JOB_ATTEMPT:              1
      AWS_BATCH_CE_NAME:                  My-Eks-CE1

...
```

AWS Batch Amazon EKS 작업이 지원하는 기능

다음은 Amazon EKS에서 실행되는 Kubernetes 작업에도 일반적인 AWS Batch 특정 기능입니다.

- 작업 종속성
- 배열 작업
- 작업 제한 시간
- 작업 자동 재시도
- 공정 공유 일정을 사용하여 작업 예약 지원

Kubernetes 및 **`SecretsServiceAccounts`**

AWS Batch 는 Kubernetes `Secrets` 및 참조를 지원합니다`ServiceAccounts`. 서비스 계정에 Amazon EKS IAM 역할을 사용하도록 포드를 구성할 수 있습니다. 자세한 내용은 Amazon EKS 사용 설명서의 포드를 구성하여 Kubernetes 서비스 계정 사용하기를 참조하세요.

관련 문서

- Amazon EKS의 AWS Batch에 대한 메모리 및 vCPU 고려 사항
- GPU 작업 실행
- `RUNNABLE` 상태에서 정체된 작업

다중 노드 병렬 작업

다중 노드 병렬 작업을 사용하면 여러 Amazon EC2 인스턴스에 걸쳐 단일 작업을 실행할 수 있습니다. AWS Batch 다중 노드 병렬 작업(단체 예약이라고도 함)을 사용하면 Amazon EC2 리소스를 직접 시작, 구성 및 관리할 필요 없이 긴밀하게 결합된 대규모 고성능 컴퓨팅 애플리케이션과 분산 GPU 모델 학습을 실행할 수 있습니다. AWS Batch 다중 노드 병렬 작업은 IP 기반 노드 간 통신을 지원하는 모든 프레임워크와 호환됩니다. Apache MXNet, TensorFlow, Caffe2 또는 Message Passing Interface (MPI) 등을 예로 들 수 있습니다.

다중 노드 병렬 작업은 단일 작업으로 제출됩니다. 하지만 작업 정의(또는 작업 제출 노드 재정의)는 작업에 생성할 노드 수와 생성할 노드 그룹을 지정합니다. 각 다중 노드 병렬 작업에는 처음에 시작되는 기본 노드가 포함됩니다. 기본 노드가 가동된 후 하위 노드가 시작됩니다. 메인 노드가 종료되는 경우에만 작업이 완료됩니다. 그러면 모든 하위 노드가 중지됩니다. 자세한 내용은 노드 그룹 단원을 참조하십시오.

다중 노드 병렬 작업 노드는 단일 테넌트입니다. 즉, 각 Amazon EC2 인스턴스에서 작업 컨테이너가 하나만 실행됩니다.

최종 작업 상태(`SUCCEEDED` 또는 `FAILED`)는 기본 노드의 최종 작업 상태에 따라 결정됩니다. 다중 노드 병렬 작업의 상태를 가져오려면 작업을 제출할 때 반환된 작업 ID를 사용하여 작업을 설명할 수 있습니다. 하위 노드에 대한 세부 정보가 필요한 경우 각 하위 노드를 개별적으로 설명해야 합니다. *`#N`* 표기법(0부터 시작)을 사용하여 노드에 주소를 지정할 수 있습니다. 예를 들어 작업의 두 번째 노드에 대한 세부 정보에 액세스하려면 AWS Batch DescribeJobs API 작업을 사용하여 *`aws_batch_job_id`*#1을 설명합니다. 다중 노드 병렬 작업에 대한 `started`, `stoppedAt`, `statusReason` 및 `exit` 정보는 기본 노드에서 채워집니다.

작업 재시도를 지정하는 경우 메인 노드 장애로 인해 또 다른 시도가 발생합니다. 하위 노드 장애로 인해 추가 시도가 발생하지는 않습니다. 다중 노드 병렬 작업의 새로운 시도가 발생할 때마다 연결된 하위 노드의 해당 시도가 업데이트됩니다.

다중 노드 병렬 작업을 실행하려면 AWS Batch애플리케이션 코드에 분산 통신에 필요한 프레임워크와 라이브러리가 포함되어야 합니다.

주제

- 환경 변수
- 노드 그룹
- MNP 작업의 작업 수명 주기
- 를 사용하여 MNP에 대한 컴퓨팅 환경 고려 사항 AWS Batch

환경 변수

런타임 시 각 노드는 모든 AWS Batch 작업이 수신하는 표준 환경 변수로 구성됩니다. 또한 노드는 다중 노드 병렬 작업과 관련된 다음과 같은 환경 변수로 구성됩니다.

`AWS_BATCH_JOB_MAIN_NODE_INDEX`

이 변수는 작업 기본 노드의 인덱스 번호로 설정됩니다. 애플리케이션 코드는 `AWS_BATCH_JOB_MAIN_NODE_INDEX`(을)를 개별 노드의 `AWS_BATCH_JOB_NODE_INDEX`(와)과 비교하여 이 노드가 기본 노드인지를 확인할 수 있습니다.

`AWS_BATCH_JOB_MAIN_NODE_PRIVATE_IPV4_ADDRESS`

이 변수는 다중 노드 병렬 작업 하위 노드에서만 설정됩니다. 이 변수는 메인 노드에는 없습니다. 이 변수는 작업 기본 노드의 프라이빗 IPv4 주소로 설정됩니다. 하위 노드의 애플리케이션 코드는 이 주소를 사용하여 기본 노드와 통신할 수 있습니다.

`AWS_BATCH_JOB_NODE_INDEX`

이 변수는 노드의 노드 인덱스 번호로 설정됩니다. 노드 인덱스는 0에서 시작하며, 각 노드는 고유의 인덱스 번호를 받습니다. 예를 들면 하위가 10개 있는 다중 노드 병렬 작업의 인덱스 값은 0~9입니다.

`AWS_BATCH_JOB_NUM_NODES`

이 변수는 다중 노드 병렬 작업에 대해 요청한 노드 수로 설정됩니다.

노드 그룹

노드 그룹은 모두 동일한 컨테이너 속성을 공유하는 작업 노드의 동일한 그룹입니다. AWS Batch 를 사용하여 각 작업에 대해 최대 5개의 개별 노드 그룹을 지정할 수 있습니다.

각 그룹에는 고유의 컨테이너 이미지, 명령, 환경 변수 등이 있을 수 있습니다. 예를 들어, 메인 노드용 단일 `c5.xlarge` 인스턴스와 다섯 개의 `c5.xlarge` 인스턴스 하위 노드가 필요한 작업을 제출할 수 있습니다. 이러한 개별 노드 그룹 각각은 각 작업에 대해 실행할 다른 컨테이너 이미지 또는 명령을 지정할 수 있습니다.

또는 작업의 모든 노드가 단일 노드 그룹을 사용할 수도 있습니다. 또한 애플리케이션 코드는 메인 노드 및 하위 노드와 같은 노드 역할을 구분할 수 있습니다. `AWS_BATCH_JOB_MAIN_NODE_INDEX` 환경 변수를 `AWS_BATCH_JOB_NODE_INDEX`에 대한 자체 값과 비교하여 이를 수행합니다. 단일 작업에 최대 1,000개의 노드가 있을 수 있습니다. 이는 Amazon ECS 클러스터 내 인스턴스에 대한 기본 제한입니다. 이 한도를 늘리도록 요청할 수 있습니다.

> **Note**
>
> 현재 다중 노드 병렬 작업의 모든 노드 그룹은 동일한 인스턴스 유형을 사용해야 합니다.

MNP 작업의 작업 수명 주기

다중 노드 병렬 작업을 제출하면 작업이 `SUBMITTED` 상태가 됩니다. 그런 다음 작업은 모든 작업 종속성이 완료될 때까지 대기합니다. 또한 작업이 `RUNNABLE` 상태로 전환합니다. 마지막으로, 작업을 실행하는 데 필요한 인스턴스 용량을 AWS Batch 프로비저닝하고 이러한 인스턴스를 시작합니다.

각 다중 노드 병렬 작업에는 기본 노드가 포함되어 있습니다. 기본 노드는가 제출된 다중 노드 작업의 결과를 확인하기 위해 AWS Batch 모니터링하는 단일 하위 작업입니다. 기본 노드가 처음에 시작되고 `STARTING` 상태로 이동합니다. `attemptDurationSeconds` 파라미터에 지정된 제한 시간 값은 노드가 아닌 전체 작업에 적용됩니다.

기본 노드가 `RUNNING` 상태에 도달하면(노드의 컨테이너가 실행된 후) 하위 노드가 시작되고 하위 노드도 `STARTING` 상태로 이동합니다. 하위 노드는 임의의 순서로 나타납니다. 하위 노드가 시작되는 시간이나 순서는 보장할 수 없습니다. 작업의 모든 노드가 `RUNNING` 상태인지 확인하려면(노드 컨테이너가 실행된 후) 애플리케이션 코드가 AWS Batch API를 쿼리하여 기본 노드 및 하위 노드 정보를 가져올 수 있습니다. 또는 애플리케이션 코드가 모든 노드가 온라인 상태가 될 때까지 기다린 후 분산 처리 태스크를 시작할 수도 있습니다. 기본 노드의 프라이빗 IP 주소를 각 하위 노드에서

AWS_BATCH_JOB_MAIN_NODE_PRIVATE_IPV4_ADDRESS 환경 변수로 사용할 수 있습니다. 애플리케이션 코드는 이 정보를 사용하여 각 태스크 간에 데이터를 조정하고 통신할 수 있습니다.

개별 노드가 종료되면 종료 코드에 따라 SUCCEEDED 또는 FAILED(으)로 이동합니다. 기본 노드가 종료되면 작업이 완료된 것으로 간주되고 모든 하위 노드가 중지됩니다. 하위 노드가 죽으면 작업의 다른 노드에 대해 어떤 작업도 수행하지 AWS Batch 않습니다. 감소된 수의 노드로 작업을 계속하지 않으려는 경우 이러한 요인을 애플리케이션 코드에 반영해야 합니다. 이렇게 하면 작업이 종료되거나 취소됩니다.

를 사용하여 MNP에 대한 컴퓨팅 환경 고려 사항 AWS Batch

AWS Batch(을)를 사용하여 다중 노드 병렬 작업을 실행하도록 컴퓨팅 환경을 구성할 때 몇 가지 고려할 점이 있습니다.

- 다중 노드 병렬 작업은 UNMANAGED 컴퓨팅 환경에서 지원되지 않습니다.
- 다중 노드 병렬 작업을 컴퓨팅 환경에 제출하려는 경우 단일 가용 영역에서 클러스터 배치 그룹을 생성하고 이 그룹을 컴퓨팅 리소스와 연결합니다. 이렇게 하면 높은 네트워크 흐름 잠재력으로 가깝게 인접한 인스턴스를 논리적으로 그룹화할 때 다중 노드 병렬 작업이 그대로 유지됩니다. 자세한 내용을 알아보려면 Amazon EC2 사용 설명서의 배치 그룹을 참조하세요.
- 스팟 인스턴스를 사용하는 컴퓨팅 환경에서는 다중 노드 병렬 작업이 지원되지 않습니다.
- AWS Batch 다중 노드 병렬 작업은 다중 노드 병렬 작업 컨테이너에 Amazon EC2 인스턴스와 동일한 네트워킹 속성을 제공하는 Amazon ECS awsvpc 네트워크 모드를 사용합니다. 각 다중 노드 병렬 작업 컨테이너는 고유의 탄력적 네트워크 인터페이스, 기본 프라이빗 IP 주소, 내부 DNS 호스트 이름을 가져옵니다. 네트워크 인터페이스는 호스트 컴퓨팅 리소스와 동일한 VPC 서브넷에서 생성됩니다.
- 컴퓨팅 환경에는 최대 5개의 보안 그룹을 연결할 수 있습니다. 생성되어 MNP 태스크에 연결되는 탄력적 네트워크 인터페이스는 컴퓨팅 환경에 지정된 보안 그룹을 사용합니다. 보안 그룹을 지정하지 않는 경우, VPC의 기본 보안 그룹이 사용됩니다. 일반 AWS Batch 작업과 달리 다중 노드 병렬 작업은 시작 템플릿에 지정된 보안 그룹을 사용하지 않습니다.
- awsvpc 네트워크 모드는 다중 노드 병렬 작업에 대한 탄력적 네트워크 인터페이스에 퍼블릭 IP 주소를 제공하지 않습니다. 인터넷에 액세스하려면 NAT 게이트웨이를 사용하도록 구성된 프라이빗 서브넷에서 컴퓨팅 리소스를 시작해야 합니다. 자세한 정보는 Amazon VPC 사용 설명서의 NAT 게이트웨이 섹션을 참조하세요. 노드 간 통신은 노드의 프라이빗 IP 주소 또는 DNS 호스트 이름을 사용해야 합니다. 퍼블릭 서브넷 내의 컴퓨팅 리소스에서 실행하는 다중 노드 병렬 작업은 아웃바운드 네트워크 액세스를 수행할 수 없습니다. 프라이빗 서브넷과 NAT 게이트웨이를 사용하여 VPC를 생성하려면 Virtual Private Cloud 생성 섹션을 참조하세요.

- 생성하여 컴퓨팅 리소스에 연결하는 탄력적 네트워크 인터페이스는 계정에서 수동으로 분리하거나 수정할 수 없습니다. 이러한 제한 사항은 실행 중인 작업과 연결된 탄력적 네트워크 인터페이스의 우발적인 삭제를 방지하기 위한 것입니다. 태스크에 대한 탄력적 네트워크 인터페이스를 해제하려면 작업을 종료합니다.
- 컴퓨팅 환경에는 다중 노드 병렬 작업을 지원하기에 충분한 최대 vCPU가 있어야 합니다.
- Amazon EC2 인스턴스 할당량에는 작업을 실행하는 데 필요한 인스턴스 수가 포함됩니다. 예를 들어, 작업에 30개의 인스턴스가 필요하지만 계정이 20개의 인스턴스만 실행할 수 있는 경우 작업은 지원합니다. 그러면 작업이 그대로 RUNNABLE 상태로 유지됩니다.
- 다중 노드 병렬 작업의 노드 그룹에 인스턴스 유형을 지정하는 경우 컴퓨팅 환경에서 해당 인스턴스 유형을 시작할 수 있어야 합니다.

Amazon EKS의 다중 노드 병렬 작업

Amazon Elastic Kubernetes Service AWS Batch 에서를 사용하여 관리형 Kubernetes 클러스터에서 다중 노드 병렬(MNP) 작업(그룹 스케줄링이라고도 함)을 실행할 수 있습니다. 이 옵션은 일반적으로 단일 Amazon Elastic Compute Cloud 인스턴스에서 실행할 수 없는 긴밀하게 결합된 대규모 고성능 작업에 사용됩니다. 자세한 내용은 다중 노드 병렬 작업 단원을 참조하십시오.

이 기능을 사용하여 Amazon EKS 관리형 Kubernetes 전용 고성능 컴퓨팅 애플리케이션, 대규모 언어 모델 훈련 및 기타 인공 지능(AI)/기계 학습(ML) 작업을 실행할 수 있습니다.

주제

- MNP 작업 실행
- Amazon EKS MNP 작업 정의 생성
- Amazon EKS MNP 작업 제출
- Amazon EKS MNP 작업 정의 재정의

MNP 작업 실행

AWS Batch 는 Amazon EC2를 사용하여 Amazon Elastic Container Service 및 Amazon EKS에서 MNP 작업을 지원합니다. 다음은 이 기능의 인스턴스 및 컨테이너 파라미터에 대한 자세한 내용입니다.

Amazon EKS의 MNP에 대한 인스턴스 할당량

- 단일 MNP 작업에 최대 1,000개의 인스턴스를 사용할 수 있습니다.

- 단일 Amazon EKS 클러스터에 최대 5,000개의 인스턴스를 조인할 수 있습니다.
- 최대 5개의 컴퓨팅 환경을 클러스터링하여 작업 대기열에 연결할 수 있습니다.

예를 들어, 한 작업 대기열에서 클러스터링된 컴퓨팅 환경을 최대 5개까지 확장하고 각 컴퓨팅 환경에서 인스턴스를 1,000개까지 확장할 수 있습니다.

인스턴스 파라미터 외에도 두 서비스 모두 MNP 작업에 Fargate를 사용할 수 없다는 점에 유의해야 합니다.

각 MNP 작업에는 인스턴스 유형을 하나만 사용할 수 있습니다. 컴퓨팅 환경을 업데이트하거나 새 컴퓨팅 환경을 정의할 때 인스턴스 유형을 변경할 수 있습니다. 작업 정의를 생성할 때 인스턴스 유형을 지정하고 vCPU 및 메모리 요구 사항을 제공할 수도 있습니다.

Amazon EKS의 MNP에 대한 컨테이너 할당량

- 다중 노드 병렬(MNP) 작업은 노드당 하나의 포드를 지원합니다.
- 각 포드는 최대 10개의 컨테이너(또는 10개의 Init 컨테이너)를 지원합니다. 자세한 내용은 Kubernetes 설명서의 Init 컨테이너를 참조하세요.
- 각 MNP 작업에서 최대 5개의 노드 범위를 지원합니다.
- 각 노드 범위에서 최대 10개의 고유한 컨테이너 이미지를 지원합니다.

예를 들어, 5개의 노드 범위와 총 50개의 고유 이미지가 포함된 단일 MNP 작업에서 최대 10,000개의 컨테이너를 실행할 수 있습니다.

프라이빗 Amazon VPC 및 Amazon EKS 클러스터에서 MNP 작업 실행

MNP 작업은 퍼블릭 인터넷 지원 여부와 관계없이 모든 Amazon EKS 클러스터에서 실행할 수 있습니다. 프라이빗 네트워크 액세스만 있는 Amazon EKS 클러스터를 사용하는 경우가 Amazon EKS 컨트롤 플레인 및 관리형 Kubernetes API 서버에 액세스할 AWS Batch 수 있는지 확인합니다. Amazon Virtual Private Cloud 엔드포인트를 통해 필요한 액세스 권한을 부여할 수 있습니다. 자세한 내용은 엔드포인트 서비스 구성을 참조하세요.

프라이빗 VPC는 인터넷에 액세스할 수 없으므로 Amazon EKS 클러스터 포드는 퍼블릭 소스에서 이미지를 다운로드할 수 없습니다. Amazon EKS 클러스터는 Amazon VPC 내에 있는 컨테이너 레지스트리에서 이미지를 가져와야 합니다. Amazon VPC에서 Amazon Elastic Container Registry(Amazon ECR)를 생성하고 노드에서 액세스할 수 있도록 여기에 컨테이너 이미지를 복사할 수 있습니다.

Amazon ECR을 사용하여 풀스루 캐시 규칙을 생성할 수도 있습니다. 외부 퍼블릭 레지스트리에 대한 풀스루 캐시 규칙이 생성되면 Amazon ECR 프라이빗 레지스트리 URI를 사용하여 해당 외부 퍼블릭 레지스트리에서 이미지를 가져오면 됩니다. 그러면 Amazon ECR이 리포지토리를 생성하고 해당 이미지를 캐시합니다. Amazon ECR 프라이빗 레지스트리 URI를 사용하여 캐시된 이미지를 가져오면 Amazon ECR은 원격 레지스트리를 점검하여 이미지의 새 버전이 있는지 확인하며, 최대 24시간마다 한 번씩 프라이빗 레지스트리를 업데이트합니다. 자세한 내용은 Amazon ECR에서 풀스루 캐시 규칙 생성을 참조하세요.

오류 알림

MNP 작업이 차단된 경우 AWS Management Console 및 Amazon EventBridge를 통해 알림을 받을 수 있습니다. 예를 들어, MNP 작업이 대기열 상단에서 멈춘 경우 작업 대기열 차단을 해결하기 위한 즉각적인 조치를 취할 수 있도록 문제에 대한 알림과 함께 문제 원인에 대한 정보를 받을 수 있습니다. 선택적으로, 작업 대기열 템플릿에서 정의할 수 있는 별도의 시간 내에 조치가 취해지지 않으면 MNP 작업을 자동으로 종료할 수 있습니다. 자세한 내용은 작업 대기열 차단 이벤트 섹션을 참조하세요.

Amazon EKS MNP 작업 정의 생성

Amazon EKS에서 MNP 작업을 정의하고 실행할 수 있도록 `RegisterJobDefinition` 및 `SubmitJob` API 작업 내에 새로운 파라미터가 있습니다.

- `nodeProperties` 섹션 아래에서 `eksProperties`를 사용하여 MNP 작업 정의를 정의합니다.
- `nodePropertyOverrides` 섹션 아래에서 `eksPropertiesOverride`를 사용하여 MNP 작업을 제출할 때 작업 정의에 정의된 파라미터를 재정의합니다.

이러한 작업은 API 작업 및 AWS Management Console을 통해 정의할 수 있습니다.

참조: Amazon EKS MNP 작업 정의 요청 페이로드 등록

다음 예제에서는 Amazon EKS MNP 작업 정의를 두 노드에 등록하는 방법을 보여줍니다.

```
{
  "jobDefinitionName": "MyEksMnpJobDefinition",
  "type": "multinode",
  "nodeProperties": {
    "numNodes": 2,
    "mainNode": 0,
    "nodeRangeProperties": [
      {
        "targetNodes" : "0:",
```

```
        "eksProperties": {
          "podProperties": {
            "containers": [
              {
                "name": "test-eks-container-1",
                "image": "public.ecr.aws/amazonlinux/amazonlinux:2",
                "command": [
                  "sleep",
                  "60"
                ],
                "resources": {
                  "limits": {
                    "cpu": "1",
                    "memory": "1024Mi"
                  }
                },
                "securityContext":{
                  "runAsUser":1000,
                  "runAsGroup":3000,
                  "privileged":true,
                  "readOnlyRootFilesystem":true,
                  "runAsNonRoot":true
               }
              }
            ],
            "initContainers": [
               {
                  "name":"init-ekscontainer",
                  "image": "public.ecr.aws/amazonlinux/amazonlinux:2",
                  "command": [
                     "echo",
                     "helloWorld"
                   ],
                   "resources": {
                     "limits": {
                       "cpu": "1",
                       "memory": "1024Mi"
                     }
                  }
               }
            ],
            "metadata": {
               "labels": {
                  "environment" : "test"
```

```
                }
            }
          }
        }
      }
    ]
  }
}
```

를 사용하여 작업 정의를 등록하려면 정의를 MyEksMnpJobDefinition.json이라는 로컬 파일에 AWS CLI복사하고 다음 명령을 실행합니다.

```
aws batch register-job-definition --cli-input-json file://MyEksMnpJobDefinition.json
```

다음과 같은 JSON 응답이 반환됩니다.

```
{
    "jobDefinitionName": "MyEksMnpJobDefinition",
    "jobDefinitionArn": "arn:aws:batch:us-east-1:0123456789:job-definition/
MyEksMnpJobDefinition:1",
    "revision": 1
}
```

Amazon EKS MNP 작업 제출

등록된 작업 정의를 사용하여 작업을 제출하려면 다음 명령을 입력합니다.
<EKS_JOB_QUEUE_NAME> 값을 Amazon EKS 컴퓨팅 환경과 연결된 기존 작업 대기열의 이름 또는 ARN으로 바꿉니다.

```
aws batch submit-job --job-queue <EKS_JOB_QUEUE_NAME> \
    --job-definition MyEksMnpJobDefinition \
    --job-name myFirstEksMnpJob
```

다음과 같은 JSON 응답이 반환됩니다.

```
{
    "jobArn": "arn:aws:batch:region:account:job/9b979cce-9da0-446d-90e2-ffa16d52af68",
    "jobName": "myFirstEksMnpJob",
    "jobId": "<JOB_ID>"
```

```
}
```

반환된 jobId와 다음 명령을 사용하여 작업 상태를 확인할 수 있습니다.

```
aws batch describe-jobs --jobs <JOB_ID>
```

Amazon EKS MNP 작업 정의 재정의

선택적으로, MNP 작업 크기나 하위 작업 세부 정보를 변경하는 것처럼 작업 정의 세부 정보를 재정의할 수 있습니다. 다음은 5개 노드 MNP 작업과 `test-eks-container-1` 컨테이너의 명령 변경 사항을 제출하기 위한 예제 JSON 요청 페이로드입니다.

```
{
  "numNodes": 5,
  "nodePropertyOverrides": [
    {
      "targetNodes": "0:",
      "eksPropertiesOverride": {
        "podProperties": {
          "containers": [
            {
              "name": "test-eks-container-1",
              "command": [
                "sleep",
                "150"
              ]
            }
          ]
        }
      }
    }
  ]
}
```

이러한 재정의가 있는 작업을 제출하려면 예제를 로컬 파일인 eks-mnp-job-nodeoverride.json AWS CLI 에 저장하고를 사용하여 재정의가 있는 작업을 제출합니다.

배열 작업

배열 작업은 작업 정의, vCPU 및 메모리와 같은 공통 파라미터를 공유하는 작업입니다. 여러 호스트에 걸쳐 배포될 수 있으며 동시에 실행할 수 있는 관련 개별 기본 작업의 모음으로 실행됩니다. 배열 작업

은 Monte Carlo 시뮬레이션, 파라미터 스윕 또는 대규모 렌더링 작업과 같이 고도의 병렬 작업을 처리하는 가장 효율적인 방법입니다.

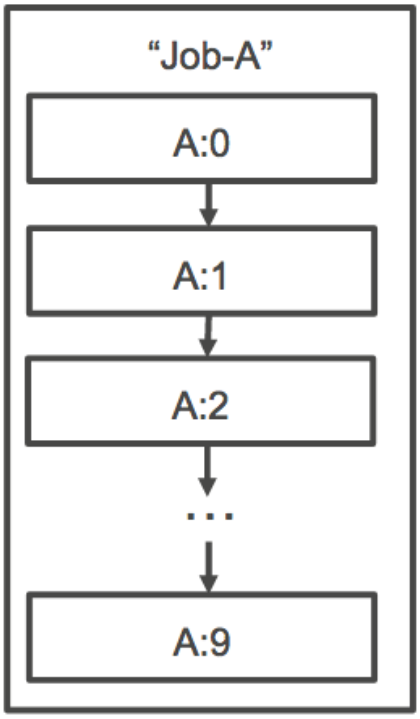
AWS Batch 배열 작업은 일반 작업과 마찬가지로 제출됩니다. 하지만 배열 크기(2~10,000)를 지정하여 얼마나 많은 하위 작업이 배열에서 실행되어야 하는지를 정의합니다. 배열 크기가 1,000인 작업을 제출하는 경우 단일 작업은 1,000개의 하위 작업을 실행 및 생성합니다. 배열 작업은 모든 하위 작업을 관리하는 참조 또는 포인터입니다. 이를 통해 단일 쿼리를 포함한 대규모 워크로드를 제출할 수 있습니다. `attemptDurationSeconds` 파라미터에 지정된 제한 시간은 각 하위 작업에 적용됩니다. 상위 배열 작업에는 제한 시간이 없습니다.

배열 작업을 제출하면 상위 배열 작업은 일반 AWS Batch 작업 ID를 가져옵니다. 각 하위 작업의 기본 ID는 동일합니다. 각 하위 작업은 동일한 기본 ID를 갖지만 하위 작업에 대한 배열 인덱스는 상위 ID의 끝에 추가됩니다(예: 배열의 첫 번째 하위 작업의 경우 *`example_job_ID`*`:0`).

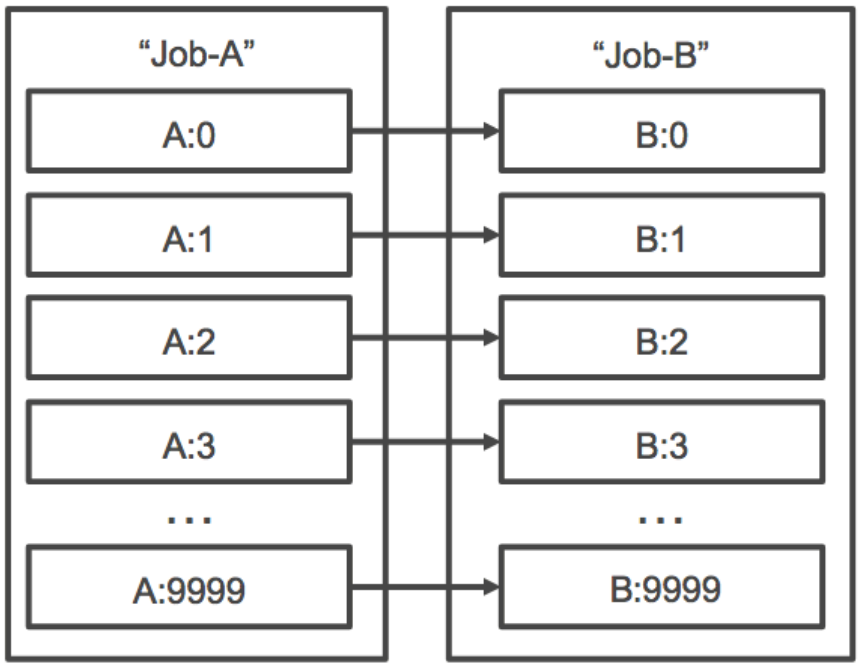
상위 배역 작업은 `SUBMITTED`, `PENDING`, `FAILED` 또는 `SUCCEEDED` 상태를 입력할 수 있습니다. 하위 작업이 `RUNNABLE`(으)로 업데이트되면 상위 배열 작업이 `PENDING`(으)로 업데이트됩니다. 이러한 종속성에 대한 자세한 내용은 작업 종속성 섹션을 참조하세요.

실행 시간 시 `AWS_BATCH_JOB_ARRAY_INDEX` 환경 변수가 컨테이너의 해당 배열 인덱스 번호에 설정됩니다. 첫 번째 배열 작업 인덱스에는 `0`(이)가 지정되고 이후 시도는 1, 2, 3 등과 같이 오름차순으로 횟수가 지정됩니다. 이 인덱스 값을 사용하여 차별화된 배열 작업 하위에 따라 관리할 수 있습니다. 자세한 내용은 배열 작업 인덱스를 사용한 작업 차별화 관리 단원을 참조하십시오.

배열 작업 종속성의 경우 `SEQUENTIAL` 또는 `N_TO_N`(와)과 같이 종속성의 유형을 지정할 수 있습니다. 작업 ID를 입력하지 않고 `SEQUENTIAL` 유형 종속성을 지정할 수 있습니다. 그러면 각 하위 배열 작업은 인덱스 0부터 순차적으로 완료됩니다. 예를 들어 배열 크기가 100인 배열 작업을 제출하고, 종속성 유형을 `SEQUENTIAL`(으)로 지정한 경우 100개의 하위 작업이 순차적으로 생성됩니다. 그리고 첫 번째 하위 작업이 성공해야 다음 하위 작업이 시작됩니다. 아래 그림은 배열 크기가 10인 배열 작업, Job A를 보여줍니다. Job A의 하위 인덱스에 있는 각 작업은 이전 하위 작업에 종속적입니다. Job A:0이 완료된 이후에 Job A:1이 시작할 수 있습니다.



배열 작업의 작업 ID로 N_TO_N 유형의 종속성을 지정할 수도 있습니다. 이 방법을 사용하면 이 작업의 각 인덱스 하위는 각 종속성의 해당 인덱스 하위가 완료될 때까지 기다린 후에만 시작할 수 있습니다. 아래 그림은 배열 크기가 각각 10,000인 2개의 배열 작업, Job A와 Job B를 보여줍니다. Job B의 하위 인덱스에 있는 각 작업은 Job A의 해당 인덱스에 종속적입니다. Job B:1은 Job A:1이 완료될 때까지 시작할 수 없습니다.



상위 배열 작업을 취소 또는 종료한 경우 모든 하위 작업이 함께 취소 또는 종료됩니다. 다른 하위 작업에 영향을 미치지 않고 개별 하위 작업을 취소 또는 종료할 수 있습니다(FAILED 상태로 변경됨). 하지만 하위 배열 작업이 실패(자체적으로 실패하거나 수동으로 작업을 취소 또는 종료하여 실패)한 경우 상위 작업 또한 실패합니다. 이 시나리오에서는 모든 하위 작업이 완료될 때 상위 작업이 FAILED로 전환됩니다.

배열 작업 검색 및 필터링에 대한 자세한 내용은 섹션을 참조하세요에서 작업 대기열의 작업 검색.

주제

- 배열 작업 워크플로의 예
- 배열 작업 인덱스를 사용한 작업 차별화 관리

배열 작업 워크플로의 예

AWS Batch 고객의 일반적인 워크플로는 사전 조건 설정 작업을 실행하고 많은 수의 입력 작업에 대해 일련의 명령을 실행한 다음 결과를 집계하고 요약 데이터를 Amazon S3, DynamoDB, Amazon Redshift 또는 Aurora에 쓰는 작업으로 끝나는 것입니다.

예제:

- JobA: Amazon S3 버킷인 BucketA에 있는 객체의 빠른 나열과 메타데이터 검증을 수행하는 표준 비 배열 작업입니다. SubmitJob JSON 구문은 다음과 같습니다.

```
{
    "jobName": "JobA",
    "jobQueue": "ProdQueue",
    "jobDefinition": "JobA-list-and-validate:1"
}
```

- JobB: JobA에 종속적인 10,000개의 작업 부수가 포함된 배열 작업. BucketA의 각 객체에 대한 CPU 집약적 명령을 실행하고 BucketB에 결과를 업로드합니다. SubmitJob JSON 구문은 다음과 같습니다.

```
{
    "jobName": "JobB",
    "jobQueue": "ProdQueue",
    "jobDefinition": "JobB-CPU-Intensive-Processing:1",
    "containerOverrides": {
        "resourceRequirements": [
            {
```

```
                "type": "MEMORY",
                "value": "4096"
            },
            {
                "type": "VCPU",
                "value": "32"
            }
        ]
   }
    "arrayProperties": {
        "size": 10000
    },
    "dependsOn": [
        {
            "jobId": "JobA_job_ID"
  }
    ]
}
```

- JobC: N_TO_N 종속성 모델을 사용하여 JobB에 종속되는 또 다른 10,000 부 배열 작업입니다. 이 작업은 BucketB의 각 항목에 대한 메모리 집약적 명령을 실행하고, 메타데이터를 DynamoDB에 쓴 다음, 결과 출력을 BucketC에 업로드합니다. SubmitJob JSON 구문은 다음과 같습니다.

```
{
    "jobName": "JobC",
    "jobQueue": "ProdQueue",
    "jobDefinition": "JobC-Memory-Intensive-Processing:1",
    "containerOverrides": {
        "resourceRequirements": [
            {
                "type": "MEMORY",
                "value": "32768"
            },
            {
                "type": "VCPU",
                "value": "1"
            }
        ]
   }
    "arrayProperties": {
        "size": 10000
    },
    "dependsOn": [
```

```
        {
            "jobId": "JobB_job_ID",
            "type": "N_TO_N"
        }
    ]
}
```

- JobD: 10개의 검증 단계를 수행하는 배열 작업입니다. 각 단계는 DynamoDB에 쿼리해야 하고, 위의 Amazon S3 버킷 중 어떠한 버킷과도 상호 작용할 수 있습니다. JobD의 각 단계는 동일한 명령을 실행합니다. 그러나 동작은 작업의 컨테이너 내 AWS_BATCH_JOB_ARRAY_INDEX 환경 변수의 값에 따라 다릅니다. 이러한 검증 단계는 순차적으로 작동합니다(예: JobD:0 이후 JobD:1). SubmitJob JSON 구문은 다음과 같습니다.

```
{
    "jobName": "JobD",
    "jobQueue": "ProdQueue",
    "jobDefinition": "JobD-Sequential-Validation:1",
    "containerOverrides": {
        "resourceRequirements": [
            {
                "type": "MEMORY",
                "value": "32768"
            },
            {
                "type": "VCPU",
                "value": "1"
            }
        ]
   }
    "arrayProperties": {
        "size": 10
    },
    "dependsOn": [
        {
            "jobId": "JobC_job_ID"
        },
        {
            "type": "SEQUENTIAL"
        },

    ]
}
```

- JobE: 몇 가지 단순한 정리 작업을 수행하고 파이프라인이 완료되었다는 메시지와 출력 URL 링크가 포함된 Amazon SNS 알림을 전송하는 최종 비 배열 작업입니다. SubmitJob JSON 구문은 다음과 같습니다.

```
{
    "jobName": "JobE",
    "jobQueue": "ProdQueue",
    "jobDefinition": "JobE-Cleanup-and-Notification:1",
    "parameters": {
        "SourceBucket": "s3://amzn-s3-demo-source-bucket",
        "Recipient": "pipeline-notifications@mycompany.com"
    },
    "dependsOn": [
        {
            "jobId": "JobD_job_ID"
        }
    ]
}
```

배열 작업 인덱스를 사용한 작업 차별화 관리

이 자습서에서는 AWS_BATCH_JOB_ARRAY_INDEX 환경 변수를 사용하여 하위 작업을 구분하는 방법을 설명합니다. 각 하위 작업이 이 변수에 할당됩니다. 이 예제에서는 하위 작업의 인덱스 번호를 사용하여 파일의 특정 줄을 읽습니다. 그런 다음 해당 줄 번호와 관련된 파라미터를 작업 컨테이너 내의 명령으로 대체합니다. 그 결과 동일한 Docker 이미지 및 명령 인수를 실행하는 여러 AWS Batch 작업이 있을 수 있습니다. 하지만 배열 작업 인덱스가 한정자로 사용되므로 결과가 달라집니다.

이 자습서에서는 각 줄에 무지개색 텍스트 파일을 만듭니다. 그런 다음 Docker 컨테이너용 진입점 스크립트를 만들어 색상 파일의 줄 번호에 사용 가능한 값(인덱스는 0부터 시작하지만 줄 번호는 1부터 시작)으로 인덱스를 변환합니다. 인덱스는 0에서 시작하지만 줄 번호는 1부터 시작합니다. 색상과 인덱스 파일을 컨테이너 이미지에 복사하고 이미지의 ENTRYPOINT(을)를 진입점 스크립트에 지정하는 Dockerfile을 만듭니다. Dockerfile 및 리소스를 도커 이미지에 만들어 Amazon ECR로 푸시합니다. 그런 다음 새 컨테이너 이미지를 사용하는 작업 정의를 등록하고, 해당 작업 정의와 함께 AWS Batch 배열 작업을 제출하고, 결과를 봅니다.

주제

- 사전 조건
- 컨테이너 이미지 빌드

- 이미지를 Amazon ECR로 푸시
- 작업 정의 생성 및 등록
- AWS Batch 배열 작업 제출
- 배열 작업 로그 보기

사전 조건

이 자습서 워크플로의 사전 요구 사항은 다음과 같습니다.

- AWS Batch 컴퓨팅 환경. 자세한 내용은 컴퓨팅 환경 생성 단원을 참조하십시오.
- AWS Batch 작업 대기열 및 관련 컴퓨팅 환경. 자세한 내용은 작업 대기열 생성 단원을 참조하십시오.
- 로컬 시스템에 AWS CLI 설치된 . 자세한 내용은 AWS Command Line Interface 사용 설명서의 AWS Command Line Interface설치를 참조하세요.
- 로컬 시스템에 설치한 도커입니다. 자세한 내용은 Docker 설명서의 Docker CE 소개를 참조하세요.

컨테이너 이미지 빌드

명령 파라미터의 작업 정의에 `AWS_BATCH_JOB_ARRAY_INDEX`를 사용할 수 있습니다. 하지만 진입점 스크립트에서 변수를 사용하는 컨테이너 이미지를 대신 생성하는 것이 좋습니다. 이 섹션에서는 이러한 컨테이너 이미지를 만드는 방법을 설명합니다.

Docker 컨테이너 이미지를 빌드하려면

1. 도커 이미지 작업 영역으로 사용할 새 디렉터리를 만들고 여기로 이동합니다.
2. 작업 영역 디렉터리에 이름이 `colors.txt`인 파일을 만들고 그 아래에 내용을 붙여 넣습니다.

   ```
   red
   orange
   yellow
   green
   blue
   indigo
   violet
   ```

3. 작업 영역 디렉터리에 이름이 `print-color.sh`인 파일을 만들고 그 아래에 내용을 붙여 넣습니다.

> **Note**
>
> 배열 인덱스는 0에서 시작하지만 줄 번호는 1에서 시작하므로 LINE 변수가 AWS_BATCH_JOB_ARRAY_INDEX+1로 지정됩니다. COLOR 변수가 줄 번호와 연결된 colors.txt의 색으로 지정됩니다.

```
#!/bin/sh
LINE=$((AWS_BATCH_JOB_ARRAY_INDEX + 1))
COLOR=$(sed -n ${LINE}p /tmp/colors.txt)
echo My favorite color of the rainbow is $COLOR.
```

4. 작업 영역 디렉터리에 이름이 Dockerfile인 파일을 만들고 다음 콘텐츠를 붙여 넣습니다. 이 Dockerfile이 이전 파일을 컨테이너에 복사하고 컨테이너가 시작하면 진입점 스크립트를 실행하도록 지정합니다.

```
FROM busybox
COPY print-color.sh /tmp/print-color.sh
COPY colors.txt /tmp/colors.txt
RUN chmod +x /tmp/print-color.sh
ENTRYPOINT /tmp/print-color.sh
```

5. 도커 이미지를 빌드합니다.

```
$ docker build -t print-color .
```

6. 다음 스크립트로 컨테이너를 테스트합니다. 이 스크립트는 로컬에서 AWS_BATCH_JOB_ARRAY_INDEX 변수를 0으로 지정한 다음 증분하여 하위가 7개인 배열 작업을 시뮬레이션합니다.

```
$ AWS_BATCH_JOB_ARRAY_INDEX=0
while [ $AWS_BATCH_JOB_ARRAY_INDEX -le 6 ]
do
    docker run -e AWS_BATCH_JOB_ARRAY_INDEX=$AWS_BATCH_JOB_ARRAY_INDEX print-color
    AWS_BATCH_JOB_ARRAY_INDEX=$((AWS_BATCH_JOB_ARRAY_INDEX + 1))
done
```

출력 값은 다음과 같습니다.

```
My favorite color of the rainbow is red.
My favorite color of the rainbow is orange.
My favorite color of the rainbow is yellow.
My favorite color of the rainbow is green.
My favorite color of the rainbow is blue.
My favorite color of the rainbow is indigo.
My favorite color of the rainbow is violet.
```

이미지를 Amazon ECR로 푸시

이제 Docker 컨테이너를 구축하고 테스트했으므로 이미지 리포지토리에 푸시해야 합니다. 이 예제는 Amazon ECR을 사용하지만, DockerHub 같은 다른 레지스트리를 선택해도 됩니다.

1. 컨테이너 이미지를 저장할 Amazon ECR 이미지 리포지토리를 생성합니다. 이 예제에서는 만 사용하지만 AWS CLI도 사용할 수 있습니다 AWS Management Console. 자세한 내용은 Amazon Elastic Container Registry 사용 설명서의 리포지토리 생성을 참조하세요.

   ```
   $ aws ecr create-repository --repository-name print-color
   ```

2. 이전 단계에서 반환된 Amazon ECR 리포지토리 URI로 `print-color` 이미지에 태그를 지정합니다.

   ```
   $ docker tag print-color aws_account_id.dkr.ecr.region.amazonaws.com/print-color
   ```

3. Amazon ECR 레지스트리에 로그인합니다. 자세한 내용은 Amazon Elastic Container Registry 사용 설명서의 레지스트리 권한을 참조하세요.

   ```
   $ aws ecr get-login-password \
       --region region | docker login \
       --username AWS \
       --password-stdin aws_account_id.dkr.ecr.region.amazonaws.com
   ```

4. 이미지를 Amazon ECR로 푸시합니다.

   ```
   $ docker push aws_account_id.dkr.ecr.region.amazonaws.com/print-color
   ```

작업 정의 생성 및 등록

이제 Docker 이미지가 이미지 레지스트리에 있으므로 AWS Batch 작업 정의에서 지정할 수 있습니다. 그런 다음 나중에 이를 사용하여 배열 작업을 실행할 수 있습니다. 이 예제에서는 AWS CLI(을)를 사용합니다. 하지만 AWS Management Console(을)를 사용할 수도 있습니다. 자세한 내용은 단일 노드 작업 정의 생성 단원을 참조하십시오.

작업 정의를 생성하려면

1. 작업 영역 디렉터리에 이름이 `print-color-job-def.json`인 파일을 만들고 그 아래에 내용을 붙여 넣습니다. 이미지 리포지토리 URI를 고유의 이미지 URI로 바꿉니다.

```
{
  "jobDefinitionName": "print-color",
  "type": "container",
  "containerProperties": {
    "image": "aws_account_id.dkr.ecr.region.amazonaws.com/print-color",
    "resourceRequirements": [
        {
            "type": "MEMORY",
            "value": "250"
        },
        {
            "type": "VCPU",
            "value": "1"
        }
    ]
  }
}
```

2. 작업 정의를에 등록합니다 AWS Batch.

```
$ aws batch register-job-definition --cli-input-json file://print-color-job-
def.json
```

AWS Batch 배열 작업 제출

작업 정의를 등록한 후 새 컨테이너 이미지를 사용하는 AWS Batch 배열 작업을 제출할 수 있습니다.

AWS Batch 배열 작업을 제출하려면

1. 작업 영역 디렉터리에 이름이 `print-color-job.json`인 파일을 만들고 그 아래에 내용을 붙여 넣습니다.

   > **Note**
   >
   > 이 예에서는 the section called “사전 조건” 섹션에 언급된 작업 대기열을 사용합니다.

   ```
   {
     "jobName": "print-color",
     "jobQueue": "existing-job-queue",
     "arrayProperties": {
       "size": 7
     },
     "jobDefinition": "print-color"
   }
   ```

2. 작업을 AWS Batch 작업 대기열에 제출합니다. 출력에서 반환된 작업 ID를 기록합니다.

   ```
   $ aws batch submit-job --cli-input-json file://print-color-job.json
   ```

3. 작업의 상태를 설명하고 이 작업이 SUCCEEDED(으)로 이동하기를 기다립니다.

배열 작업 로그 보기

작업이 SUCCEEDED 상태에 도달한 후 작업의 컨테이너에서 CloudWatch Logs를 볼 수 있습니다.

CloudWatch Logs에서 작업 로그를 보려면

1. https://console.aws.amazon.com/batch/ AWS Batch 콘솔을 엽니다.
2. 왼쪽 탐색 창에서 작업을 선택합니다.
3. Job queue(작업 대기열)에서 대기열을 선택합니다.
4. 상태 섹션에서 성공을 선택합니다.
5. 배열 작업의 하위 작업을 모두 표시하려면 이전 섹션에서 반환된 작업 ID를 선택합니다.
6. 작업의 컨테이너에서 로그를 보려면 하위 작업 중 하나 선택하고 로그 보기를 선택합니다.

7. 다른 하위 작업의 로그를 봅니다. 각 작업은 다른 무지개색을 반환합니다 .

GPU 작업 실행

GPU 작업을 통해 인스턴스의 GPU를 사용하는 작업을 실행할 수 있습니다.

다음과 같은 Amazon EC2 GPU 기반 인스턴스 유형이 지원됩니다. 자세한 내용은 Amazon EC2 G3 인스턴스, Amazon EC2 G4 인스턴스, Amazon EC2 G5 인스턴스,Amazon EC2 G6 인스턴스 , Amazon EC2 P2 인스턴스, Amazon EC2 P3 인스턴스, Amazon EC2 P4d 인스턴스, Amazon EC2 P5 인스턴스, Amazon EC2 P6 인스턴스, Amazon EC2 Trn1 인스턴스, Amazon EC2 Trn2 인스턴스, Amazon EC2 Inf1 인스턴스, Amazon EC2 Inf2 인스턴스, Amazon EC2 DI1 인스턴스 및 Amazon EC2 DI2 인스턴스를 참조하세요.

인스턴스 유형	GPU	GPU 메모리	vCPU	Memory	네트워크 대역폭
g3s.xlarge	1	8GiB	4	30.5GiB	10Gbps
g3.4xlarge	1	8GiB	16	122GiB	최대 10Gbps
g3.8xlarge	2	16GiB	32	244GiB	10Gbps
g3.16xlarge	4	32GiB	64	488GiB	25Gbps
g4dn.xlarge	1	16GiB	4	16GiB	최대 25Gbps
g4dn.2xlarge	1	16GiB	8	32GiB	최대 25Gbps
g4dn.4xlarge	1	16GiB	16	64GiB	최대 25Gbps
g4dn.8xlarge	1	16GiB	32	128GiB	50Gbps

인스턴스 유형	GPU	GPU 메모리	vCPU	Memory	네트워크 대역폭
g4dn.12xlarge	4	64GiB	48	192GiB	50Gbps
g4dn.16xlarge	1	16GiB	64	256GiB	50Gbps
g5.xlarge	1	24GiB	4	16GiB	최대 10Gbps
g5.2xlarge	1	24GiB	8	32GiB	최대 10Gbps
g5.4xlarge	1	24GiB	16	64GiB	최대 25Gbps
g5.8xlarge	1	24GiB	32	128GiB	25Gbps
g5.16xlarge	1	24GiB	64	256GiB	25Gbps
g5.12xlarge	4	96GiB	48	192GiB	40Gbps
g5.24xlarge	4	96GiB	96	384GiB	50Gbps
g5.48xlarge	8	192GiB	192	768GiB	100Gbps
g5g.xlarge	1	16GiB	4	8GiB	최대 10Gbps
g5g.2xlarge	1	16GiB	8	16GiB	최대 10Gbps
g5g.4xlarge	1	16GiB	16	32GiB	최대 10Gbps
g5g.8xlarge	1	16GiB	32	64GiB	12Gbps
g5g.16xlarge	2	32GiB	64	128GiB	25Gbps
g5g.metal	2	32GiB	64	128GiB	25Gbps
g6.xlarge	1	24GiB	4	16GiB	최대 10Gbps
g6.2xlarge	1	24GiB	8	32GiB	최대 10Gbps
g6.4xlarge	1	24GiB	16	64GiB	최대 25Gbps
g6.8xlarge	1	24GiB	32	128GiB	25Gbps

인스턴스 유형	GPU	GPU 메모리	vCPU	Memory	네트워크 대역폭
g6.16xlarge	1	24GiB	64	256GiB	25Gbps
g6.12xlarge	4	96GiB	48	192GiB	40Gbps
g6.24xlarge	4	96GiB	96	384GiB	50Gbps
g6.48xlarge	8	192GiB	192	768GiB	100Gbps
g6e.xlarge	1	48GiB	4	32GiB	최대 20Gbps
g6e.2xlarge	1	48GiB	8	64GiB	최대 20Gbps
g6e.4xlarge	1	48GiB	16	128GiB	20Gbps
g6e.8xlarge	1	48GiB	32	256GiB	25Gbps
g6e.16xlarge	1	48GiB	64	512GiB	35Gbps
g6e.12xlarge	4	192GiB	48	384 GiB	100Gbps
g6e.24xlarge	4	192GiB	96	768GiB	200Gbps
g6e.48xlarge	8	384 GiB	192	1536GiB	400Gbps
gr6.4xlarge	1	24GiB	16	128GiB	최대 25Gbps
gr6.8xlarge	1	24GiB	32	256GiB	25Gbps
p2.xlarge	1	12GiB	4	61GiB	높음
p2.8xlarge	8	96GiB	32	488GiB	10Gbps
p2.16xlarge	16	192GiB	64	732GiB	20Gbps
p3.2xlarge	1	16GiB	8	61GiB	최대 10Gbps
p3.8xlarge	4	64GiB	32	244GiB	10Gbps
p3.16xlarge	8	128GiB	64	488GiB	25Gbps

인스턴스 유형	GPU	GPU 메모리	vCPU	Memory	네트워크 대역폭
p3dn.24xlarge	8	256GiB	96	768GiB	100Gbps
p4d.24xlarge	8	320GiB	96	1152GiB	400Gbps
p4de.24xlarge	8	640GiB	96	1152GiB	400Gbps
p5.48xlarge	8	640GiB	192	2TiB	3200Gbps
p5e.48xlarge	8	1128GiB	192	2TiB	3200Gbps
p5en.48xlarge	8	1128GiB	192	2TiB	3200Gbps
p6-b200.48xlarge	8	1440GiB	192	2TiB	100Gbps
trn1.2xlarge	1	32GiB	8	32GiB	최대 1.25Gbps
trn1.32xlarge	16	512GiB	128	512GiB	800Gbps
trn1n.32xlarge	16	512GiB	128	512GiB	1600Gbps
trn2.48xlarge	16	1.5TiB	192	2TiB	3.2Tbps
inf1.xlarge	1	8GiB	4	8GiB	최대 25Gbps
inf1.2xlarge	1	8GiB	8	16GiB	최대 25Gbps
inf1.6xlarge	4	32GiB	24	48GiB	25Gbps
inf1.24xlarge	16	128GiB	96	192GiB	100Gbps
inf2.xlarge	1	32GiB	4	16GiB	최대 15Gbps
inf2.8xlarge	1	32GiB	32	128GiB	최대 25Gbps
inf2.24xlarge	6	192GiB	96	384GiB	50Gbps
inf2.48xlarge	12	384 GiB	192	768GiB	100Gbps
dl1.24xlarge	8	256GiB	96	768GiB	400Gbps

인스턴스 유형	GPU	GPU 메모리	vCPU	Memory	네트워크 대역폭
dl2q.24xlarge	8	128GiB	96	768GiB	100Gbps

> **Note**
>
> GPU 작업의 경우 NVIDIA GPUs가 있는 인스턴스 유형 AWS Batch 만 지원합니다. 예를 들어 G4ad 패밀리는 GPU 예약에 지원되지 않습니다. 작업 정의에서 vcpu 및 메모리 요구 사항만 정의 AWS Batch 한 다음 Amazon ECS 또는 Amazon EKS 컴퓨팅 최적화 AMI 또는 AMD GPUs 사용을 위한 사용자 지정 AMI를 사용하여 Amazon EC2 시작 템플릿 사용자 데이터의 사용자 지정을 통해 호스트 GPU에 직접 액세스하여 G4ad에서를 계속 사용할 수 있습니다. Amazon EC2 GPUs
>
> ARM64 아키텍처를 사용하는 인스턴스 유형은 AWS Batch 또는 Amazon EC2 사용자 데이터에 제공된 사용자 지정 AMIs의 GPU 작업에서 사용자 지정 코드 및 구성을 통해 GPUs에 액세스할 수 있도록 지원됩니다. G5g 인스턴스 패밀리를 그 예로 들 수 있습니다.

작업 정의에 대한 resourceRequirements 파라미터는 컨테이너에 고정되며 해당 작업 기간 동안 해당 인스턴스에서 실행 중인 다른 작업에서는 사용할 수 없는 GPU 수를 지정합니다. 이 수의 GPU는 해당 작업이 진행되는 동안 해당 인스턴스에서 실행되는 다른 작업에서는 사용할 수 없습니다. GPU 작업을 실행하는 컴퓨팅 환경의 모든 인스턴스 유형은 p3, p4, p5, p6, g3, g3s, g4, g5 또는 g6 인스턴스 패밀리의 인스턴스 유형이어야 합니다. 그렇지 않으면 GPU 작업이 RUNNABLE 상태로 고착될 수 있습니다.

GPU를 사용하지 않는 작업은 GPU 인스턴스에서 실행할 수 있습니다. 하지만 유사한 비 GPU 인스턴스보다 GPU 인스턴스에서 실행하는 경우 비용이 더 많이 들 수 있습니다. 특정 vCPU, 메모리 및 필요한 시간에 따라 이러한 비 GPU 작업은 GPU 작업의 실행을 차단할 수 있습니다.

주제

- Amazon EKS에서 GPU 기반 Kubernetes 클러스터 생성
- Amazon EKS GPU 작업 정의 생성
- Amazon EKS 클러스터에서 GPU 작업 실행

Amazon EKS에서 GPU 기반 Kubernetes 클러스터 생성

Amazon EKS에서 GPU 기반 Kubernetes 클러스터를 생성하려면 먼저 Amazon EKS AWS Batch 에서 시작하기의 단계를 완료해야 합니다. 추가적으로 다음 사항도 고려하세요.

- AWS Batch 는 NVIDIA GPUs에서 인스턴스 유형을 지원합니다.
- 기본적으로는 Amazon EKS 클러스터 컨트롤 플레인 Kubernetes 버전과 일치하는 버전으로 Amazon EKS 가속 AMI를 AWS Batch 선택합니다.

```
$ cat <<EOF > ./batch-eks-gpu-ce.json
{
  "computeEnvironmentName": "My-Eks-GPU-CE1",
  "type": "MANAGED",
  "state": "ENABLED",
  "eksConfiguration": {
    "eksClusterArn": "arn:aws:eks:<region>:<account>:cluster/<cluster-name>",
    "kubernetesNamespace": "my-aws-batch-namespace"
  },
  "computeResources": {
    "type": "EC2",
    "allocationStrategy": "BEST_FIT_PROGRESSIVE",
    "minvCpus": 0,
    "maxvCpus": 1024,
    "instanceTypes": [
      "p3dn.24xlarge",
      "p4d.24xlarge"
    ],
    "subnets": [
        "<eks-cluster-subnets-with-access-to-internet-for-image-pull>"
    ],
    "securityGroupIds": [
        "<eks-cluster-sg>"
    ],
    "instanceRole": "<eks-instance-profile>"
  }
}
EOF

$ aws batch create-compute-environment --cli-input-json file://./batch-eks-gpu-ce.json
```

AWS Batch 는 사용자를 대신하여 NVIDIA GPU 디바이스 플러그인을 관리하지 않습니다. 이 플러그인을 Amazon EKS 클러스터에 설치하고 AWS Batch 노드를 대상으로 지정할 수 있도록 허용해야 합니다. 자세한 내용은 GitHub의 Kubernetes에서 GPU 지원 활성화를 참조하세요.

AWS Batch 노드를 대상으로 하도록 NVIDIA 디바이스 플러그인(DaemonSet)을 구성하려면 다음 명령을 실행합니다.

```
# pull nvidia daemonset spec
$ curl -O https://raw.githubusercontent.com/NVIDIA/k8s-device-plugin/v0.12.2/nvidia-device-plugin.yml
# using your favorite editor, add Batch node toleration
# this will allow the DaemonSet to run on Batch nodes
- key: "batch.amazonaws.com/batch-node"
  operator: "Exists"

$ kubectl apply -f nvidia-device-plugin.yml
```

동일한 컴퓨팅 환경과 작업 대기열의 쌍에서 컴퓨팅 기반(CPU 및 메모리) 워크로드와 GPU 기반 워크로드를 함께 사용하지 않는 것이 좋습니다. 컴퓨팅 작업이 GPU 용량을 소모할 수 있기 때문입니다.

작업 대기열을 연결하려면 다음 명령을 실행합니다.

```
$ cat <<EOF > ./batch-eks-gpu-jq.json
 {
    "jobQueueName": "My-Eks-GPU-JQ1",
    "priority": 10,
    "computeEnvironmentOrder": [
      {
        "order": 1,
        "computeEnvironment": "My-Eks-GPU-CE1"
      }
    ]
  }
EOF

$ aws batch create-job-queue --cli-input-json file://./batch-eks-gpu-jq.json
```

Amazon EKS GPU 작업 정의 생성

현재는 nvidia.com/gpu만 지원되며, 설정하는 리소스 값은 정수여야 합니다. GPU의 일부만 사용할 수는 없습니다. 자세한 내용은 Kubernetes 설명서의 GPU 예약을 참조하세요.

Amazon EKS용 GPU 작업 정의를 등록하려면 다음 명령을 실행합니다.

```
$ cat <<EOF > ./batch-eks-gpu-jd.json
{
    "jobDefinitionName": "MyGPUJobOnEks_Smi",
    "type": "container",
    "eksProperties": {
        "podProperties": {
            "hostNetwork": true,
            "containers": [
                {
                    "image": "nvcr.io/nvidia/cuda:10.2-runtime-centos7",
                    "command": ["nvidia-smi"],
                    "resources": {
                        "limits": {
                            "cpu": "1",
                            "memory": "1024Mi",
                            "nvidia.com/gpu": "1"
                        }
                    }
                }
            ]
        }
    }
}
EOF

$ aws batch register-job-definition --cli-input-json file://./batch-eks-gpu-jd.json
```

Amazon EKS 클러스터에서 GPU 작업 실행

GPU 리소스는 압축할 수 없습니다. 요청 값이 제한 값과 동일한 GPU 작업에 대한 포드 사양을 AWS Batch 생성합니다. 이는 Kubernetes 요구 사항입니다.

작업을 다시 시작하려면 다음 명령을 실행합니다.

```
$ aws batch submit-job --job-queue My-Eks-GPU-JQ1 --job-definition MyGPUJobOnEks_Smi --
job-name My-Eks-GPU-Job

# locate information that can help debug or find logs (if using Amazon CloudWatch Logs
 with Fluent Bit)
$ aws batch describe-jobs --job <job-id> | jq '.jobs[].eksProperties.podProperties |
 {podName, nodeName}'
```

```
{
  "podName": "aws-batch.f3d697c4-3bb5-3955-aa6c-977fcf1cb0ca",
  "nodeName": "ip-192-168-59-101.ec2.internal"
}
```

AWS Batch 작업 대기열에서 작업 보기

AWS Batch에서 작업을 보고 필터링할 수 있습니다. 이 기능은 기존 작업 대기열을 보고 세 가지 옵션 중 하나를 사용하여 작업을 필터링하는 옵션을 제공합니다.

검색 및 필터는 최종 상태(SUCCEEDED 또는 FAILED)가 아닌 작업을 검색할 수 있습니다. 작업 상태가 SUCCEEDED 또는 FAILED가 되면 최대 7일 동안 작업을 검색할 수 있습니다. 여전히 작업의 CloudWatch 또는 Amazon EventBridge 로그를 볼 수 있습니다.

이 절차를 사용하여 AWS Batch 콘솔의 작업 대기열에 있는 모든 작업을 나열합니다. 선택적으로, 필터 결과 필드를 지정한 기준에 따라 결과의 범위를 좁힙니다.

1. AWS Batch 콘솔로 이동합니다.
2. 탐색 창에서 작업을 선택합니다.
3. 작업 대기열 드롭다운 목록을 확장하고 검색할 작업 대기열을 선택합니다.

   > **Note**
   >
   > 한 번에 하나의 작업 대기열 내에서만 작업을 검색할 수 있습니다.

4. 결과 필터링 필드에 결과에 포함할 키워드를 입력합니다. 이 필드를 사용하여 작업 이름, 상태 또는 작업 ID를 기준으로 필터링할 수 있습니다. 속성에 따라서는 같음(=) 또는 포함(^)과 같은 추가 연산자를 정의해야 할 수 있습니다.

   > **Note**
   >
   > SageMaker 훈련 작업 대기열은 작업 이름 및 작업 ID별 필터링만 지원합니다.

5. 검색을 선택합니다.

작업 대기열에서 작업 AWS Batch 검색

작업 검색을 AWS Batch 사용하여에서 작업을 검색하고 필터링할 수 있습니다. 이 기능은 기존 작업 대기열에서 작업을 검색하고 필터링하는 옵션을 제공합니다.

검색 및 필터는 최종 상태(SUCCEEDED 또는 FAILED)가 아닌 작업을 검색할 수 있습니다. 작업 상태가 SUCCEEDED 또는 FAILED가 되면 최대 7일 동안 작업을 검색할 수 있습니다. 여전히 작업의 CloudWatch 또는 Amazon EventBridge 로그를 볼 수 있습니다.

여러 기준을 동시에 사용하여 검색하려면 고급 검색 기능을 사용합니다. 예를 들어 상태, 날짜 범위 및 추가 기준(예: 작업 이름, 작업 정의, 작업 ID)과 같은 필터 중 일부 또는 전부를 포함할 수 있습니다.

검색 AWS Batch 작업(AWS 콘솔)

이 절차를 사용하여 AWS Batch 콘솔의 작업 대기열에서 작업을 검색합니다.

1. AWS Batch 콘솔로 이동합니다.
2. 탐색 창에서 작업을 선택합니다.
3. 고급 검색을 켭니다.
4. 작업 대기열 드롭다운 목록을 확장하고 검색할 작업 대기열을 선택합니다.

   > **Note**
   >
   > 한 번에 하나의 작업 대기열 내에서만 작업을 검색할 수 있습니다.

5. 검색 옵션:

   a. 상태 드롭다운 목록에서 필터링할 상태를 하나 이상 선택할 수 있습니다. 자세한 내용은 작업 상태 및 서비스 작업 상태 섹션을 참조하세요.

      > **Note**
      >
      > 배열 작업 상위는 하위 작업이 로 업데이트될 PENDING 때 로 업데이트되고 하위 작업이 실행되는 동안 PENDING 상태를 RUNNABLE 유지합니다. 이러한 작업을 보려면 모든 하위 작업이 터미널 상태에 도달할 때까지 PENDING 상태를 기준으로 필터링합니다.

   b. 날짜 범위를 선택하여 날짜 및 시간 범위를 기준으로 결과를 필터링합니다.

- 상대 모드를 선택하여 현재 날짜 및 시간에서 역순으로 계산한 시간 범위 내에 생성 날짜를 가진 작업을 검색합니다.
- 절대 모드를 선택하여 지정한 날짜 및 시간 범위 내에 생성 날자를 가진 작업을 검색합니다.

c. 추가 기준 필드에 검색 결과에 포함할 키워드를 입력합니다. 예를 들어이 필드를 사용하여 작업 이름, 작업 정의, 작업 ID 또는 공유 식별자로 검색할 수 있습니다. 속성에 따라서는 같음(=) 또는 포함(^)과 같은 추가 연산자를 정의해야 할 수 있습니다.

> **Note**
>
> SageMaker 훈련 작업 대기열은 작업 이름 및 작업 ID별 필터링만 지원합니다.

> **Note**
>
> 공유 식별자를 기준으로 필터링할 때 작업 상태를 지정할 수도 있습니다. 이는 다른 필터가 작업 상태 필터링을 제외하는 제한에 대한 예외입니다.

6. 검색을 선택합니다.

AWS Batch 작업 검색 및 필터링(AWS CLI)

이 절차를 사용하여 AWS CLI로 작업 대기열에 있는 모든 작업을 나열합니다. 선택적으로, -filters 파라미터를 사용하여 지정한 기준에 따라 결과의 범위를 좁힙니다.

Search job queue (AWS CLI)

list-jobs 명령을 사용하여 작업 대기열을 검색하고 필터링할 수 있습니다.

예를 들어 작업 이름을 기반으로 작업 대기열을 검색할 수 있습니다.

```
aws batch list-jobs \
    --job-queue my-job-queue \
    --filters name=JOB_NAME,values="my-job"
```

공유 식별자를 기준으로 작업을 필터링합니다.

```
aws batch list-jobs \
```

```
    --job-queue my-job-queue \
    --filters name=SHARE_IDENTIFIER,values="my-share"
```

공유 식별자를 기준으로 필터링할 때 작업 상태를 포함할 수 있습니다.

```
aws batch list-jobs \
    --job-queue my-job-queue \
    --job-status RUNNING \
    --filters name=SHARE_IDENTIFIER,values="my-share"
```

위의 명령에서 다음과 같이 변경합니다.

- *my-job-queue*를 작업 대기열의 이름으로 바꿉니다.
- *my-job*을 작업의 이름으로 바꿉니다.
- *my-share*를 필터링하려는 공유 식별자로 바꿉니다.

Search service job queue (AWS CLI)

list-service-jobs 명령을 사용하여 서비스 작업 대기열을 검색하고 필터링할 수 있습니다.

예를 들어 작업 이름을 기반으로 서비스 작업 대기열을 검색할 수 있습니다.

```
aws batch list-service-jobs \
    --job-queue my-sm-queue \
    --filters name=JOB_NAME,values="my-sm-job"
```

공유 식별자를 기준으로 서비스 작업을 필터링합니다.

```
aws batch list-service-jobs \
    --job-queue my-sm-queue \
    --filters name=SHARE_IDENTIFIER,values="my-share"
```

위의 명령에서 다음과 같이 변경합니다.

- *my-sm-queue*를 서비스 작업 대기열의 이름으로 바꿉니다.
- *my-sm-job*을 서비스 작업의 이름으로 바꿉니다.
- *my-share*를 필터링하려는 공유 식별자로 바꿉니다.

AWS Batch 작업에 대한 네트워킹 모드

다음 표에서는 네트워킹 모드와 AWS Batch 작업 유형의 일반적인 사용에 대해 설명합니다. 고려 사항 및 동작에 대한 자세한 내용은 '작업 유형' 열의 링크를 참조하세요.

작업 유형	지원되는 네트워크 모드	일반적인 사용
ECS-EC2 단순 작업	`host`	컴퓨팅 환경에 정의된 vpc로의 송신만 필요한 가장 확장 가능한 처치 곤란 병렬 배치 워크로드에 사용됩니다.
ECS-EC2 다중 노드 병렬 작업	`awsvpc`	태스크 노드 간의 조정된 통신을 통해 단일 작업으로 모델링된 긴밀하게 결합된 다중 호스트(노드) 분산 워크로드에 사용됩니다.
ECS-Fargate 단순 작업	`awsvpc`	처치 곤란 병렬 배치 워크로드를 위한 트루 서버리스입니다. 일반적으로 가장 낮은 TCO 및 가장 높은 컨테이너 격리 작업 모델입니다.
EKS-EC2 단순 작업	호스트 및 포드	컴퓨팅 환경에 정의된 vpc로의 송신만 필요한 고확장성의 처치 곤란 병렬 배치 워크로드에 사용됩니다. 기본값은 호스트 네트워킹입니다.
EKS-EC2 다중 노드 병렬 작업	호스트 및 포드	포드 노드 간의 조정된 통신을 통해 단일 작업으로 모델링된 긴밀하게 결합된 다중 호스트(노드) 분산 워크로드에 사용됩니다. 기본값은 호스트 네트워킹입니다.

CloudWatch Logs에서 AWS Batch 작업 로그 보기

Amazon CloudWatch Logs로 로그 정보를 전송하도록 AWS Batch 작업을 구성할 수 있습니다. 이렇게 하면 한 곳에서 작업의 다양한 로그를 편리하게 볼 수 있습니다. 자세한 내용은 에서 CloudWatch Logs 사용 AWS Batch 단원을 참조하십시오.

AWS Batch 콘솔에서 작업 로그를 사용하여 AWS Batch 작업을 모니터링하거나 문제를 해결할 수도 있습니다.

1. AWS Batch 콘솔을 엽니다.
2. 작업을 선택합니다. 작업 대기열의 정렬 및 필터링 작업에 대한 자세한 내용은 AWS Batch 작업 대기열에서 작업 보기 및 에서 작업 대기열의 작업 검색 섹션을 참조하세요.
3. 작업 대기열에서 원하는 작업 대기열을 선택합니다.

> **Tip**
>
> 작업 대기열에 여러 작업이 있는 경우 검색 및 필터링 기능을 켜서 작업을 더 빨리 찾을 수 있습니다. 자세한 내용은 작업 대기열에서 작업 AWS Batch 검색 단원을 참조하십시오.

4. 상태에서 원하는 작업 상태를 선택합니다.
5. 원하는 작업을 선택하면 세부 정보 페이지가 열립니다.
6. 세부 정보 페이지에서 아래의 로그 스트림 이름으로 스크롤하여 링크를 선택합니다. 링크를 클릭하면 작업에 대한 Amazon CloudWatch Logs 페이지가 열립니다.
7. (선택 사항) 이번이 로그를 처음 보는 경우라면 승인 요청 메시지가 표시될 수 있습니다.

   필요한 승인에 **OK**(을)를 입력한 다음 승인을 선택하여 Amazon CloudWatch 요금을 수락합니다.

> **Note**
>
> CloudWatch 요금에 대한 승인을 취소하려면:
>
> 1. 왼쪽 탐색 창에서 권한을 선택합니다.
> 2. 작업 로그에서 편집을 선택합니다.
> 3. CloudWatch를 사용하도록 배치에 승인 확인란을 해제합니다.
> 4. 변경 사항 저장을 선택합니다.

AWS Batch 작업 정보 검토

상태, 작업 정의, 컨테이너 정보와 같은 AWS Batch 작업 정보를 검토할 수 있습니다.

1. AWS Batch 콘솔을 엽니다.
2. 작업을 선택합니다.
3. 작업 대기열에서 원하는 작업 대기열을 선택합니다.

> **Tip**
>
> 작업 대기열에 여러 작업이 있는 경우 검색 및 필터를 설정하여 작업을 더 빨리 찾을 수 있습니다. 자세한 내용은 작업 대기열에서 작업 AWS Batch 검색 단원을 참조하십시오.

4. 원하는 색상을 선택합니다.

> **Note**
>
> AWS Command Line Interface (AWS CLI)를 사용하여 AWS Batch 작업에 대한 세부 정보를 볼 수도 있습니다. 자세한 내용은 AWS CLI 명령 레퍼런스에서 describe-domain을 참조하세요.

의 보안 AWS Batch

의 클라우드 보안 AWS 이 최우선 순위입니다. AWS 고객은 보안에 가장 민감한 조직의 요구 사항을 충족하도록 구축된 데이터 센터 및 네트워크 아키텍처의 이점을 누릴 수 있습니다.

보안은 AWS 와 사용자 간의 공동 책임입니다. 공동 책임 모델은 이 사항을 클라우드의 보안 및 클라우드 내 보안으로 설명합니다.

- 클라우드 보안 - AWS 는에서 AWS 서비스를 실행하는 인프라를 보호할 책임이 있습니다 AWS 클라우드. AWS 또한는 안전하게 사용할 수 있는 서비스를 제공합니다. 타사 감사자는 AWS 규정 준수 프로그램 일환으로 보안의 효과를 정기적으로 테스트하고 확인합니다. 에 적용되는 규정 준수 프로그램에 대한 자세한 내용은 규정 준수 프로그램 AWS 제공 범위 내 서비스규정 준수 프로그램 제공 범위 내 서비스를 AWS Batch참조하세요.
- 클라우드 내 보안 – 귀하의 책임은 귀하가 사용하는 AWS 서비스에 따라 결정됩니다. 또한 귀하는 데이터의 민감도, 회사 요구 사항, 관련 법률 및 규정을 비롯한 기타 요소에 대해서도 책임이 있습니다.

이 설명서는를 사용할 때 공동 책임 모델을 적용하는 방법을 이해하는 데 도움이 됩니다 AWS Batch. 다음 주제에서는 보안 및 규정 준수 목표를 충족하도록 AWS Batch 를 구성하는 방법을 보여줍니다. 또한 AWS Batch 리소스를 모니터링하고 보호하는 데 도움이 되는 다른 AWS 서비스를 사용하는 방법을 알아봅니다.

주제

- 에 대한 자격 증명 및 액세스 관리 AWS Batch
- AWS Batch IAM 정책, 역할 및 권한
- AWS Batch IAM 실행 역할
- Virtual Private Cloud 생성
- 인터페이스 엔드포인트를 사용하여 액세스 AWS Batch
- 에 대한 규정 준수 검증 AWS Batch
- 의 인프라 보안 AWS Batch
- 교차 서비스 혼동된 대리자 방지
- 를 사용하여 AWS Batch API 호출 로깅 AWS CloudTrail
- AWS Batch IAM 문제 해결

에 대한 자격 증명 및 액세스 관리 AWS Batch

AWS Identity and Access Management (IAM)는 관리자가 AWS 리소스에 대한 액세스를 안전하게 제어하는 데 도움이 AWS 서비스 되는 입니다. IAM 관리자는 누가 AWS Batch 리소스를 사용할 수 있는 인증(로그인) 및 권한(권한 있음)을 받을 수 있는지 제어합니다. IAM은 추가 비용 없이 사용할 수 AWS 서비스 있는 입니다.

주제

대상

AWS Identity and Access Management (IAM)를 사용하는 방법은 역할에 따라 다릅니다.

- 서비스 사용자 - 기능에 액세스할 수 없는 경우 관리자에게 권한 요청(참조AWS Batch IAM 문제 해결)
- 서비스 관리자 - 사용자 액세스 결정 및 권한 요청 제출(AWS Batch 에서 IAM을 사용하는 방법 참조)
- IAM 관리자 - 액세스를 관리하기 위한 정책 작성(에 대한 자격 증명 기반 정책 예제 AWS Batch 참조)

ID를 통한 인증

인증은 자격 증명 자격 증명을 AWS 사용하여에 로그인하는 방법입니다. AWS 계정 루트 사용자, IAM 사용자 또는 IAM 역할을 수임하여 인증되어야 합니다.

AWS IAM Identity Center (IAM Identity Center), Single Sign-On 인증 또는 Google/Facebook 자격 증명과 같은 자격 증명 소스의 자격 증명을 사용하여 페더레이션 자격 증명으로 로그인할 수 있습니다. 로그인하는 방법에 대한 자세한 내용은 AWS 로그인 사용 설명서의 AWS 계정에 로그인하는 방법 섹션을 참조하세요.

프로그래밍 방식 액세스를 위해는 요청에 암호화 방식으로 서명할 수 있는 SDK 및 CLI를 AWS 제공합니다. 자세한 내용은 IAM 사용 설명서의 API 요청용AWS Signature Version 4 섹션을 참조하세요.

AWS 계정 루트 사용자

를 생성할 때 모든 AWS 서비스 및 리소스에 대한 완전한 액세스 권한이 있는 AWS 계정 theroot 사용자라는 하나의 로그인 자격 증명으로 AWS 계정시작합니다. 일상적인 태스크에 루트 사용자를 사용하지 않을 것을 강력히 권장합니다. 루트 사용자 자격 증명이 필요한 작업은 IAM 사용 설명서의 루트 사용자 자격 증명이 필요한 작업 섹션을 참조하세요.

페더레이션 ID

가장 좋은 방법은 인간 사용자가 자격 증명 공급자와의 페더레이션을 사용하여 임시 자격 증명을 AWS 서비스 사용하여에 액세스하도록 요구하는 것입니다.

페더레이션 자격 증명은 엔터프라이즈 디렉터리, 웹 자격 증명 공급자 또는 자격 증명 소스의 자격 증명을 AWS 서비스 사용하여 Directory Service 에 액세스하는 사용자입니다. 페더레이션 ID는 임시 자격 증명을 제공하는 역할을 수임합니다.

중앙 집중식 액세스 관리를 위해 AWS IAM Identity Center를 추천합니다. 자세한 정보는 AWS IAM Identity Center 사용 설명서의 What is IAM Identity Center?를 참조하세요.

IAM 사용자 및 그룹

IAM 사용자는 단일 개인 또는 애플리케이션에 대한 특정 권한을 가진 ID입니다. 장기 자격 증명이 있는 IAM 사용자 대신 임시 자격 증명을 사용하는 것이 좋습니다. 자세한 내용은 IAM 사용 설명서의 자격 증명 공급자와의 페더레이션을 사용하여 임시 자격 증명을 AWS 사용하여에 액세스하도록 인간 사용자에게 요구하기를 참조하세요.

IAM 그룹은 IAM 사용자 모음을 지정하고 대규모 사용자 집합에 대한 관리 권한을 더 쉽게 만듭니다. 자세한 내용은 IAM 사용 설명서의 IAM 사용자 사용 사례 섹션을 참조하세요.

IAM 역할

IAM 역할은 임시 자격 증명을 제공하는 특정 권한이 있는 자격 증명입니다. 사용자에서 IAM 역할(콘솔)로 전환하거나 또는 API 작업을 호출하여 역할을 수임할 수 있습니다. AWS CLI AWS 자세한 내용은 IAM 사용 설명서의 역할 수임 방법을 참조하세요.

IAM 역할은 페더레이션 사용자 액세스, 임시 IAM 사용자 권한, 교차 계정 액세스, 교차 서비스 액세스 및 Amazon EC2에서 실행되는 애플리케이션에 유용합니다. 자세한 내용은 IAM 사용 설명서의 교차 계정 리소스 액세스를 참조하세요.

정책을 사용하여 액세스 관리

정책을 AWS 생성하고 자격 증명 또는 리소스에 연결하여 AWS 에서 액세스를 제어합니다. 정책은 자격 증명 또는 리소스와 연결될 때 권한을 정의합니다.는 보안 주체가 요청할 때 이러한 정책을 AWS 평가합니다. 대부분의 정책은에 JSON 문서 AWS 로 저장됩니다. JSON 정책 문서에 대한 자세한 내용은 IAM 사용 설명서의 JSON 정책 개요 섹션을 참조하세요.

정책을 사용하여 관리자는 어떤 보안 주체가 어떤 리소스에 대해 어떤 조건에서 작업을 수행할 수 있는지 정의하여 누가 무엇을 액세스할 수 있는지 지정합니다.

기본적으로 사용자 및 역할에는 어떠한 권한도 없습니다. IAM 관리자는 IAM 정책을 생성하고 사용자가 수임할 수 있는 역할에 추가합니다. IAM 정책은 작업을 수행하기 위해 사용하는 방법과 관계없이 작업에 대한 권한을 정의합니다.

ID 기반 정책

ID 기반 정책은 ID(사용자, 사용자 그룹 또는 역할)에 연결하는 JSON 권한 정책 문서입니다. 이러한 정책은 자격 증명이 수행할 수 있는 작업, 대상 리소스 및 이에 관한 조건을 제어합니다. ID 기반 정책을 생성하는 방법을 알아보려면 IAM 사용 설명서에서 고객 관리형 정책으로 사용자 지정 IAM 권한 정의를 참조하세요.

ID 기반 정책은 인라인 정책(단일 ID에 직접 포함) 또는 관리형 정책(여러 ID에 연결된 독립 실행형 정책)일 수 있습니다. 관리형 정책 또는 인라인 정책을 선택하는 방법을 알아보려면 IAM 사용 설명서의 관리형 정책 및 인라인 정책 중에서 선택 섹션을 참조하세요.

리소스 기반 정책

리소스 기반 정책은 리소스에 연결하는 JSON 정책 설명서입니다. 예를 들어 IAM 역할 신뢰 정책 및 Amazon S3 버킷 정책이 있습니다. 리소스 기반 정책을 지원하는 서비스에서 서비스 관리자는 이러한 정책을 사용하여 특정 리소스에 대한 액세스를 통제할 수 있습니다. 리소스 기반 정책에서 보안 주체를 지정해야 합니다.

리소스 기반 정책은 해당 서비스에 있는 인라인 정책입니다. 리소스 기반 정책에서는 IAM의 AWS 관리형 정책을 사용할 수 없습니다.

기타 정책 유형

AWS 는 보다 일반적인 정책 유형에서 부여한 최대 권한을 설정할 수 있는 추가 정책 유형을 지원합니다.

- 권한 경계 - ID 기반 정책에서 IAM 엔터티에 부여할 수 있는 최대 권한을 설정합니다. 자세한 정보는 IAM 사용 설명서의 IAM 엔터티의 권한 범위를 참조하세요.
- 서비스 제어 정책(SCP) - AWS Organizations내 조직 또는 조직 단위에 대한 최대 권한을 지정합니다. 자세한 내용은AWS Organizations 사용 설명서의 서비스 제어 정책을 참조하세요.
- 리소스 제어 정책(RCP) – 계정의 리소스에 사용할 수 있는 최대 권한을 설정합니다. 자세한 내용은 AWS Organizations 사용 설명서의 리소스 제어 정책(RCP)을 참조하세요.
- 세션 정책 – 역할 또는 페더레이션 사용자에 대해 임시 세션을 프로그래밍 방식으로 생성할 때 파라미터로 전달하는 고급 정책입니다. 자세한 내용은 IAM 사용 설명서의 세션 정책을 참조하세요.

여러 정책 유형

여러 정책 유형이 요청에 적용되는 경우, 결과 권한은 이해하기가 더 복잡합니다. 에서 여러 정책 유형이 관련될 때 요청을 허용할지 여부를 AWS 결정하는 방법을 알아보려면 IAM 사용 설명서의 정책 평가 로직을 참조하세요.

AWS Batch 에서 IAM을 사용하는 방법

IAM을 사용하여에 대한 액세스를 관리하기 전에 사용할 수 있는 IAM 기능을 AWS Batch알아봅니다 AWS Batch.

에서 사용할 수 있는 IAM 기능 AWS Batch

IAM 특성	AWS Batch 지원
자격 증명 기반 정책	예
리소스 기반 정책	아니요
정책 작업	예
정책 리소스	예
정책 조건 키	예
ACL	아니요
ABAC(정책의 태그)	예
임시 보안 인증	예

IAM 특성	AWS Batch 지원
엔터티 권한	예
서비스 역할	예
서비스 연결 역할	예

AWS Batch 및 기타 AWS 서비스가 대부분의 IAM 기능과 작동하는 방법을 개괄적으로 알아보려면 IAM 사용 설명서의 AWS IAM으로 작업하는 서비스를 참조하세요.

에 대한 자격 증명 기반 정책 AWS Batch

ID 기반 정책 지원: 예

ID 기반 정책은 IAM 사용자, 사용자 그룹 또는 역할과 같은 ID에 연결할 수 있는 JSON 권한 정책 문서입니다. 이러한 정책은 사용자 및 역할이 어떤 리소스와 어떤 조건에서 어떤 작업을 수행할 수 있는지를 제어합니다. 자격 증명 기반 정책을 생성하는 방법을 알아보려면 IAM 사용 설명서에서 고객 관리형 정책으로 사용자 지정 IAM 권한 정의를 참조하세요.

IAM ID 기반 정책을 사용하면 허용되거나 거부되는 작업과 리소스뿐 아니라 작업이 허용되거나 거부되는 조건을 지정할 수 있습니다. JSON 정책에서 사용할 수 있는 모든 요소에 대해 알아보려면 IAM 사용 설명서의 IAM JSON 정책 요소 참조를 참조하세요.

에 대한 자격 증명 기반 정책 예제 AWS Batch

자격 AWS Batch 증명 기반 정책의 예를 보려면 섹션을 참조하세요에 대한 자격 증명 기반 정책 예제 AWS Batch.

에 대한 정책 작업 AWS Batch

정책 작업 지원: 예

관리자는 AWS JSON 정책을 사용하여 누가 무엇에 액세스할 수 있는지 지정할 수 있습니다. 즉, 어떤 보안 주체가 어떤 리소스와 어떤 조건에서 작업을 수행할 수 있는지를 지정할 수 있습니다.

JSON 정책의 `Action`요소는 정책에서 액세스를 허용하거나 거부하는 데 사용할 수 있는 작업을 설명합니다. 연결된 작업을 수행할 수 있는 권한을 부여하기 위한 정책에 작업을 포함하세요.

AWS Batch 작업 목록을 보려면 서비스 승인 참조의 에서 정의한 작업을 AWS Batch 참조하세요.

의 정책 작업은 작업 앞에 다음 접두사를 AWS Batch 사용합니다.

```
batch
```

단일 문에서 여러 작업을 지정하려면 쉼표로 구분합니다.

```
"Action": [
      "batch:action1",
      "batch:action2"
         ]
```

와일드카드(*)를 사용하여 여러 작업을 지정할 수 있습니다. 예를 들어, `Describe`라는 단어로 시작하는 모든 작업을 지정하려면 다음 작업을 포함합니다.

```
"Action": "batch:Describe*"
```

자격 AWS Batch 증명 기반 정책의 예를 보려면 섹션을 참조하세요에 대한 자격 증명 기반 정책 예제 AWS Batch.

에 대한 정책 리소스 AWS Batch

정책 리소스 지원: 예

관리자는 AWS JSON 정책을 사용하여 누가 무엇에 액세스할 수 있는지 지정할 수 있습니다. 즉, 어떤 보안 주체가 어떤 리소스와 어떤 조건에서 작업을 수행할 수 있는지를 지정할 수 있습니다.

`Resource` JSON 정책 요소는 작업이 적용되는 하나 이상의 객체를 지정합니다. 모범 사례에 따라 Amazon 리소스 이름(ARN)을 사용하여 리소스를 지정합니다. 리소스 수준 권한을 지원하지 않는 작업의 경우, 와일드카드(*)를 사용하여 해당 문이 모든 리소스에 적용됨을 나타냅니다.

```
"Resource": "*"
```

AWS Batch 리소스 유형 및 해당 ARNs 목록을 보려면 서비스 승인 참조의 에서 정의한 리소스를 AWS Batch 참조하세요. 각 리소스의 ARN을 지정할 수 있는 작업을 알아보려면 AWS Batch가 정의한 작업을 참조하세요.

AWS Batch정책 조건 키

서비스별 정책 조건 키 지원: 예

관리자는 AWS JSON 정책을 사용하여 누가 무엇에 액세스할 수 있는지 지정할 수 있습니다. 즉, 어떤 보안 주체가 어떤 리소스와 어떤 조건에서 작업을 수행할 수 있는지를 지정할 수 있습니다.

`Condition` 요소는 정의된 기준에 따라 문이 실행되는 시기를 지정합니다. 같음(equals) 또는 미만(less than)과 같은 조건 연산자를 사용하여 정책의 조건을 요청의 값과 일치시키는 조건식을 생성할 수 있습니다. 모든 AWS 전역 조건 키를 보려면 IAM 사용 설명서의 AWS 전역 조건 컨텍스트 키를 참조하세요.

AWS Batch 조건 키 목록을 보려면 서비스 승인 참조의에 대한 조건 키를 참조하세요 AWS Batch. 조건 키를 사용할 수 있는 작업과 리소스를 알아보려면 에서 정의한 작업을 AWS Batch 참조하세요.

를 사용한 ABAC(속성 기반 액세스 제어) AWS Batch

ABAC 지원(정책의 태그): 예

속성 기반 액세스 제어(ABAC)는 태그라고 불리는 속성을 기반으로 권한을 정의하는 권한 부여 전략입니다. IAM 엔터티 및 AWS 리소스에 태그를 연결한 다음 보안 주체의 태그가 리소스의 태그와 일치할 때 작업을 허용하는 ABAC 정책을 설계할 수 있습니다.

태그에 근거하여 액세스를 제어하려면 `aws:ResourceTag/key-name`, `aws:RequestTag/key-name`또는 `aws:TagKeys`조건 키를 사용하여 정책의 조건 요소에 태그 정보를 제공합니다.

서비스가 모든 리소스 유형에 대해 세 가지 조건 키를 모두 지원하는 경우, 값은 서비스에 대해 예입니다. 서비스가 일부 리소스 유형에 대해서만 세 가지 조건 키를 모두 지원하는 경우, 값은 부분적입니다.

ABAC에 대한 자세한 내용은 IAM 사용 설명서의 ABAC 권한 부여를 통한 권한 정의를 참조하세요. ABAC 설정 단계가 포함된 자습서를 보려면 IAM 사용 설명서의 속성 기반 액세스 제어(ABAC) 사용을 참조하세요.

에서 임시 자격 증명 사용 AWS Batch

임시 자격 증명 지원: 예

임시 자격 증명은 AWS 리소스에 대한 단기 액세스를 제공하며 페더레이션 또는 전환 역할을 사용할 때 자동으로 생성됩니다. 장기 액세스 키를 사용하는 대신 임시 자격 증명을 동적으로 생성하는 것이 AWS 좋습니다. 자세한 내용은 IAM 사용 설명서의 IAM의 임시 보안 자격 증명 및 IAM으로 작업하는 AWS 서비스 섹션을 참조하세요.

에 대한 교차 서비스 보안 주체 권한 AWS Batch

전달 액세스 세션(FAS) 지원: 예

전달 액세스 세션(FAS)은를 호출하는 보안 주체의 권한을 다운스트림 서비스에 대한 요청 AWS 서비스 과 AWS 서비스함께 사용합니다. FAS 요청 시 정책 세부 정보는 전달 액세스 세션을 참조하세요.

에 대한 서비스 역할 AWS Batch

서비스 역할 지원: 예

서비스 역할은 서비스가 사용자를 대신하여 작업을 수행하는 것으로 가정하는 IAM 역할입니다. IAM 관리자는 IAM 내에서 서비스 역할을 생성, 수정 및 삭제할 수 있습니다. 자세한 내용은 IAM 사용 설명서의 AWS 서비스 AWS에 권한을 위임할 역할 생성을 참조하세요.

> **Warning**
>
> 서비스 역할에 대한 권한을 변경하면 AWS Batch 기능이 중단될 수 있습니다. AWS Batch 이 그 일을 하라는 지침을 제공하는 경우에만 서비스 역할을 편집합니다.

에 대한 서비스 연결 역할 AWS Batch

서비스 연결 역할 지원: 예

서비스 연결 역할은에 연결된 서비스 역할의 한 유형입니다 AWS 서비스. 서비스는 사용자를 대신하여 태스크를 수행하기 위해 역할을 수임할 수 있습니다. 서비스 연결 역할은에 나타나 AWS 계정 며 서비스가 소유합니다. IAM 관리자는 서비스 연결 역할의 권한을 볼 수 있지만 편집은 할 수 없습니다.

서비스 연결 역할 생성 또는 관리에 대한 자세한 내용은 IAM으로 작업하는AWS 서비스를 참조하세요. 서비스 연결 역할 열에서 Yes가 포함된 서비스를 테이블에서 찾습니다. 해당 서비스에 대한 서비스 연결 역할 설명서를 보려면 예(Yes) 링크를 선택합니다.

에 대한 자격 증명 기반 정책 예제 AWS Batch

기본적으로 사용자 및 역할에는 AWS Batch 리소스를 생성하거나 수정할 수 있는 권한이 없습니다. 사용자에게 사용자가 필요한 리소스에서 작업을 수행할 권한을 부여하려면 IAM 관리자가 IAM 정책을 생성하면 됩니다.

이러한 예제 JSON 정책 문서를 사용하여 IAM ID 기반 정책을 생성하는 방법을 알아보려면 IAM 사용 설명서의 IAM 정책 생성(콘솔)을 참조하세요.

각 리소스 유형에 대한 ARNs 형식을 포함하여 AWS Batch에서 정의한 작업 및 리소스 유형에 대한 자세한 내용은 서비스 승인 참조의에 대한 작업, 리소스 및 조건 키를 참조하세요 AWS Batch.

주제

- 정책 모범 사례
- AWS Batch 콘솔 사용
- 사용자가 자신의 고유한 권한을 볼 수 있도록 허용

정책 모범 사례

자격 증명 기반 정책에 따라 계정에서 사용자가 AWS Batch 리소스를 생성, 액세스 또는 삭제할 수 있는지 여부가 결정됩니다. 이 작업으로 인해 AWS 계정에 비용이 발생할 수 있습니다. ID 기반 정책을 생성하거나 편집할 때는 다음 지침과 권장 사항을 따르세요.

- AWS 관리형 정책을 시작하고 최소 권한으로 전환 - 사용자 및 워크로드에 권한 부여를 시작하려면 많은 일반적인 사용 사례에 대한 권한을 부여하는 AWS 관리형 정책을 사용합니다. 에서 사용할 수 있습니다 AWS 계정. 사용 사례에 맞는 AWS 고객 관리형 정책을 정의하여 권한을 추가로 줄이는 것이 좋습니다. 자세한 내용은 IAM 사용 설명서의 AWS 관리형 정책 또는 AWS 직무에 대한 관리형 정책을 참조하세요.
- 최소 권한 적용 – IAM 정책을 사용하여 권한을 설정하는 경우, 작업을 수행하는 데 필요한 권한만 부여합니다. 이렇게 하려면 최소 권한으로 알려진 특정 조건에서 특정 리소스에 대해 수행할 수 있는 작업을 정의합니다. IAM을 사용하여 권한을 적용하는 방법에 대한 자세한 정보는 IAM 사용 설명서에 있는 IAM의 정책 및 권한을 참조하세요.
- IAM 정책의 조건을 사용하여 액세스 추가 제한 – 정책에 조건을 추가하여 작업 및 리소스에 대한 액세스를 제한할 수 있습니다. 예를 들어, SSL을 사용하여 모든 요청을 전송해야 한다고 지정하는 정책 조건을 작성할 수 있습니다. AWS 서비스와 같은 특정를 통해 사용되는 경우 조건을 사용하여 서비스 작업에 대한 액세스 권한을 부여할 수도 있습니다 CloudFormation. 자세한 내용은 IAM 사용 설명서의 IAM JSON 정책 요소: 조건을 참조하세요.
- IAM Access Analyzer를 통해 IAM 정책을 확인하여 안전하고 기능적인 권한 보장 - IAM Access Analyzer에서는 IAM 정책 언어(JSON)와 모범 사례가 정책에서 준수되도록 새로운 및 기존 정책을 확인합니다. IAM Access Analyzer는 100개 이상의 정책 확인 항목과 실행 가능한 추천을 제공하여 안전하고 기능적인 정책을 작성하도록 돕습니다. 자세한 내용은 IAM 사용 설명서의 IAM Access Analyzer에서 정책 검증을 참조하세요.
- 다중 인증(MFA) 필요 -에서 IAM 사용자 또는 루트 사용자가 필요한 시나리오가 있는 경우 추가 보안을 위해 MFA를 AWS 계정켭니다. API 작업을 직접적으로 호출할 때 MFA가 필요하면 정책에 MFA 조건을 추가합니다. 자세한 내용은 IAM 사용 설명서의 MFA를 통한 보안 API 액세스를 참조하세요.

IAM의 모범 사례에 대한 자세한 내용은 IAM 사용 설명서의 IAM의 보안 모범 사례를 참조하세요.

AWS Batch 콘솔 사용

AWS Batch 콘솔에 액세스하려면 최소 권한 집합이 있어야 합니다. 이러한 권한은의 AWS Batch 리소스에 대한 세부 정보를 나열하고 볼 수 있도록 허용해야 합니다 AWS 계정. 최소 필수 권한보다 더 제한적인 ID 기반 정책을 생성하는 경우, 콘솔이 해당 정책에 연결된 엔티티(사용자 또는 역할)에 대해 의도대로 작동하지 않습니다.

AWS CLI 또는 AWS API만 호출하는 사용자에게 최소 콘솔 권한을 허용할 필요는 없습니다. 대신, 수행하려는 API 작업과 일치하는 작업에만 액세스할 수 있도록 합니다.

사용자와 역할이 AWS Batch 콘솔을 계속 사용할 수 있도록 하려면 `ConsoleAccess` 또는 `ReadOnly` AWS 관리형 AWS Batch 정책도 엔터티에 연결합니다. 자세한 내용은 IAM 사용 설명서의 사용자에게 권한 추가를 참조하세요.

사용자가 자신의 고유한 권한을 볼 수 있도록 허용

이 예제는 IAM 사용자가 자신의 사용자 ID에 연결된 인라인 및 관리형 정책을 볼 수 있도록 허용하는 정책을 생성하는 방법을 보여줍니다. 이 정책에는 콘솔에서 또는 AWS CLI 또는 AWS API를 사용하여 프로그래밍 방식으로이 작업을 완료할 수 있는 권한이 포함되어 있습니다.

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Sid": "ViewOwnUserInfo",
            "Effect": "Allow",
            "Action": [
                "iam:GetUserPolicy",
                "iam:ListGroupsForUser",
                "iam:ListAttachedUserPolicies",
                "iam:ListUserPolicies",
                "iam:GetUser"
            ],
            "Resource": ["arn:aws:iam::*:user/${aws:username}"]
        },
        {
            "Sid": "NavigateInConsole",
            "Effect": "Allow",
            "Action": [
                "iam:GetGroupPolicy",
                "iam:GetPolicyVersion",
                "iam:GetPolicy",
```

```
                "iam:ListAttachedGroupPolicies",
                "iam:ListGroupPolicies",
                "iam:ListPolicyVersions",
                "iam:ListPolicies",
                "iam:ListUsers"
            ],
            "Resource": "*"
        }
    ]
}
```

AWS 에 대한 관리형 정책 AWS Batch

AWS 관리형 정책을 사용하여 팀 및 프로비저닝된 AWS 리소스의 ID 액세스 관리를 간소화할 수 있습니다. AWS 관리형 정책은 다양한 일반 사용 사례를 다루고, AWS 계정에서 기본적으로 사용할 수 있으며, 사용자를 대신하여 유지 관리 및 업데이트됩니다. AWS 관리형 정책에서는 권한을 변경할 수 없습니다. 유연성이 더 필요한 경우 IAM 고객 관리형 정책을 생성할 수도 있습니다. 이렇게 하면 팀에 필요한 정확한 권한만 제공하여 프로비저닝된 리소스를 팀에 제공할 수 있습니다.

AWS 관리형 정책에 대한 자세한 내용은 IAM 사용 설명서의 AWS 관리형 정책을 참조하세요.

AWS 서비스는 사용자를 대신하여 AWS 관리형 정책을 유지 관리하고 업데이트합니다. 정기적으로 AWS 서비스는 AWS 관리형 정책에 추가 권한을 추가합니다. AWS 관리형 정책은 새 기능 시작 또는 작업을 사용할 수 있게 되면 업데이트될 가능성이 높습니다. 이 유형의 업데이트는 정책이 연결된 모든 자격 증명(사용자, 그룹 및 역할)에 적용됩니다. 하지만 권한을 제거하거나 기존 권한을 손상시키지는 않습니다.

또한는 여러 서비스에 걸쳐 있는 직무에 대한 관리형 정책을 AWS 지원합니다. 예를 들어 관리`ReadOnlyAccess` AWS 형 정책은 모든 AWS 서비스 및 리소스에 대한 읽기 전용 액세스를 제공합니다. 서비스가 새 기능을 시작하면는 새 작업 및 리소스에 대한 읽기 전용 권한을 AWS 추가합니다. 직무 정책의 목록과 설명은 IAM 사용 설명서의 직무에 관한AWS 관리형 정책을 참조하세요.

AWS 관리형 정책: BatchServiceRolePolicy

BatchServiceRolePolicy 관리형 IAM 정책은 `AWSServiceRoleForBatch` 서비스 연결 역할에서 사용됩니다. 이렇게 하면 AWS Batch 가 사용자를 대신하여 작업을 수행할 수 있습니다. IAM 엔터티에 이

정책을 연결할 수 없습니다. 자세한 내용은 에 대한 서비스 연결 역할 사용 AWS Batch 단원을 참조하십시오.

이 정책은가 특정 리소스에서 다음 작업을 완료 AWS Batch 하도록 허용합니다.

- `autoscaling` - AWS Batch 가 Amazon EC2 Auto Scaling 리소스를 생성하고 관리할 수 있도록 허용합니다.는 대부분의 컴퓨팅 환경에 대해 Amazon EC2 Auto Scaling 그룹을 AWS Batch 생성하고 관리합니다.
- `ec2` - AWS Batch 가 Amazon EC2 인스턴스의 수명 주기를 제어하고 시작 템플릿 및 태그를 생성 및 관리할 수 있도록 허용합니다.는 일부 EC2 스팟 컴퓨팅 환경에 대한 EC2 스팟 플릿 요청을 AWS Batch 생성하고 관리합니다.
- `ecs` -가 작업 실행 AWS Batch 을 위한 Amazon ECS 클러스터, 작업 정의 및 작업을 생성하고 관리할 수 있도록 허용합니다.
- `eks` -가 검증 AWS Batch 을 위해 Amazon EKS 클러스터 리소스를 설명할 수 있도록 허용합니다.
- `iam` - 소유자가 제공한 역할을 검증하고 Amazon EC2, Amazon EC2 Auto Scaling 및 Amazon ECS에 전달할 수 AWS Batch 있습니다.
- `logs` - AWS Batch 가 AWS Batch 작업에 대한 로그 그룹 및 로그 스트림을 생성하고 관리할 수 있도록 허용합니다.

정책의 JSON을 보려면 AWS 관리형 정책 참조 안내서의 BatchServiceRolePolicy를 참조하세요.

AWS 관리형 정책: AWSBatchServiceRolePolicyForSageMaker

`AWSServiceRoleForAWSBatchWithSagemaker`는 AWS Batch 가 사용자를 대신하여 작업을 수행할 수 있도록 허용합니다. IAM 엔터티에 이 정책을 연결할 수 없습니다. 자세한 내용은 에 대한 서비스 연결 역할 사용 AWS Batch 단원을 참조하십시오.

이 정책은가 특정 리소스에서 다음 작업을 완료 AWS Batch 하도록 허용합니다.

- `sagemaker` - AWS Batch 가 SageMaker AI 훈련 작업 및 기타 SageMaker AI 리소스를 관리할 수 있도록 허용합니다.
- `iam:PassRole` - AWS Batch 가 작업 실행을 위해 고객 정의 실행 역할을 SageMaker AI에 전달할 수 있도록 허용합니다. 리소스 제약 조건은 역할을 SageMaker AI 서비스에 전달하도록 허용합니다.

정책에 대한 JSON을 보려면 AWS 관리형 정책 참조 안내서의 AWSBatchServiceRolePolicyForSageMaker를 참조하세요.

AWS 관리형 정책: AWSBatchServiceRole 정책

AWSBatchServiceRole이라는 역할 권한 정책은 AWS Batch 가 특정 리소스에 대해 다음 작업을 수행하도록 허용합니다.

AWSBatchServiceRole 관리형 IAM 정책은 AWSBatchServiceRole이라는 역할에서 자주 사용되며 다음 권한을 포함합니다. 최소 권한 부여에 대한 표준 보안 권고 사항에 따라 AWSBatchServiceRole 관리형 정책을 가이드로 사용할 수 있습니다. 자신의 사용 사례에서 관리형 정책에서 부여된 권한이 필요하지 않은 경우 사용자 지정 정책을 생성하고 필요한 권한만 추가합니다. 이 AWS Batch 관리형 정책 및 역할은 대부분의 컴퓨팅 환경 유형에서 사용할 수 있지만 오류가 발생하기 쉽고 범위가 더 잘 지정되며 관리형 환경이 개선되기 때문에 서비스 연결 역할 사용이 선호됩니다.

- `autoscaling` - AWS Batch 가 Amazon EC2 Auto Scaling 리소스를 생성하고 관리할 수 있도록 허용합니다.는 대부분의 컴퓨팅 환경에 대해 Amazon EC2 Auto Scaling 그룹을 AWS Batch 생성하고 관리합니다.
- `ec2` - AWS Batch 가 Amazon EC2 인스턴스의 수명 주기를 관리하고 시작 템플릿 및 태그를 생성 및 관리할 수 있도록 허용합니다.는 일부 EC2 스팟 컴퓨팅 환경에 대한 EC2 스팟 플릿 요청을 AWS Batch 생성하고 관리합니다.
- `ecs` -가 작업 실행 AWS Batch 을 위한 Amazon ECS 클러스터, 작업 정의 및 작업을 생성하고 관리할 수 있도록 허용합니다.
- `iam` - 소유자가 제공한 역할을 검증하고 Amazon EC2, Amazon EC2 Auto Scaling 및 Amazon ECS에 전달할 수 AWS Batch 있습니다.
- `logs` - AWS Batch 가 AWS Batch 작업에 대한 로그 그룹 및 로그 스트림을 생성하고 관리할 수 있도록 허용합니다.

정책에 대한 JSON을 보려면 AWS 관리형 정책 참조 안내서의 AWSBatchServiceRole을 참조하세요.

AWS 관리형 정책: AWSBatchFullAccess

AWSBatchFullAccess 정책은 AWS Batch 작업에 AWS Batch 리소스에 대한 모든 액세스 권한을 부여합니다. 또한 Amazon EC2, Amazon ECS, Amazon EKS, CloudWatch 및 IAM 서비스에 대한 설명 및 목록 작업 액세스 권한을 부여합니다. 이는 사용자 또는 역할인 IAM 자격 증명이 사용자를 대신하여 생성된 AWS Batch 관리형 리소스를 볼 수 있도록 하기 위한 것입니다. 마지막으로, 이 정책은 선택된 IAM 역할을 해당 서비스에 전달할 수도 있도록 허용합니다.

AWSBatchFullAccess를 IAM 엔터티에 연결할 수 있습니다. AWS Batch 는가 사용자를 대신하여 작업을 AWS Batch 수행하도록 허용하는 서비스 역할에도이 정책을 연결합니다.

정책에 대한 JSON을 보려면 AWS 관리형 정책 참조 안내서의 AWSBatchFullAccess를 참조하세요.

AWS Batch AWS 관리형 정책에 대한 업데이트

이 서비스가 이러한 변경 사항을 추적하기 시작한 AWS Batch 이후부터의 AWS 관리형 정책 업데이트에 대한 세부 정보를 봅니다. 이 페이지의 변경 사항에 대한 자동 알림을 받으려면 AWS Batch 문서 기록 페이지에서 RSS 피드를 구독하세요.

변경	설명	Date
AWSBatchServiceRolePolicyForSageMaker 정책 업데이트	가 SageMaker AI 훈련 작업을 삭제할 수 AWS Batch 있는 `sagemaker:DeleteTrainingJob` 권한을 추가하도록 업데이트되었습니다.	2026년 4월 16일
AWSBatchServiceRolePolicyForSageMaker 정책 추가	가 사용자를 대신하여 SageMaker AI를 AWS 관리할 수 있도록 허용하는 AWSBatchServiceRolePolicyForSageMaker 서비스 연결 역할에 대한 새로운 AWS Batch 관리형 정책이 추가되었습니다. AWSBatchServiceRolePolicyForSageMaker SageMaker	2025년 7월 31일
BatchServiceRolePolicy 정책 업데이트	스팟 플릿 요청 기록 및 Amazon EC2 Auto Scaling 활동 설명을 위한 지원을 추가하도록 업데이트되었습니다.	2023년 12월 5일
AWSBatchServiceRole 정책 추가	문 IDs 추가`ec2:DescribeSpotFleetRequestHistory` 하고 및에 AWS Batch 권한을 부여하도록 업데이트되었습니다`autoscaling:DescribeScalingActivities` .	2023년 12월 5일
BatchServiceRolePolicy 정책 업데이트	Amazon EKS 클러스터 설명에 대한 지원을 추가하도록 업데이트되었습니다.	2022년 10월 20일

변경	설명	Date
AWSBatchFullAccess 정책 업데이트	Amazon EKS 클러스터 나열 및 설명에 대한 지원을 추가하도록 업데이트되었습니다.	2022년 10월 20일
BatchServiceRolePolicy 정책 업데이트	AWS Resource Groups에서 관리하는 Amazon EC2 용량 예약 그룹에 대한 지원을 추가하도록 업데이트되었습니다. 자세한 내용은 Amazon EC2 사용 설명서의 용량 예약 그룹 작업을 참조하세요.	2022년 5월 18일
BatchServiceRolePolicy 및 AWSBatchServiceRole 정책 업데이트	비정상 인스턴스가 교체되도록 Amazon EC2의 AWS Batch 관리형 인스턴스 상태를 설명하는 지원을 추가하도록 업데이트되었습니다.	2021년 12월 6일
BatchServiceRolePolicy 정책 업데이트	Amazon EC2의 배치 그룹, 용량 예약, 탄력적 GPU 및 Elastic Inference 리소스에 대한 지원을 추가하도록 업데이트되었습니다.	2021년 3월 26일
BatchServiceRolePolicy 정책 추가	AWSServiceRoleForBatch 서비스 연결 역할에 대한 BatchServiceRolePolicy 관리형 정책을 사용하면 AWS Batch에서 관리하는 서비스 연결 역할을 사용할 수 있습니다. 이 정책을 사용하면 컴퓨팅 환경에서 사용하기 위한 자체 역할을 유지 관리할 필요가 없습니다.	2021년 3월 10일
AWSBatchFullAccess - 서비스 연결 역할을 추가할 수 있는 권한 추가	AWSServiceRoleForBatch 서비스 연결 역할이 계정에 추가될 수 있도록 IAM 권한을 추가합니다.	2021년 3월 10일
AWS Batch 에서 변경 내용 추적 시작	AWS Batch 가 AWS 관리형 정책에 대한 변경 내용 추적을 시작했습니다.	2021년 3월 10일

AWS Batch IAM 정책, 역할 및 권한

기본적으로 사용자는 AWS Batch 리소스를 생성 또는 수정하거나 AWS Batch API, AWS Batch 콘솔 또는를 사용하여 작업을 수행할 수 있는 권한이 없습니다 AWS CLI. 사용자가 이러한 작업을 수행하도록 허용하려면 특정 리소스 및 API 작업을 위한 사용자 권한을 부여하는 IAM 정책을 생성합니다. 그런 다음, 해당 권한이 필요한 사용자 또는 그룹에 이러한 정책을 연결합니다.

사용자 또는 그룹에 정책을 연결하면 해당 정책은 지정된 리소스에서 지정된 태스크를 수행할 권한을 사용자에게 허용하거나 거부합니다. 자세한 내용은 IAM 사용 설명서의 권한 및 정책을 참조하세요. 사용자 지정 IAM 정책 관리 및 생성에 대한 자세한 내용은 IAM 정책 관리 섹션을 참조하세요.

AWS Batch 는 사용자를 대신하여 다른를 호출 AWS 서비스 합니다. 따라서 AWS Batch 는 자격 증명을 사용하여 인증해야 합니다. 보다 구체적으로,는 이러한 권한을 제공하는 IAM 역할 및 정책을 생성하여 AWS Batch 인증합니다. 그런 다음, 컴퓨팅 환경을 생성할 때 역할을 컴퓨팅 환경과 연결합니다. 자세한 내용은 IAM 사용 설명서의 Amazon ECS 인스턴스 역할, IAM 역할, 서비스 연결 역할 사용 및 AWS 서비스에 대한 권한 위임을 위한 역할 생성을 참조하세요. https://docs.aws.amazon.com/IAM/latest/UserGuide/roles-toplevel.html

주제

- IAM 정책 구조
- 리소스:에 대한 정책 예제 AWS Batch
- Resource: AWS Batch managed 정책

IAM 정책 구조

다음 항목에서는 IAM 정책의 구조에 대해 설명합니다.

주제

- 정책 구문
- 에 대한 API 작업 AWS Batch
- 에 대한 Amazon 리소스 이름 AWS Batch
- 사용자에게 필요한 권한이 있는지 확인

정책 구문

IAM 정책은 하나 이상의 문으로 구성된 JSON 문서입니다. 각 명령문의 구조는 다음과 같습니다.

```
{
  "Statement":[{
    "Effect":"effect",
    "Action":"action",
    "Resource":"arn",
    "Condition":{
      "condition":{
    "key":"value"
    }
      }
    }
  ]
}
```

명령문을 구성하는 네 가지 기본 요소가 있습니다.

- 효과(Effect): 효과(effect)는 `Allow` 또는 `Deny`일 수 있습니다. 기본적으로 사용자에게는 리소스 및 API 작업을 사용할 권한이 없으므로 모든 요청이 거부됩니다. 따라서 모든 요청을 거부합니다. 명시적 허용은 기본 설정을 무시합니다. 명시적 거부는 모든 허용을 무시합니다.
- 작업: 작업(action)은 권한을 부여하거나 거부할 특정 API 작업입니다. 작업(action)을 지정하는 방법에 대한 지침은 에 대한 API 작업 AWS Batch 섹션을 참조하세요.
- 리소스: 작업의 영향을 받는 리소스입니다. 일부 AWS Batch API 작업의 경우 작업이 생성하거나 수정할 수 있는 리소스를 정책에 구체적으로 포함할 수 있습니다. 설명문에서 리소스를 지정하려면 Amazon 리소스 이름(ARN)을 사용합니다. 자세한 내용은 AWS Batch API 작업에 지원되는 리소스 수준 권한 및 에 대한 Amazon 리소스 이름 AWS Batch 섹션을 참조하세요. AWS Batch API 작업이 현재 리소스 수준 권한을 지원하지 않는 경우 와일드카드(*)를 포함하여 모든 리소스가 작업의 영향을 받을 수 있도록 지정합니다.
- 조건(Condition): 조건(Condition)은 선택 사항입니다. 정책이 적용되는 시점을 제어하는 데 사용할 수 있습니다.

예제 IAM 정책 설명에 대한 자세한 내용은 섹션을 AWS Batch참조하세요리소스:에 대한 정책 예제 AWS Batch.

에 대한 API 작업 AWS Batch

IAM 정책 설명에는 IAM을 지원하는 모든 서비스의 모든 API 작업을 지정할 수 있습니다. 에는 API 작업의 이름과 함께 접두사를 AWS Batch사용합니다`batch:`(예: `batch:SubmitJob` 및 `batch:CreateComputeEnvironment`).

단일 문에서 여러 작업을 지정하려면 각 작업을 쉼표로 구분합니다.

```
"Action": ["batch:action1", "batch:action2"]
```

와일드카드(*)를 포함하여 여러 작업을 지정할 수도 있습니다. 예를 들어 이름이 'Describe'로 시작되는 모든 작업을 지정할 수 있습니다.

```
"Action": "batch:Describe*"
```

모든 AWS Batch API 작업을 지정하려면 와일드카드(*)를 포함합니다.

```
"Action": "batch:*"
```

AWS Batch 작업 목록은 AWS Batch API 참조의 작업을 참조하세요.

에 대한 Amazon 리소스 이름 AWS Batch

각 IAM 정책 설명은 Amazon 리소스 이름(ARN)을 사용하여 지정한 리소스에 적용됩니다.

Amazon 리소스 이름(ARN)의 일반 구문은 다음과 같습니다.

```
arn:aws:[service]:[region]:[account]:resourceType/resourcePath
```

서비스

서비스(예: `batch`)입니다.

리전

리소스 AWS 리전 의 입니다(예: `us-east-2`).

account

하이픈이 없는 AWS 계정 ID입니다(예: `123456789012`).

resourceType

리소스의 유형(예: `compute-environment`)입니다.

resourcePath

리소스를 식별하는 경로입니다. 경로에 와일드카드(*)를 사용할 수 있습니다.

AWS Batch API 작업은 현재 여러 API 작업에 대한 리소스 수준 권한을 지원합니다. 자세한 내용은 AWS Batch API 작업에 지원되는 리소스 수준 권한 단원을 참조하십시오. 모든 리소스를 지정해야 하거나 특정 API 작업이 ARN을 지원하지 않는 경우 다음과 같이 `Resource` 요소에 와일드카드(*)를 포함합니다.

```
"Resource": "*"
```

사용자에게 필요한 권한이 있는지 확인

IAM 정책을 프로덕션에 적용하기 전에, 해당 정책이 사용자에게 필요한 특정 API 작업 및 리소스를 사용할 권한을 부여하는지 확인하는 것이 좋습니다.

먼저, 테스트용으로 사용자를 생성하고 정책을 테스트 사용자에게 연결합니다. 그런 다음 테스트 사용자 자격으로 요청을 수행합니다. 콘솔 또는 AWS CLI에서 테스트 요청을 수행할 수 있습니다.

> **Note**
>
> IAM 정책 시뮬레이터로도 정책을 테스트할 수 있습니다. 정책 시뮬레이터에 대한 자세한 내용은 IAM 사용 설명서의 IAM 정책 시뮬레이터 작업 섹션을 참조하세요.

정책이 사용자에게 예상한 권한을 부여하지 않거나 과도한 권한을 부여하는 경우, 필요에 따라 정책을 조정할 수 있습니다. 원하는 결과를 얻을 때까지 다시 테스트합니다.

> **Important**
>
> 변경된 정책이 전파되어 효력을 발휘하려면 몇 분이 걸릴 수 있습니다. 따라서 정책을 업데이트한 경우 최소 5분간 기다린 후에 테스트하는 것이 좋습니다.

요청 시 권한 부여 확인에 실패하면 진단 정보가 포함된 인코딩 메시지가 반환됩니다. `DecodeAuthorizationMessage` 작업을 사용하여 메시지를 디코딩할 수 있습니다. 자세한 내용은 AWS Security Token Service API 레퍼런스의 DecodeAuthorizationMessage와 AWS CLI 명령 참조의 decode-authorization-message를 참조하세요.

리소스:에 대한 정책 예제 AWS Batch

사용자는 특정 IAM 정책을 생성하여 계정의 사용자가 액세스할 수 있는 호출 및 리소스를 제한합니다. 그런 다음 이러한 정책을 사용자에 연결할 수 있습니다.

사용자 또는 사용자 그룹에 정책을 연결하면 사용자의 특정 리소스에서 지정된 태스크를 수행할 권한이 허용되거나 거부됩니다. 자세한 내용은 IAM 사용 설명서의 권한 및 정책을 참조하세요. 사용자 지정 IAM 정책을 관리하고 생성하는 방법에 대한 지침은 IAM 정책 관리를 참조하세요.

다음 예제는 사용자가 갖는 AWS Batch권한을 제어하는 데 사용할 수 있는 정책 명령문을 보여줍니다.

예제

- 리소스:에 대한 읽기 전용 액세스 AWS Batch
- 리소스: 작업 제출 시 POSIX 사용자, Docker 이미지, 권한 수준 및 역할로 제한
- 리소스: 작업 제출 시 작업 정의 접두사로 제한
- 리소스: 작업 대기열로 제한
- 모든 조건이 문자열과 일치하는 경우 작업 거부
- 리소스: 문자열과 일치하는 조건 키가 있는 경우 작업 거부
- 리소스: batch:ShareIdentifier 조건 키 사용
- 를 사용하여 SageMaker AI 리소스 관리 AWS Batch
- 리소스: 작업 정의 및 작업 대기열의 리소스 태그로 작업 제출 제한

리소스:에 대한 읽기 전용 액세스 AWS Batch

다음 정책은 사용자에게 `Describe` 및 로 시작하는 이름으로 모든 AWS Batch API 작업을 사용할 수 있는 권한을 부여합니다`List`.

다른 명령문으로 해당 권한을 부여하지 않으면 사용자는 리소스에 대한 작업 수행 권한을 가지지 않습니다. 기본적으로 사용자의 API 작업 사용 권한은 거부됩니다.

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "batch:Describe*",
                "batch:List*",
                "batch:Get*"
```

```
            ],
            "Resource": "*"
        }
    ]
}
```

리소스: 작업 제출 시 POSIX 사용자, Docker 이미지, 권한 수준 및 역할로 제한

다음 정책은 POSIX 사용자가 고유의 작업 제한 정의 세트를 관리할 수 있도록 허용합니다.

첫 번째 및 두 번째 문을 사용하여 이름에 *JobDefA_*라는 접두사가 있는 모든 작업 정의를 등록하고 등록 취소합니다.

또한 첫 번째 설명문은 조건 컨텍스트 키를 사용하여 작업 정의의 `containerProperties` 내에 있는 POSIX 사용자, 권한 상태, 컨테이너 이미지 값을 제한합니다. 자세한 내용은 AWS Batch API 참조의 RegisterJobDefinition을 참조하세요. 이 예시에서는 POSIX 사용자가 `nobody`로 설정된 경우에만 작업 정의를 등록할 수 있습니다. 권한이 있는 플래그는 `false`로 설정됩니다. 마지막으로, 이 이미지는 Amazon ECR 리포지토리에서 `myImage`로 설정되어 있습니다.

> **Important**
>
> 도커는 `user` 파라미터를 컨테이너 이미지 내에 있는 해당 사용자 `uid`로 확인합니다. 대부분의 경우 이러한 사례는 컨테이너 이미지 내의 `/etc/passwd` 파일에서 찾을 수 있습니다. 이러한 이름 확인은 작업 정의 및 관련 IAM 정책에 직접 `uid` 값을 사용하면 생기지 않습니다. AWS Batch API 작업 및 `batch:User` IAM 조건 키는 숫자 값을 지원합니다.

세 번째 설명문을 사용하여 작업 정의에서 특정 역할만 하도록 제한합니다.

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "batch:RegisterJobDefinition"
            ],
```

```
            "Resource": [
                "arn:aws:batch:us-east-2:999999999999:job-definition/JobDefA_*"
            ],
            "Condition": {
                "StringEquals": {
                    "batch:User": [
                        "nobody"
                    ],
                    "batch:Image": [
                        "999999999999.dkr.ecr.us-east-2.amazonaws.com/myImage"
                    ]
                },
                "Bool": {
                    "batch:Privileged": "false"
                }
            }
        },
        {
            "Effect": "Allow",
            "Action": [
                "batch:DeregisterJobDefinition"
            ],
            "Resource": [
                "arn:aws:batch:us-east-2:999999999999:job-definition/JobDefA_*"
            ]
        },
        {
            "Effect": "Allow",
            "Action": [
                "iam:PassRole"
            ],
            "Resource": [
                "arn:aws:iam::999999999999:role/MyBatchJobRole"
            ]
        }
    ]
}
```

리소스: 작업 제출 시 작업 정의 접두사로 제한

다음 정책을 사용하여 *JobDefA*로 시작하는 작업 정의 이름이 있는 작업 만 모든 작업 대기열에 제출하도록 합니다.

> **Important**
>
> 작업 제출을 위한 리소스 수준 액세스의 범위를 지정할 때 작업 대기열 및 작업 정의 리소스 유형을 모두 제공해야 합니다.

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "batch:SubmitJob"
            ],
            "Resource": [
                "arn:aws:batch:us-east-2:111122223333:job-definition/JobDefA_*",
                "arn:aws:batch:us-east-2:111122223333:job-queue/*"
            ]
        }
    ]
}
```

작업 정의 개정에 대한 고려

> **Important**
>
> 요청 ARN에 개정(예: `job-definition/my-job-def`)이 포함`SubmitJob`되므로 개정 번호 또는 와일드카드 없이 작업 정의 이름만 참조하는 정책(예:)은 요청과 일치하지 않습니다`job-definition/my-job-def:1`. 와일드카드를 사용하여 모든 개정을 일치시킵니다.

다음 예제에서는 `SubmitJob` 작업에 대한 리소스 ARNs에서 와일드카드 및 개정 번호를 사용하는 방법을 보여줍니다.

예: 특정 작업 정의 개정 허용

다음 정책은 지정된 작업 정의의 개정 1만 사용하는 작업 제출을 허용합니다.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": "batch:SubmitJob",
      "Resource": [
        "arn:aws:batch:us-east-1:111122223333:job-definition/my-job-def:1",
        "arn:aws:batch:us-east-1:111122223333:job-queue/*"
      ]
    }
  ]
}
```

예: 작업 정의의 모든 개정 허용

다음 정책은 지정된 작업 정의의 모든 개정을 사용하여 작업 제출을 허용합니다. `:*` 패턴은 모든 개정 번호와 일치합니다.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": "batch:SubmitJob",
      "Resource": [
        "arn:aws:batch:us-east-1:111122223333:job-definition/my-job-def:*",
        "arn:aws:batch:us-east-1:111122223333:job-queue/*"
      ]
    }
  ]
}
```

리소스: 작업 대기열로 제한

다음 정책을 사용하여 queue1이라는 이름의 특정 작업 대기열에 모든 작업 정의 작업을 제출하게 합니다.

> **Important**
>
> 작업 제출을 위한 리소스 수준 액세스의 범위를 지정할 때 작업 대기열 및 작업 정의 리소스 유형을 모두 제공해야 합니다.

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "batch:SubmitJob"
            ],
            "Resource": [
                "arn:aws:batch:us-east-2:888888888888:job-definition/*",
                "arn:aws:batch:us-east-2:888888888888:job-queue/queue1"
            ]
        }
    ]
}
```

모든 조건이 문자열과 일치하는 경우 작업 거부

다음 정책은 batch:Image (컨테이너 이미지 ID) 조건 키가 "*string1*"이고 batch:LogDriver (컨테이너 로그 드라이버) 조건 키가 "*string2*"인 경우 RegisterJobDefinition API 작업에 대한 액세스를 거부합니다.는 각 컨테이너의 조건 키를 AWS Batch 평가합니다. 다중 노드 병렬 작업과 같이 작업이 여러 컨테이너에 걸쳐 있는 경우 컨테이너의 구성이 다를 수 있습니다. 한 명령문에서 여러 조건 키를 평가하는 경우 AND 로직을 사용하여 병합됩니다. 따라서 여러 조건 키 중 하나라도 컨테이너와 일치하지 않는 경우 해당 컨테이너에는 Deny 효과가 적용되지 않습니다. 오히려 같은 작업에 있는 다른 컨테이너가 거부될 수 있습니다.

에 대한 조건 키 목록은 서비스 승인 AWS Batch참조의에 대한 조건 키를 AWS Batch 참조하세요. batch:ShareIdentifier을 제외한 모든 batch 조건 키는 이 방법으로 사용될 수 있습니다. batch:ShareIdentifier 조건 키는 작업 정의가 아니라 작업에 대해 정의됩니다.

JSON

```
{
  "Version":"2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "batch:RegisterJobDefinition"
      ],
      "Resource": [
        "*"
      ]
    },
    {
      "Effect": "Deny",
      "Action": "batch:RegisterJobDefinition",
      "Resource": "*",
      "Condition": {
        "StringEquals": {
          "batch:Image": "string1",
          "batch:LogDriver": "string2"
        }
      }
    }
  ]
}
```

리소스: 문자열과 일치하는 조건 키가 있는 경우 작업 거부

다음 정책은 batch:Image(컨테이너 이미지 ID) 조건 키가 '*string1*'이거나 batch:LogDriver(컨테이너 로그 드라이버) 조건 키가 '*string2*'인 경우 RegisterJobDefinition API 작업에 대한 액세스를 거부합니다. 다중 노드 병렬 작업과 같이 작업이 여러 컨테이너에 걸쳐 있는 경우 컨테이너의 구성이 다를 수 있습니다. 한 명령문에서 여러 조건 키를 평가하는 경우 AND 로직을 사용하여 병합됩니다. 따라서 여러 조건 키 중 하나라도 컨테이너와 일치하지 않는 경우 해당 컨테이너에는 Deny 효과가 적용되지 않습니다. 오히려 같은 작업에 있는 다른 컨테이너가 거부될 수 있습니다.

에 대한 조건 키 목록은 서비스 승인 AWS Batch참조의에 대한 조건 키를 AWS Batch 참조하세요. batch:ShareIdentifier을 제외한 모든 batch 조건 키는 이 방법으로 사용될 수 있습니다. (batch:ShareIdentifier 조건 키는 작업 정의가 아니라 작업에 대해 정의됩니다.)

JSON

```
{
  "Version":"2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "batch:RegisterJobDefinition"
      ],
      "Resource": [
        "*"
      ]
    },
    {
      "Effect": "Deny",
      "Action": [
        "batch:RegisterJobDefinition"
      ],
      "Resource": [
        "*"
      ],
      "Condition": {
        "StringEquals": {
          "batch:Image": [
            "string1"
          ]
        }
      }
    },
    {
      "Effect": "Deny",
      "Action": [
        "batch:RegisterJobDefinition"
      ],
      "Resource": [
        "*"
      ],
      "Condition": {
        "StringEquals": {
          "batch:LogDriver": [
            "string2"
          ]
```

```
        }
      }
    }
  ]
}
```

리소스: **batch:ShareIdentifier** 조건 키 사용

다음 정책을 사용하여 `jobDefA` 작업 정의를 사용하는 작업을 `lowCpu` 공유 식별자와 함께 `jobqueue1` 작업 대기열에 제출하세요.

JSON

```
{
  "Version":"2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "batch:SubmitJob"
      ],
      "Resource": [
        "arn:aws:batch:us-east-2:555555555555:job-definition/JobDefA",
        "arn:aws:batch:us-east-2:555555555555:job-queue/jobqueue1"
      ],
      "Condition": {
        "StringEquals": {
          "batch:ShareIdentifier": [
            "lowCpu"
          ]
        }
      }
    }
  ]
}
```

를 사용하여 SageMaker AI 리소스 관리 AWS Batch

이 정책은 AWS Batch 가 SageMaker AI 리소스를 관리할 수 있도록 허용합니다.

JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "batch:*"
            ],
            "Resource": "*"
        },
        {
            "Effect": "Allow",
            "Action": [
                "iam:CreateServiceLinkedRole"
            ],
            "Resource": "arn:aws:iam::*:role/
*AWSServiceRoleForAWSBatchWithSagemaker",
            "Condition": {
                 "StringEquals": {
                     "iam:AWSServiceName": "sagemaker-
queuing.batch.amazonaws.com"
                 }
             }
         },
         {
             "Effect": "Allow",
             "Action": "iam:PassRole",
             "Resource": "*",
             "Condition": {
                 "StringEquals": {
                     "iam:PassedToService": [
                         "sagemaker.amazonaws.com"
                      ]
                  }
              }
          }
    ]
}
```

리소스: 작업 정의 및 작업 대기열의 리소스 태그로 작업 제출 제한

작업 대기열에 태그가 `Environment=dev` 있고 작업 정의에 태그가 있는 경우에만 다음 정책을 사용하여 작업을 제출합니다`Project=calc`. 이 정책은 작업 제출 중에 리소스 태그를 사용하여 AWS Batch 리소스에 대한 액세스를 제어하는 방법을 보여줍니다.

> **Important**
>
> 작업 정의 리소스 태그를 평가하는 정책이 있는 작업을 제출할 때는 작업 정의 개정 형식()을 사용하여 작업을 제출해야 합니다`job-definition:revision`. 개정을 지정하지 않고 작업을 제출하면 작업 정의 태그가 평가되지 않으므로 의도한 액세스 제어를 우회할 수 있습니다. 리소스 ARN의 `*:*` 패턴은 제출에 개정이 포함되어야 함을 적용하여 태그 정책이 항상 효과적으로 적용되도록 합니다.

이 정책은 서로 다른 리소스 유형에 서로 다른 태그 조건을 적용하기 때문에 두 개의 개별 문을 사용합니다. 작업 제출을 위한 리소스 수준 액세스의 범위를 지정할 때 작업 대기열 및 작업 정의 리소스 유형을 모두 제공해야 합니다.

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": "batch:SubmitJob",
      "Resource": "arn:aws:batch:*:*:job-queue/*",
      "Condition": {
        "StringEquals": {
          "aws:ResourceTag/Environment": "dev"
        }
      }
    },
    {
      "Effect": "Allow",
      "Action": "batch:SubmitJob",
      "Resource": "arn:aws:batch:*:*:job-definition/*:*",
      "Condition": {
        "StringEquals": {
          "aws:ResourceTag/Project": "calc"
        }
      }
```

```
    }
  ]
}
```

Resource: AWS Batch managed 정책

AWS Batch 는 사용자에게 연결할 수 있는 관리형 정책을 제공합니다. 이 정책은 AWS Batch 리소스 및 API 작업을 사용할 수 있는 권한을 제공합니다. 이 정책은 직접 적용할 수도 있고, 사용자 고유의 정책을 생성하기 위한 시작 지점으로 사용할 수도 있습니다. 이러한 정책에 언급되는 각 API 작업에 대한 자세한 내용은 AWS Batch API 참조의 작업 섹션을 참조하세요.

AWSBatchFullAccess

이 정책은에 대한 전체 관리자 액세스를 허용합니다 AWS Batch.

정책에 대한 JSON을 보려면 AWS 관리형 정책 참조 안내서의 AWSBatchFullAccess를 참조하세요.

AWS Batch IAM 실행 역할

실행 역할은 Amazon ECS 컨테이너 에이전트 및 AWS Fargate 에이전트에게 사용자를 대신하여 AWS API를 호출할 수 있는 권한을 부여합니다.

> **Note**
>
> 실행 역할은 Amazon ECS 컨테이너 에이전트 버전 1.16.0 이상에서 지원됩니다.

태스크의 요구 사항에 따라 IAM 실행 역할이 필요합니다. 계정과 연결된 다른 용도 및 서비스에 사용할 여러 실행 역할이 있을 수 있습니다.

> **Note**
>
> Amazon ECS 인스턴스 역할에 대한 자세한 내용을 알아보려면 Amazon ECS 인스턴스 역할(을)를 참조하세요. 서비스 역할에 대한 자세한 내용은 AWS Batch 에서 IAM을 사용하는 방법(을)를 참조하세요.

Amazon ECS는 `AmazonECSTaskExecutionRolePolicy` 관리형 정책을 제공합니다. 이 정책은 위에서 설명한 일반 사용 사례에서 필요로 하는 권한을 포함하고 있습니다. 아래에 간단히 설명한 특수 사용 사례에서는 인라인 정책을 실행 역할에 추가해야 할 수도 있습니다.

JSON

```
{
  "Version":"2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "ecr:GetAuthorizationToken",
        "ecr:BatchCheckLayerAvailability",
        "ecr:GetDownloadUrlForLayer",
        "ecr:BatchGetImage",
        "logs:CreateLogStream",
        "logs:PutLogEvents"
      ],
      "Resource": "*"
    }
  ]
}
```

AWS Batch API 작업에 지원되는 리소스 수준 권한

리소스 수준 권한이라는 용어는 사용자가에서 작업을 수행할 수 있는 리소스를 지정하는 기능을 말합니다. AWS Batch 는 리소스 수준 권한을 부분적으로 지원합니다. 일부 AWS Batch 작업의 경우 사용자가 충족해야 하는 조건에 따라 해당 작업을 사용할 수 있는 시기를 제어할 수 있습니다. 또한 사용자가 사용할 수 있는 특정 리소스를 기반으로 제어할 수도 있습니다. 예를 들어 특정 작업 정의만 사용하여 특정 작업 대기열에만 작업을 제출할 수 있는 권한을 사용자에게 부여할 수 있습니다.

각 리소스 유형에 대한 ARNs 형식을 포함하여 AWS Batch에서 정의한 작업 및 리소스 유형에 대한 자세한 내용은 서비스 승인 참조의 AWS Batch에 대한 작업, 리소스 및 조건 키를 참조하세요.

자습서: IAM 실행 역할 생성

계정에 아직 IAM 실행 역할이 없는 경우 다음 단계를 사용하여 역할을 생성합니다.

1. IAM 콘솔(https://console.aws.amazon.com/iam/)을 엽니다.
2. 탐색 창에서 역할을 선택합니다.
3. 역할 생성을 선택합니다.
4. 신뢰할 수 있는 엔터티 유형에 AWS 서비스(을)를 선택합니다.

5. 서비스 또는 사용 사례에서 Elastic Container Service를 선택합니다. 그런 다음 Elastic Container Service 태스크를 다시 선택합니다.
6. 다음을 선택합니다.
7. 권한 정책(Permissions policies)에 대해서는 AmazonECSTaskExecutionRolePolicy를 검색하세요.
8. AmazonECSTaskExecutionRolePolicy 정책 왼쪽의 확인란을 선택하고 다음을 선택합니다.
9. 역할 이름에 `ecsTaskExecutionRole`(을)를 입력한 다음 역할 생성을 선택합니다.

자습서: IAM 실행 역할 확인

다음 절차를 사용하여 사용자 계정에 이미 IAM 실행 역할이 있는지 확인하고 필요할 경우 관리형 IAM 정책을 연결합니다.

1. IAM 콘솔(https://console.aws.amazon.com/iam/)을 엽니다.
2. 탐색 창에서 역할을 선택합니다.
3. 역할 목록에서 `ecsTaskExecutionRole`(을)를 검색합니다. 역할을 찾을 수 없는 경우 자습서: IAM 실행 역할 생성 섹션을 참조하세요. 역할을 찾을 경우 해당 역할을 선택하여 연결된 정책을 확인합니다.
4. 권한(Permissions) 탭에서 AmazonECSTaskExecutionRolePolicy 관리형 정책이 역할에 연결되었는지 검증합니다. 정책이 연결된 경우, 실행 역할이 적절히 구성된 것입니다. 그렇지 않다면 아래의 하위 단계에 따라 정책을 연결합니다.

   a. 권한 추가를 선택하고 정책 연결을 선택합니다.
   b. AmazonECSTaskExecutionRolePolicy를 검색합니다.
   c. AmazonECSTaskExecutionRolePolicy 정책 왼쪽의 확인란을 선택하고 정책 연결을 선택합니다.
5. 신뢰 관계를 선택합니다.
6. 신뢰 관계에 다음 정책이 포함되어 있는지 확인합니다. 신뢰 관계가 아래 정책과 일치하면 역할이 올바르게 구성된 것입니다. 신뢰 관계가 일치하지 않으면 신뢰 관계 편집을 선택하고 다음을 입력한 뒤 정책 업데이트를 선택합니다.

   JSON

```
{
  "Version":"2012-10-17",
```

```
  "Statement": [
    {
      "Sid": "",
      "Effect": "Allow",
      "Principal": {
        "Service": "ecs-tasks.amazonaws.com"
      },
      "Action": "sts:AssumeRole"
    }
  ]
}
```

에 대한 서비스 연결 역할 사용 AWS Batch

AWS Batch 는 AWS Identity and Access Management (IAM) 서비스 연결 역할을 사용합니다. 서비스 연결 역할은 직접 연결된 고유한 유형의 IAM 역할입니다 AWS Batch. 서비스 연결 역할은에서 사전 정의 AWS Batch 하며 서비스가 사용자를 대신하여 다른 AWS 서비스를 호출하는 데 필요한 모든 권한을 포함합니다.

AWS Batch 는 두 가지 서비스 연결 역할을 사용합니다.

- AWSServiceRoleForBatch - 컴퓨팅 환경을 포함한 AWS Batch 작업에 사용됩니다.
- AWSServiceRoleForAWSBatchWithSagemaker - SageMaker AI 워크로드 관리 및 대기열에 사용됩니다.

주제

- 에 역할 사용 AWS Batch
- SageMaker AI AWS Batch 에서에 대한 역할 사용

에 역할 사용 AWS Batch

AWS Batch 는 AWS Identity and Access Management (IAM) 서비스 연결 역할을 사용합니다. 서비스 연결 역할은 직접 연결된 고유한 유형의 IAM 역할입니다 AWS Batch. 서비스 연결 역할은에서 사전 정의 AWS Batch 하며 서비스가 사용자를 대신하여 다른 AWS 서비스를 호출하는 데 필요한 모든 권한을 포함합니다.

필요한 권한을 수동으로 추가할 필요가 없으므로 서비스 연결 역할을 더 AWS Batch 쉽게 설정할 수 있습니다.는 서비스 연결 역할의 권한을 AWS Batch 정의하며, 달리 정의되지 않은 한 만 해당 역할을 수

임 AWS Batch 할 수 있습니다. 정의된 권한에는 신뢰 정책과 권한 정책이 포함되며 이 권한 정책은 다른 IAM 엔터티에 연결할 수 없습니다.

> **Note**
>
> 다음 중 하나를 수행하여 AWS Batch 컴퓨팅 환경에 대한 서비스 역할을 지정합니다.
>
> - 서비스 역할에는 빈 문자열을 사용합니다. 이렇게 하면가 서비스 역할을 AWS Batch 생성할 수 있습니다.
> - `arn:aws:iam::account_number:role/aws-service-role/batch.amazonaws.com/AWSServiceRoleForBatch` 형식으로 서비스 역할을 지정합니다.
>
> 자세한 내용은 AWS Batch 사용 설명서부정확한 역할 이름 또는 ARN의 섹션을 참조하세요.

> **Important**
>
> 서비스 연결 역할 자동 생성은 `MANAGED` 컴퓨팅 환경에만 적용됩니다. `UNMANAGED` 컴퓨팅 환경의 경우를 명시적으로 지정해야 합니다`serviceRole`. `UNMANAGED` 컴퓨팅 환경의 역할을 생략하면 `ServiceRole is required` 오류가 발생합니다.

먼저 관련 리소스를 삭제한 후에만 서비스 연결 역할을 삭제할 수 있습니다. 이렇게 하면 AWS Batch 리소스에 대한 액세스 권한을 실수로 제거할 수 없기 때문에 리소스가 보호됩니다.

서비스 연결 역할을 지원하는 다른 서비스에 대한 자세한 내용은 AWS IAM으로 작업하는 서비스를 참조하고 서비스 연결 역할 열에서 예인 서비스를 찾습니다. 해당 서비스에 대한 서비스 연결 역할 설명서를 보려면 예 링크를 선택합니다.

에 대한 서비스 연결 역할 권한 AWS Batch

AWS Batch 는 AWSServiceRoleForBatch라는 서비스 연결 역할을 사용합니다. AWS Batch 는 사용자를 대신하여 AWS 리소스를 생성하고 관리할 수 있습니다.

AWSServiceRoleForBatch 서비스 연결 역할은 역할을 수임하기 위해 다음 서비스를 신뢰합니다.

- `batch.amazonaws.com`

BatchServiceRolePolicy라는 역할 권한 정책은가 지정된 리소스에서 다음 작업을 완료 AWS Batch 하도록 허용합니다.

- `autoscaling` - AWS Batch 가 Amazon EC2 Auto Scaling 리소스를 생성하고 관리할 수 있도록 허용합니다.는 대부분의 컴퓨팅 환경에 대해 Amazon EC2 Auto Scaling 그룹을 AWS Batch 생성하고 관리합니다.
- `ec2` - AWS Batch 가 Amazon EC2 인스턴스의 수명 주기를 제어하고 시작 템플릿 및 태그를 생성 및 관리할 수 있도록 허용합니다.는 일부 EC2 스팟 컴퓨팅 환경에 대한 EC2 스팟 플릿 요청을 AWS Batch 생성하고 관리합니다.
- `ecs` -가 작업 실행 AWS Batch 을 위한 Amazon ECS 클러스터, 작업 정의 및 작업을 생성하고 관리할 수 있도록 허용합니다.
- `eks` -가 검증 AWS Batch 을 위해 Amazon EKS 클러스터 리소스를 설명할 수 있도록 허용합니다.
- `iam` - 소유자가 제공한 역할을 검증하고 Amazon EC2, Amazon EC2 Auto Scaling 및 Amazon ECS에 전달할 수 AWS Batch 있습니다.
- `logs` - AWS Batch 가 AWS Batch 작업에 대한 로그 그룹 및 로그 스트림을 생성하고 관리할 수 있도록 허용합니다.

사용자, 그룹 또는 역할이 서비스 연결 역할을 생성, 편집 또는 삭제할 수 있도록 사용 권한을 구성해야 합니다. 자세한 내용은 IAM 사용 설명서의 서비스 연결 역할 권한을 참조하세요.

에 대한 서비스 연결 역할 생성 AWS Batch

서비스 연결 역할은 수동으로 생성할 필요가 없습니다. AWS Management Console, AWS CLI또는 AWS API에서 컴퓨팅 환경을 생성하면 AWS Batch 가 서비스 연결 역할을 생성합니다.

> **Note**
>
> 이 자동 서비스 연결 역할 생성은 `MANAGED` 컴퓨팅 환경에만 적용됩니다. `UNMANAGED` 컴퓨팅 환경의 경우를 호출할 `serviceRole` 때를 명시적으로 제공해야 합니다`CreateComputeEnvironment`.

> **Important**
>
> 이러한 서비스 연결 역할은 해당 역할이 지원하는 기능을 사용하는 다른 서비스에서 작업을 완료했을 경우 계정에 나타날 수 있습니다. AWS Batch 서비스 연결 역할을 지원하기 시작한 2021년 3월 10일 이전에 서비스를 사용 중이었다면에서 계정에 AWSServiceRoleForBatch 역

할을 AWS Batch 생성했습니다. 자세한 내용은 내에 표시되는 새 역할을 참조하세요 AWS 계정.

이 서비스 연결 역할을 삭제했다가 다시 생성해야 하는 경우 동일한 프로세스를 사용하여 계정에서 역할을 다시 생성할 수 있습니다. 컴퓨팅 환경을 생성하면가 서비스 연결 역할을 다시 AWS Batch 생성합니다.

에 대한 서비스 연결 역할 편집 AWS Batch

AWS Batch 에서는 AWSServiceRoleForBatch 서비스 연결 역할을 편집할 수 없습니다. 서비스 연결 역할을 생성한 후에는 다양한 개체가 역할을 참조할 수 있기 때문에 역할 이름을 변경할 수 없습니다. 하지만 IAM을 사용하여 역할의 설명을 편집할 수 있습니다. 자세한 내용은 IAM 사용 설명서의 서비스 연결 역할 편집을 참조하세요.

IAM 엔터티가 AWSServiceRoleForBatch 서비스 연결 역할의 설명을 편집할 수 있도록 하려면

권한 정책에 다음 설명을 추가합니다. IAM 엔터티가 서비스 연결 역할의 설명을 편집할 수 있도록 허용합니다.

```
{
    "Effect": "Allow",
    "Action": [
        "iam:UpdateRoleDescription"
    ],
    "Resource": "arn:aws:iam::*:role/aws-service-role/batch.amazonaws.com/
AWSServiceRoleForBatch",
    "Condition": {"StringLike": {"iam:AWSServiceName": "batch.amazonaws.com"}}
}
```

에 대한 서비스 연결 역할 삭제 AWS Batch

서비스 연결 역할이 필요한 기능 또는 서비스가 더 이상 필요 없는 경우에는 해당 역할을 삭제할 것을 권합니다. 그렇게 하면 적극적으로 모니터링하거나 유지 관리하지 않은 미사용 엔터티가 없습니다. 단, 서비스 연결 역할을 정리해야 수동으로 삭제할 수 있습니다.

IAM 엔터티가 AWSServiceRoleForBatch 서비스 연결 역할을 생성하도록 허용하는 방법

권한 정책에 다음 설명을 추가합니다. IAM 엔터티가 서비스 연결 역할을 삭제하도록 허용합니다.

```
{
```

```
    "Effect": "Allow",
    "Action": [
        "iam:DeleteServiceLinkedRole",
        "iam:GetServiceLinkedRoleDeletionStatus"
    ],
    "Resource": "arn:aws:iam::*:role/aws-service-role/batch.amazonaws.com/
AWSServiceRoleForBatch",
    "Condition": {"StringLike": {"iam:AWSServiceName": "batch.amazonaws.com"}}
}
```

서비스 연결 역할을 정리

IAM을 사용하여 서비스 연결 역할을 삭제하려면 먼저 역할에 활성 세션이 없는지 확인하고 단일 파티션의 모든 AWS 리전에서 역할을 사용하는 모든 AWS Batch 컴퓨팅 환경을 삭제해야 합니다.

서비스 연결 역할에 활성 세션이 있는지 확인하는 방법

1. IAM 콘솔(https://console.aws.amazon.com/iam/)을 엽니다.
2. 탐색 창에서 역할을 선택한 다음 확인란이 아닌 AWSServiceRoleForBatch 이름을 선택합니다.
3. 요약 페이지에서 액세스 고문을 선택하고 서비스 연결 역할의 최근 활동을 검토합니다.

> **Note**
>
> AWS Batch 가 AWSServiceRoleForBatch 역할을 사용하는지 여부를 모르는 경우 역할을 삭제해 볼 수 있습니다. 서비스에서 역할을 사용하는 경우에는 역할이 삭제되지 않습니다. 역할이 사용되고 있는 리전을 볼 수 있습니다. 역할이 사용 중인 경우에는 세션이 종료될 때까지 기다렸다가 역할을 삭제해야 합니다. 서비스 연결 역할에 대한 세션은 취소할 수 없습니다.

AWSServiceRoleForBatch 서비스 연결 역할에서 사용하는 AWS Batch 리소스를 제거하려면

AWSServiceRoleForBatch 역할을 삭제하려면 먼저 모든 AWS 리전에서 AWSServiceRoleForBatch 역할을 사용하는 모든 AWS Batch 컴퓨팅 환경을 삭제해야 합니다.

1. https://console.aws.amazon.com/batch/ AWS Batch 콘솔을 엽니다.
2. 탐색 모음에서 사용할 리전을 선택합니다.
3. 탐색 창에서 컴퓨팅 환경를 선택합니다.
4. 컴퓨팅 환경을 선택합니다.

5. 비활성화를 선택합니다. 상태가 비활성화됨(DISABLED)으로 변경될 때까지 기다립니다.
6. 컴퓨팅 환경을 선택합니다.
7. 삭제를 선택합니다. 컴퓨팅 환경 삭제를 선택하여 컴퓨팅 환경의 삭제를 확인합니다.
8. 모든 리전에서 서비스 연결 역할을 사용하는 모든 컴퓨팅 환경에 대해 1~7단계를 반복합니다.

IAM에서 서비스 연결 역할 삭제(콘솔)

IAM 콘솔을 사용하여 서비스 연결 역할을 삭제할 수 있습니다.

서비스 연결 역할을 삭제하는 방법(콘솔)

1. 에 로그인 AWS Management Console 하고 https://console.aws.amazon.com/iam/ IAM 콘솔을 엽니다.
2. IAM 콘솔의 탐색 창에서 역할을 선택합니다. 이름이나 행 자체가 아닌 AWSServiceRoleForBatch 옆의 확인란을 선택합니다.
3. 역할 삭제을 선택합니다.
4. 확인 대화 상자가 나타나면 서비스 마지막 액세스 데이터를 검토합니다. 이 데이터는 선택한 각 역할이 AWS 서비스(을)를 마지막으로 액세스한 일시를 보여줍니다. 이를 통해 역할이 현재 활동 중인지를 확인할 수 있습니다. 계속 진행하려면 예, 삭제합니다를 선택하여 삭제할 서비스 연결 역할을 제출합니다.
5. IAM 콘솔 알림을 보고 서비스 연결 역할 삭제 진행 상황을 모니터링합니다. IAM 서비스 연결 역할 삭제는 비동기이므로 삭제할 역할을 제출한 후에 삭제 태스크가 성공하거나 실패할 수 있습니다.

    - 태스크에 성공하면 목록에서 역할이 제거되고 성공 알림이 페이지 상단에 나타납니다.
    - 태스크에 실패할 경우 알림의 세부 정보 보기 또는 리소스 보기를 선택하면 삭제 실패 이유를 확인할 수 있습니다. 역할에서 서비스 리소스를 사용 중이어서 삭제에 실패한 경우에는 알림에 리소스 목록이 포함됩니다(서비스에서 해당 정보를 반환할 경우). 이후 리소스를 정리하고 삭제를 다시 제출할 수 있습니다.

    > **Note**
    >
    > 서비스에서 반환하는 정보에 따라 이 과정을 여러 번 반복해야 할 수 있습니다. 예를 들어, 서비스 연결 역할에서 6개의 리소스를 사용할 수 있으며, 서비스에서 이중 5개에 관한 정보를 반환할 수 있습니다. 5개 리소스를 정리하고 삭제할 역할을 다시 제출할 경우 삭제에 실패하고 서비스에서 나머지 1개의 리소스를 보고합니다. 서비스에서 리소스 전부를 반환하거나, 일부만 반환하거나, 리소스를 보고하지 않을 수 있습니다.

- 태스크에 실패했는데 알림에 리소스 목록이 포함되지 않을 경우에는 서비스에서 해당 정보를 반환하지 않을 수 있습니다. 해당 서비스의 리소스를 정리하는 방법은 IAM으로 작업하는AWS 서비스를 참조하세요. 표에서 서비스를 확인하고 예(Yes) 링크를 선택하여 해당 서비스의 서비스 연결 역할 문서를 확인합니다.

IAM에서 서비스 연결 역할 삭제(AWS CLI)

에서 IAM 명령을 사용하여 서비스 연결 역할을 AWS Command Line Interface 삭제할 수 있습니다.

서비스 연결 역할을 삭제하는 방법(CLI)

1. 서비스 연결 역할이 사용되지 않거나 연결된 리소스가 없는 경우에는 서비스 연결 역할을 삭제할 수 없으므로 삭제 요청을 제출해야 합니다. 이러한 조건이 충족되지 않으면 요청이 거부될 수 있습니다. 삭제 태스크 상태를 확인하려면 응답의 `deletion-task-id`(을)를 캡처해야 합니다. 다음 명령을 입력하여 서비스 연결 역할 삭제 요청을 제출합니다.

```
$ aws iam delete-service-linked-role --role-name AWSServiceRoleForBatch
```

2. 다음 명령을 사용하여 삭제 태스크의 상태를 확인합니다.

```
$ aws iam get-service-linked-role-deletion-status --deletion-task-id deletion-task-id
```

삭제 태스크는 `NOT_STARTED`, `IN_PROGRESS`, `SUCCEEDED` 또는 `FAILED` 상태일 수 있습니다. 삭제에 실패할 경우 문제를 해결할 수 있도록 실패 이유가 호출에 반환됩니다. 역할에서 서비스 리소스를 사용 중이어서 삭제에 실패한 경우에는 알림에 리소스 목록이 포함됩니다(서비스에서 해당 정보를 반환할 경우). 이후 리소스를 정리하고 삭제를 다시 제출할 수 있습니다.

> **Note**
>
> 서비스에서 반환하는 정보에 따라 이 과정을 여러 번 반복해야 할 수 있습니다. 예를 들어, 서비스 연결 역할에서 6개의 리소스를 사용할 수 있으며, 서비스에서 이중 5개에 관한 정보를 반환할 수 있습니다. 5개 리소스를 정리하고 삭제할 역할을 다시 제출할 경우 삭제에 실패하고 서비스에서 나머지 1개의 리소스를 보고합니다. 서비스에서 리소스 전부를 반환하거나, 일부만 반환할 수 있습니다. 또는 리소스를 보고하지 않을 수도 있습니다. 어떤 리소스도 보고하지 않는 서비스의 리소스를 정리하는 방법은 IAM으로 작업하는AWS 서비

스 섹션을 참조하세요. 표에서 서비스를 확인하고 예(Yes) 링크를 선택하여 해당 서비스의 서비스 연결 역할 문서를 확인합니다.

IAM에서 서비스 연결 역할 삭제(AWS API)

IAM API를 사용하여 서비스 연결 역할을 삭제할 수 있습니다.

서비스 연결 역할(API)을 삭제하는 방법

1. 서비스 연결 역할 삭제 요청을 제출하려면 DeleteServiceLinkedRole를 호출합니다. 요청에 AWSServiceRoleForBatch 역할 이름을 지정합니다.

   서비스 연결 역할이 사용되지 않거나 연결된 리소스가 없는 경우에는 서비스 연결 역할을 삭제할 수 없으므로 삭제 요청을 제출해야 합니다. 이러한 조건이 충족되지 않으면 요청이 거부될 수 있습니다. 삭제 태스크 상태를 확인하려면 응답의 `DeletionTaskId`(을)를 캡처해야 합니다.
2. 삭제 상태를 확인하려면 GetServiceLinkedRoleDeletionStatus를 호출합니다. 요청에 `DeletionTaskId`(을)를 지정합니다.

   삭제 태스크는 `NOT_STARTED`, `IN_PROGRESS`, `SUCCEEDED` 또는 `FAILED` 상태일 수 있습니다. 삭제에 실패할 경우 문제를 해결할 수 있도록 실패 이유가 호출에 반환됩니다. 역할에서 서비스 리소스를 사용 중이어서 삭제에 실패한 경우에는 알림에 리소스 목록이 포함됩니다(서비스에서 해당 정보를 반환할 경우). 이후 리소스를 정리하고 삭제를 다시 제출할 수 있습니다.

> **Note**
>
> 서비스에서 반환하는 정보에 따라 이 과정을 여러 번 반복해야 할 수 있습니다. 예를 들어, 서비스 연결 역할에서 6개의 리소스를 사용할 수 있으며, 서비스에서 이중 5개에 관한 정보를 반환할 수 있습니다. 5개 리소스를 정리하고 삭제할 역할을 다시 제출할 경우 삭제에 실패하고 서비스에서 나머지 1개의 리소스를 보고합니다. 서비스에서 리소스 전부를 반환하거나, 일부만 반환하거나, 리소스를 보고하지 않을 수 있습니다. 어떤 리소스도 보고하지 않는 서비스의 리소스를 정리하는 방법은 IAM으로 작업하는AWS 서비스(을)를 참조하세요. 표에서 서비스를 확인하고 예(Yes) 링크를 선택하여 해당 서비스의 서비스 연결 역할 문서를 확인합니다.

AWS Batch 서비스 연결 역할에 지원되는 리전

AWS Batch 는 서비스를 사용할 수 있는 모든 리전에서 서비스 연결 역할 사용을 지원합니다. 자세한 내용은 AWS Batch 엔드포인트를 참조하세요.

SageMaker AI AWS Batch 에서에 대한 역할 사용

AWS Batch 는 AWS Identity and Access Management (IAM) 서비스 연결 역할을 사용합니다. 서비스 연결 역할은 직접 연결된 고유한 유형의 IAM 역할입니다 AWS Batch. 서비스 연결 역할은에서 사전 정의 AWS Batch 하며 서비스가 사용자를 대신하여 다른 AWS 서비스를 호출하는 데 필요한 모든 권한을 포함합니다.

필요한 권한을 수동으로 추가할 필요가 없으므로 서비스 연결 역할을 더 AWS Batch 쉽게 설정할 수 있습니다.는 서비스 연결 역할의 권한을 AWS Batch 정의하며, 달리 정의되지 않은 한 만 해당 역할을 수임 AWS Batch 할 수 있습니다. 정의된 권한에는 신뢰 정책과 권한 정책이 포함되며 이 권한 정책은 다른 IAM 엔터티에 연결할 수 없습니다.

먼저 관련 리소스를 삭제한 후에만 서비스 연결 역할을 삭제할 수 있습니다. 이렇게 하면 AWS Batch 리소스에 대한 액세스 권한을 실수로 제거할 수 없기 때문에 리소스가 보호됩니다.

서비스 연결 역할을 지원하는 다른 서비스에 대한 자세한 내용은 AWS IAM으로 작업하는 서비스를 참조하고 서비스 연결 역할 열에서 예인 서비스를 찾습니다. 해당 서비스에 대한 서비스 연결 역할 설명서를 보려면 예 링크를 선택합니다.

에 대한 서비스 연결 역할 권한 AWS Batch

AWS Batch 는 AWSServiceRoleForAWSBatchWithSagemaker라는 서비스 연결 역할을 사용합니다. AWS Batch 는 사용자를 대신하여 SageMaker 훈련 작업을 대기열에 넣고 관리할 수 있습니다.

AWSServiceRoleForAWSBatchWithSagemaker 서비스 연결 역할은 역할을 수임하기 위해 다음 서비스를 신뢰합니다.

- `sagemaker-queuing.batch.amazonaws.com`

역할 권한 정책은가 지정된 리소스에서 다음 작업을 완료 AWS Batch 하도록 허용합니다.

- `sagemaker` - AWS Batch 가 SageMaker 훈련 작업, 변환 작업 및 기타 SageMaker AI 리소스를 관리할 수 있도록 허용합니다.
- `iam:PassRole` - AWS Batch 가 작업 실행을 위해 고객 정의 실행 역할을 SageMaker AI에 전달할 수 있도록 허용합니다. 리소스 제약 조건은 역할을 SageMaker AI 서비스에 전달하도록 허용합니다.

사용자, 그룹 또는 역할이 서비스 연결 역할을 생성, 편집 또는 삭제할 수 있도록 사용 권한을 구성해야 합니다. 자세한 내용은 IAM 사용 설명서의 서비스 연결 역할 권한을 참조하세요.

에 대한 서비스 연결 역할 생성 AWS Batch

서비스 연결 역할은 수동으로 생성할 필요가 없습니다. AWS Management Console, AWS CLI또는 AWS API`CreateServiceEnvironment`에서를 사용하여 서비스 환경을 생성하면 AWS Batch 가 서비스 연결 역할을 생성합니다.

이 서비스 연결 역할을 삭제했다가 다시 생성해야 하는 경우 동일한 프로세스를 사용하여 계정에서 역할을 다시 생성할 수 있습니다. 를 사용하여 서비스 환경을 생성하면가 서비스 연결 역할을 다시 `CreateServiceEnvironment` AWS Batch 생성합니다.

정책에 대한 JSON을 보려면 AWS 관리형 정책 참조 안내서의 AWSBatchServiceRolePolicyForSageMaker를 참조하세요.

에 대한 서비스 연결 역할 편집 AWS Batch

AWS Batch 에서는 AWSServiceRoleForAWSBatchWithSagemaker 서비스 연결 역할을 편집할 수 없습니다. 서비스 연결 역할을 생성한 후에는 다양한 개체가 역할을 참조할 수 있기 때문에 역할 이름을 변경할 수 없습니다. 하지만 IAM을 사용하여 역할의 설명을 편집할 수 있습니다. 자세한 내용은 IAM 사용 설명서의 서비스 연결 역할 편집을 참조하세요.

에 대한 서비스 연결 역할 삭제 AWS Batch

서비스 연결 역할이 필요한 기능 또는 서비스가 더 이상 필요 없는 경우에는 해당 역할을 삭제할 것을 권합니다. 그렇게 하면 적극적으로 모니터링하거나 유지 관리하지 않은 미사용 엔터티가 없습니다. 단, 서비스 연결 역할을 정리해야 수동으로 삭제할 수 있습니다.

서비스 연결 역할을 정리

IAM을 사용하여 서비스 연결 역할을 삭제하려면 먼저 역할에 활성 세션이 없는지 확인하고 단일 파티션의 모든 AWS 리전에서 역할을 사용하는 모든 서비스 환경을 삭제해야 합니다.

서비스 연결 역할에 활성 세션이 있는지 확인하는 방법

1. IAM 콘솔(https://console.aws.amazon.com/iam/)을 엽니다.
2. 탐색 창에서 역할을 선택한 다음 확인란이 아닌 AWSServiceRoleForAWSBatchWithSagemaker 이름을 선택합니다.
3. 요약 페이지에서 액세스 고문을 선택하고 서비스 연결 역할의 최근 활동을 검토합니다.

> **Note**
>
> AWS Batch 가 AWSServiceRoleForAWSBatchWithSagemaker 역할을 사용하는지 여부를 모르는 경우 역할을 삭제해 볼 수 있습니다. 서비스에서 역할을 사용하는 경우에는 역할이 삭제되지 않습니다. 역할이 사용되고 있는 리전을 볼 수 있습니다. 역할이 사용 중인 경우에는 세션이 종료될 때까지 기다렸다가 역할을 삭제해야 합니다. 서비스 연결 역할에 대한 세션은 취소할 수 없습니다.

AWSServiceRoleForAWSBatchWithSagemaker 서비스 연결 역할에서 사용하는 AWS Batch 리소스를 제거하려면

AWSServiceRoleForAWSBatchWithSagemaker 역할을 삭제하려면 먼저 모든 리전에서 AWSServiceRoleForAWSBatchWithSagemaker AWSServiceRoleForAWSBatchWithSagemaker 역할을 사용하는 모든 서비스 환경을 모든 서비스 환경에서 모든 작업 대기열의 연결을 해제해야 합니다. AWS

1. https://console.aws.amazon.com/batch/ AWS Batch 콘솔을 엽니다.
2. 탐색 모음에서 사용할 리전을 선택합니다.
3. 탐색 창에서 환경을 선택한 다음 서비스 환경을 선택합니다.
4. 모든 서비스 환경을 선택합니다.
5. 비활성화를 선택합니다. 상태가 비활성화됨(DISABLED)으로 변경될 때까지 기다립니다.
6. 서비스 환경을 선택합니다.
7. 삭제를 선택합니다. 서비스 환경 삭제를 선택하여 컴퓨팅 환경을 삭제하고자 함을 확인합니다.
8. 모든 리전에서 서비스 연결 역할을 사용하는 모든 서비스 환경에 대해 1~7단계를 반복합니다.

IAM에서 서비스 연결 역할 삭제(콘솔)

IAM 콘솔을 사용하여 서비스 연결 역할을 삭제할 수 있습니다.

서비스 연결 역할을 삭제하는 방법(콘솔)

1. 에 로그인 AWS Management Console 하고 https://console.aws.amazon.com/iam/ IAM 콘솔을 엽니다.
2. IAM 콘솔의 탐색 창에서 역할을 선택합니다. 그런 다음 이름이나 행 자체가 아닌 AWSServiceRoleForAWSBatchWithSagemaker 옆의 확인란을 선택합니다.

3. 역할 삭제을 선택합니다.
4. 확인 대화 상자가 나타나면 서비스 마지막 액세스 데이터를 검토합니다. 이 데이터는 선택한 각 역할이 AWS 서비스(을)를 마지막으로 액세스한 일시를 보여줍니다. 이를 통해 역할이 현재 활동 중인지를 확인할 수 있습니다. 계속 진행하려면 예, 삭제합니다를 선택하여 삭제할 서비스 연결 역할을 제출합니다.
5. IAM 콘솔 알림을 보고 서비스 연결 역할 삭제 진행 상황을 모니터링합니다. IAM 서비스 연결 역할 삭제는 비동기이므로 삭제할 역할을 제출한 후에 삭제 태스크가 성공하거나 실패할 수 있습니다.

   - 태스크에 성공하면 목록에서 역할이 제거되고 성공 알림이 페이지 상단에 나타납니다.
   - 태스크에 실패할 경우 알림의 세부 정보 보기 또는 리소스 보기를 선택하면 삭제 실패 이유를 확인할 수 있습니다. 역할에서 서비스 리소스를 사용 중이어서 삭제에 실패한 경우에는 알림에 리소스 목록이 포함됩니다(서비스에서 해당 정보를 반환할 경우). 이후 리소스를 정리하고 삭제를 다시 제출할 수 있습니다.

     > **Note**
     >
     > 서비스에서 반환하는 정보에 따라 이 과정을 여러 번 반복해야 할 수 있습니다. 예를 들어, 서비스 연결 역할에서 6개의 리소스를 사용할 수 있으며, 서비스에서 이중 5개에 관한 정보를 반환할 수 있습니다. 5개 리소스를 정리하고 삭제할 역할을 다시 제출할 경우 삭제에 실패하고 서비스에서 나머지 1개의 리소스를 보고합니다. 서비스에서 리소스 전부를 반환하거나, 일부만 반환하거나, 리소스를 보고하지 않을 수 있습니다.

   - 태스크에 실패했는데 알림에 리소스 목록이 포함되지 않을 경우에는 서비스에서 해당 정보를 반환하지 않을 수 있습니다. 해당 서비스의 리소스를 정리하는 방법은 IAM으로 작업하는AWS 서비스를 참조하세요. 표에서 서비스를 확인하고 예(Yes) 링크를 선택하여 해당 서비스의 서비스 연결 역할 문서를 확인합니다.

IAM에서 서비스 연결 역할 삭제(AWS CLI)

에서 IAM 명령을 사용하여 서비스 연결 역할을 AWS Command Line Interface 삭제할 수 있습니다.

서비스 연결 역할을 삭제하는 방법(CLI)

1. 서비스 연결 역할이 사용되지 않거나 연결된 리소스가 없는 경우에는 서비스 연결 역할을 삭제할 수 없으므로 삭제 요청을 제출해야 합니다. 이러한 조건이 충족되지 않으면 요청이 거부될 수 있습니다. 삭제 태스크 상태를 확인하려면 응답의 `deletion-task-id`(을)를 캡처해야 합니다. 다음 명령을 입력하여 서비스 연결 역할 삭제 요청을 제출합니다.

```
$ aws iam delete-service-linked-role --role-name
 AWSServiceRoleForAWSBatchWithSagemaker
```

2. 다음 명령을 사용하여 삭제 태스크의 상태를 확인합니다.

```
$ aws iam get-service-linked-role-deletion-status --deletion-task-id deletion-task-id
```

삭제 태스크는 NOT_STARTED, IN_PROGRESS, SUCCEEDED 또는 FAILED 상태일 수 있습니다. 삭제에 실패할 경우 문제를 해결할 수 있도록 실패 이유가 호출에 반환됩니다. 역할에서 서비스 리소스를 사용 중이어서 삭제에 실패한 경우에는 알림에 리소스 목록이 포함됩니다(서비스에서 해당 정보를 반환할 경우). 이후 리소스를 정리하고 삭제를 다시 제출할 수 있습니다.

> **Note**
>
> 서비스에서 반환하는 정보에 따라 이 과정을 여러 번 반복해야 할 수 있습니다. 예를 들어, 서비스 연결 역할에서 6개의 리소스를 사용할 수 있으며, 서비스에서 이중 5개에 관한 정보를 반환할 수 있습니다. 5개 리소스를 정리하고 삭제할 역할을 다시 제출할 경우 삭제에 실패하고 서비스에서 나머지 1개의 리소스를 보고합니다. 서비스에서 리소스 전부를 반환하거나, 일부만 반환할 수 있습니다. 또는 리소스를 보고하지 않을 수도 있습니다. 어떤 리소스도 보고하지 않는 서비스의 리소스를 정리하는 방법은 IAM으로 작업하는AWS 서비스 섹션을 참조하세요. 표에서 서비스를 확인하고 예(Yes) 링크를 선택하여 해당 서비스의 서비스 연결 역할 문서를 확인합니다.

IAM에서 서비스 연결 역할 삭제(AWS API)

IAM API를 사용하여 서비스 연결 역할을 삭제할 수 있습니다.

서비스 연결 역할(API)을 삭제하는 방법

1. 서비스 연결 역할 삭제 요청을 제출하려면 DeleteServiceLinkedRole를 호출합니다. 요청에 AWSServiceRoleForAWSBatchWithSagemaker 역할 이름을 지정합니다.

   서비스 연결 역할이 사용되지 않거나 연결된 리소스가 없는 경우에는 서비스 연결 역할을 삭제할 수 없으므로 삭제 요청을 제출해야 합니다. 이러한 조건이 충족되지 않으면 요청이 거부될 수 있습니다. 삭제 태스크 상태를 확인하려면 응답의 DeletionTaskId(을)를 캡처해야 합니다.

2. 삭제 상태를 확인하려면 GetServiceLinkedRoleDeletionStatus를 호출합니다. 요청에 `DeletionTaskId`(을)를 지정합니다.

   삭제 태스크는 `NOT_STARTED`, `IN_PROGRESS`, `SUCCEEDED` 또는 `FAILED` 상태일 수 있습니다. 삭제에 실패할 경우 문제를 해결할 수 있도록 실패 이유가 호출에 반환됩니다. 역할에서 서비스 리소스를 사용 중이어서 삭제에 실패한 경우에는 알림에 리소스 목록이 포함됩니다(서비스에서 해당 정보를 반환할 경우). 이후 리소스를 정리하고 삭제를 다시 제출할 수 있습니다.

   > **Note**
   >
   > 서비스에서 반환하는 정보에 따라 이 과정을 여러 번 반복해야 할 수 있습니다. 예를 들어, 서비스 연결 역할에서 6개의 리소스를 사용할 수 있으며, 서비스에서 이중 5개에 관한 정보를 반환할 수 있습니다. 5개 리소스를 정리하고 삭제할 역할을 다시 제출할 경우 삭제에 실패하고 서비스에서 나머지 1개의 리소스를 보고합니다. 서비스에서 리소스 전부를 반환하거나, 일부만 반환하거나, 리소스를 보고하지 않을 수 있습니다. 어떤 리소스도 보고하지 않는 서비스의 리소스를 정리하는 방법은 IAM으로 작업하는AWS 서비스(을)를 참조하세요. 표에서 서비스를 확인하고 예(Yes) 링크를 선택하여 해당 서비스의 서비스 연결 역할 문서를 확인합니다.

AWS Batch 서비스 연결 역할에 지원되는 리전

AWS Batch 는 서비스를 사용할 수 있는 모든 리전에서 서비스 연결 역할 사용을 지원합니다. 자세한 내용은 AWS Batch 엔드포인트를 참조하세요.

Amazon ECS 인스턴스 역할

AWS Batch 컴퓨팅 환경은 Amazon ECS 컨테이너 인스턴스로 채워집니다. Amazon ECS 컨테이너 에이전트를 로컬에서 실행합니다. Amazon ECS 컨테이너 에이전트는 사용자를 대신하여 다양한 AWS API 작업을 호출합니다. 그러므로 에이전트를 실행하는 컨테이너 인스턴스는 해당 에이전트의 소유자를 확인하기 위해 이 서비스에 대한 IAM 정책과 역할을 요구합니다. 컨테이너 인스턴스를 시작할 때 사용할 IAM 역할과 인스턴스 프로파일을 생성해야 합니다. 그렇지 않으면 컴퓨팅 환경을 생성하고 해당 환경에서 컨테이너 인스턴스를 시작할 수 없습니다. 이 요구 사항은 Amazon에서 제공한 Amazon ECS 최적화 AMI를 사용하거나 사용하지 않고 시작된 컨테이너 인스턴스에 적용됩니다. 자세한 내용은 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS 인스턴스 역할 섹션을 참조하세요.

주제

- 계정의 Amazon ECS 인스턴스 역할 확인

계정의 Amazon ECS 인스턴스 역할 확인

Amazon ECS 인스턴스 역할과 인스턴스 프로파일은 콘솔 처음 실행 환경에서 자동으로 생성됩니다. 하지만 다음 단계를 수행하여 계정에 이미 Amazon ECS 인스턴스 역할과 인스턴스 프로파일이 있는지 확인할 수 있습니다. 다음 단계는 관리형 IAM 정책을 연결하는 방법도 다룹니다.

자습서: IAM 콘솔에서 **`ecsInstanceRole`** 확인

1. IAM 콘솔(https://console.aws.amazon.com/iam/)을 엽니다.
2. 탐색 창에서 역할을 선택합니다.
3. 역할 목록에서 `ecsInstanceRole`(을)를 검색합니다. 역할이 존재하지 않을 경우, 아래 단계를 사용하여 역할을 생성합니다.
   a. 역할 생성을 선택합니다.
   b. 신뢰할 수 있는 엔터티 유형에 AWS 서비스(을)를 선택합니다.
   c. 일반 사용 사례에서 EC2를 선택합니다.
   d. 다음을 선택합니다.
   e. 권한 정책에 대해서는 AmazonEC2ContainerServiceforEC2Role을 검색하세요.
   f. AmazonEC2ContainerServiceforEC2Role 옆에 있는 확인란을 선택한 후 다음을 선택합니다.
   g. 역할 이름에 `ecsInstanceRole`(을)를 입력하고 역할 생성을 선택합니다.

> **Note**
>
> AWS Management Console 를 사용하여 Amazon EC2에 대한 역할을 생성하는 경우 콘솔은 역할과 동일한 이름의 인스턴스 프로파일을 생성합니다.

또는 AWS CLI 를 사용하여 `ecsInstanceRole` IAM 역할을 생성할 수 있습니다. 다음 예제에서는 신뢰 정책 및 AWS 관리형 정책을 사용하여 IAM 역할을 생성합니다.

자습서: IAM 역할 및 인스턴스 프로파일 생성(AWS CLI)

1. 다음 신뢰 정책을 생성하고 `ecsInstanceRole-role-trust-policy.json`(이)라는 텍스트 파일로 저장합니다.

JSON

```
{
  "Version":"2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Principal": { "Service": "ec2.amazonaws.com"},
      "Action": "sts:AssumeRole"
    }
  ]
}
```

2. create-role 명령을 사용하여 ecsInstanceRole 역할을 생성합니다. assume-role-policy-document 파라미터에 신뢰 정책 파일 위치를 지정합니다.

```
$ aws iam create-role \
    --role-name ecsInstanceRole \
    --assume-role-policy-document file://ecsInstanceRole-role-trust-policy.json
```

3. create-instance-profile 명령을 사용하여 ecsInstanceRole(이이)라는 인스턴스 프로파일을 생성합니다.

> **Note**
>
> AWS CLI 및 AWS API에서 역할과 인스턴스 프로파일을 별도의 작업으로 생성해야 합니다.

```
$ aws iam create-instance-profile --instance-profile-name ecsInstanceRole
```

다음은 응답의 예입니다.

```
{
    "InstanceProfile": {
        "Path": "/",
        "InstanceProfileName": "ecsInstanceRole",
        "InstanceProfileId": "AIPAT46P5RDITREXAMPLE",
        "Arn": "arn:aws:iam::123456789012:instance-profile/ecsInstanceRole",
```

```
        "CreateDate": "2022-06-30T23:53:34.093Z",
        "Roles": [],    }
}
```

4. add-role-to-instance-profile 명령을 사용하여 ecsInstanceRole 인스턴스 프로파일에 ecsInstanceRole 역할을 추가합니다.

```
aws iam add-role-to-instance-profile \
    --role-name ecsInstanceRole --instance-profile-name ecsInstanceRole
```

5. attach-role-policy 명령을 사용하여 AmazonEC2ContainerServiceforEC2Role AWS 관리형 정책을 ecsInstanceRole 역할에 연결합니다.

```
$ aws iam attach-role-policy \
    --policy-arn arn:aws:iam::aws:policy/service-role/
AmazonEC2ContainerServiceforEC2Role \
    --role-name ecsInstanceRole
```

Amazon EC2 스팟 플릿 역할

Amazon EC2 스팟 플릿 인스턴스를 사용하는 관리형 컴퓨팅 환경을 생성하는 경우 AmazonEC2SpotFleetTaggingRole 정책을 생성해야 합니다. 이는 사용자 대신 인스턴스를 시작, 태그, 종료할 수 있는 스팟 플릿 권한을 정책에 부여합니다. 스팟 플릿 요청에서 권한을 지정합니다. 또한 Amazon EC2 스팟 및 스팟 플릿에 대한 AWSServiceRoleForEC2Spot 및 AWSServiceRoleForEC2SpotFleet 서비스 연결 역할이 있어야 합니다. 이 모든 역할을 생성하려면 다음 지침을 따릅니다. 자세한 내용은 IAM 사용 설명서의 서비스 연결 역할 사용 및 AWS 서비스에 대한 권한 위임을 위한 역할 생성을 참조하세요.

주제

- 에서 Amazon EC2 스팟 플릿 역할 생성 AWS Management Console
- 를 사용하여 Amazon EC2 스팟 플릿 역할 생성 AWS CLI

에서 Amazon EC2 스팟 플릿 역할 생성 AWS Management Console

Amazon EC2 스팟 플릿에 대한 **AmazonEC2SpotFleetTaggingRole** IAM 서비스 연결 역할을 생성하려면

1. IAM 콘솔(https://console.aws.amazon.com/iam/)을 엽니다.

2. 액세스 관리에서 역할을 선택합니다.
3. 역할에서 역할 생성을 선택합니다.
4. 신뢰할 수 있는 엔터티 유형에 대한 신뢰할 수 있는 엔터티 선택에서 AWS 서비스를 선택합니다.
5. 다른의 사용 사례 AWS 서비스에서 EC2를 선택한 다음 EC2 - 스팟 플릿 태그 지정을 선택합니다.
6. 다음을 선택합니다.
7. 권한 정책의 정책 이름에서는 `AmazonEC2SpotFleetTaggingRole`을 확인합니다.
8. 다음을 선택합니다.
9. 이름 지정, 검토 및 생성:

    a. 역할 이름에 역할을 식별하는 이름을 입력합니다.

    b. 설명에 정책에 대한 간단한 설명을 입력합니다.

    c. (선택 사항) 1단계: 신뢰할 수 있는 엔터티 선택에서 편집을 선택하여 코드를 수정합니다.

    d. (선택 사항) 2단계: 권한 추가에서는 편집을 선택하여 코드를 수정합니다.

    e. (선택 사항) 태그 추가에서는 태그 추가를 선택하여 리소스에 태그를 추가합니다.

    f. 역할 생성을 선택합니다.

> **Note**
>
> 과거에는 Amazon EC2 스팟 플릿 역할에 대해 두 가지 관리형 정책이 있었습니다.
>
> - AmazonEC2SpotFleetRole: 이 항목은 스팟 플릿 역할에 대한 기존 관리형 정책입니다. 그러나 더 이상 사용하지 않는 것이 좋습니다 AWS Batch. 이 정책은 `AWSServiceRoleForBatch` 서비스 연결 역할을 사용하는 데 필요한 컴퓨팅 환경의 스팟 인스턴스 태그 지정을 지원하지 않습니다. 이 정책이 있는 스팟 플릿 역할을 이전에 만든 경우, 해당 역할에 새 권장 정책을 적용합니다. 자세한 내용은 생성 시 태그가 지정되지 않은 스팟 인스턴스 단원을 참조하십시오.
> - AmazonEC2SpotFleetTaggingRole: 이 역할은 Amazon EC2 스팟 인스턴스에 태그를 지정하는 데 필요한 모든 권한을 제공합니다. 이 역할을 사용하여 AWS Batch 컴퓨팅 환경에서 스팟 인스턴스 태그 지정을 허용합니다.

를 사용하여 Amazon EC2 스팟 플릿 역할 생성 AWS CLI

스팟 플릿 컴퓨팅 환경에 대한 AmazonEC2SpotFleetTaggingRole IAM 역할을 생성하려면

1. AWS CLI를 사용하여 다음 명령을 실행합니다.

```
$ aws iam create-role --role-name AmazonEC2SpotFleetTaggingRole \
     --assume-role-policy-document '{
  "Version":"2012-10-17",
  "Statement": [
    {
      "Sid": "",
      "Effect": "Allow",
      "Principal": {
        "Service": "spotfleet.amazonaws.com"
      },
      "Action": "sts:AssumeRole"
    }
  ]
}'
```

2. AmazonEC2SpotFleetTaggingRole 관리형 IAM 정책을 AmazonEC2SpotFleetTaggingRole 역할에 연결하려면 AWS CLI를 사용하여 다음 명령을 실행합니다.

```
$ aws iam attach-role-policy \
  --policy-arn \
    arn:aws:iam::aws:policy/service-role/AmazonEC2SpotFleetTaggingRole \
  --role-name \
    AmazonEC2SpotFleetTaggingRole
```

Amazon EC2 스팟에 대한 **AWSServiceRoleForEC2Spot** IAM 서비스 연결 역할을 생성하려면

> **Note**
>
> AWSServiceRoleForEC2Spot IAM 서비스 연결 역할이 이미 있는 경우 다음과 비슷한 오류 메시지가 표시됩니다.
>
> ```
> An error occurred (InvalidInput) when calling the CreateServiceLinkedRole
>  operation:
> ```

```
Service role name AWSServiceRoleForEC2Spot has been taken in this account,
 please try a different suffix.
```

- AWS CLI를 사용하여 다음 명령을 실행합니다.

```
$ aws iam create-service-linked-role --aws-service-name spot.amazonaws.com
```

Amazon EC2 스팟 플릿에 대한 **AWSServiceRoleForEC2SpotFleet** IAM 서비스 연결 역할을 생성하려면

> **Note**
>
> AWSServiceRoleForEC2SpotFleet IAM 서비스 연결 역할이 이미 있는 경우 다음과 비슷한 오류 메시지가 표시됩니다.
>
> ```
> An error occurred (InvalidInput) when calling the CreateServiceLinkedRole
>  operation:
> Service role name AWSServiceRoleForEC2SpotFleet has been taken in this account,
>  please try a different suffix.
> ```

- AWS CLI를 사용하여 다음 명령을 실행합니다.

```
$ aws iam create-service-linked-role --aws-service-name spotfleet.amazonaws.com
```

EventBridge IAM 역할

Amazon EventBridge는 AWS 리소스의 변경 사항을 설명하는 시스템 이벤트의 스트림을 거의 실시간으로 제공합니다. AWS Batch 작업은 EventBridge 대상으로 사용할 수 있습니다. 빠르게 설정할 수 있는 단순한 규칙을 사용하여 이벤트를 일치시키고 해당 이벤트에 응답하여 AWS Batch 작업을 제출할 수 있습니다. EventBridge 규칙 및 대상으로 AWS Batch 작업을 제출하려면 먼저 EventBridge에 사용자를 대신하여 AWS Batch 작업을 실행할 수 있는 권한이 있어야 합니다.

> **Note**
>
> EventBridge 콘솔에서 AWS Batch 대기열을 대상으로 지정하는 규칙을 생성할 때이 역할을 생성할 수 있습니다. 예제 연습은 EventBridge 대상으로 사용되는 AWS Batch작업 섹션을 참조하세요. IAM 콘솔을 사용하여 EventBridge 역할을 수동으로 생성할 수 있습니다. 지침은 IAM 사용 설명서에서 사용자 지정 신뢰 정책을 사용한 역할 생성(콘솔)을 참조하세요.

EventBridge IAM 역할에 대한 신뢰 관계는 `events.amazonaws.com` 서비스 보안 주체에게 역할을 수임할 수 있는 기능을 제공해야 합니다.

JSON

```
{
  "Version":"2012-10-17",
  "Statement": [
    {
      "Sid": "",
      "Effect": "Allow",
      "Principal": {
        "Service": "events.amazonaws.com"
      },
      "Action": "sts:AssumeRole"
    }
  ]
}
```

EventBridge IAM 역할에 연결된 정책이 리소스에 대한 `batch:SubmitJob` 권한을 허용하는지 확인합니다. 다음 예시에서는 AWS Batch 는 이러한 권한을 제공하는 `AWSBatchServiceEventTargetRole` 관리형 정책을 제공합니다.

JSON

```
{
  "Version":"2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
```

```
      "Action": [
        "batch:SubmitJob"
       ],
      "Resource": "*"
    }
  ]
}
```

Virtual Private Cloud 생성

컴퓨팅 환경의 컴퓨팅 리소스는 AWS Batch 및 Amazon ECS 서비스 엔드포인트와 통신하려면 외부 네트워크에 액세스해야 합니다. 그러나 프라이빗 서브넷에서 실행하고 싶은 작업이 있을 수 있습니다. 퍼블릭 또는 프라이빗 서브넷에서 작업을 실행할 수 있는 유연성을 가지려면 퍼블릭 서브넷과 프라이빗 서브넷이 모두 있는 VPC를 생성하세요.

Amazon Virtual Private Cloud(VPC)를 사용하여 사용자가 정의한 가상 네트워크로 AWS 리소스를 시작할 수 있습니다. 이 주제에서는 Amazon VPC 마법사로 연결되는 링크와 선택할 수 있는 옵션 목록을 제공합니다.

VPC 생성

Amazon VPC 생성하는 방법에 대한 자세한 내용은 Amazon VPC 사용 설명서의 VPC만 생성을 참조하고 다음 표를 사용하여 선택할 옵션을 선택하세요.

옵션	값
생성할 리소스	VPC 전용
이름	선택적으로 VPC의 이름을 입력합니다.
IPv4 CIDR block	IPv4 CIDR 수동 입력 CIDR 블록 크기는 /16과 /28 사이여야 합니다.
IPv6 CIDR block	No IPv6 CIDR 블록
테넌시	기본값

Amazon VPC에 대한 자세한 내용은 Amazon VPC 사용 설명서의 Amazon VPC란 무엇인가? 섹션을 참조하세요.

다음 단계

VPC를 생성한 후 다음 단계를 고려합니다.

- 인바운드 네트워크 액세스가 필요한 경우 퍼블릭 및 프라이빗 리소스에 대한 보안 그룹을 만듭니다. 자세한 내용은 Amazon VPC 사용 설명서의 보안 그룹 작업을 참조하세요.
- 새 VPC로 컴퓨팅 리소스를 시작하는 AWS Batch 관리형 컴퓨팅 환경을 생성합니다. 자세한 내용은 컴퓨팅 환경 생성 단원을 참조하십시오. AWS Batch 콘솔에서 컴퓨팅 환경 생성 마법사를 사용하는 경우 방금 생성한 VPC와 인스턴스를 시작하려는 퍼블릭 또는 프라이빗 서브넷을 지정할 수 있습니다.
- 새 컴퓨팅 환경에 매핑되는 AWS Batch 작업 대기열을 생성합니다. 자세한 내용은 작업 대기열 생성 단원을 참조하십시오.
- 태스크를 실행할 태스크 정의를 만듭니다. 자세한 내용은 단일 노드 작업 정의 생성 섹션을 참조하세요.
- 태스크 정의가 있는 태스크를 새 작업 대기열에 제출합니다. 이 작업은 새 VPC와 서브넷으로 만든 컴퓨팅 환경에 전달됩니다. 자세한 내용은 자습서: 작업 제출 단원을 참조하십시오.

인터페이스 엔드포인트를 사용하여 액세스 AWS Batch

AWS PrivateLink 를 사용하여 VPC와 간에 프라이빗 연결을 생성할 수 있습니다 AWS Batch. 인터넷 게이트웨이, NAT 디바이스, VPN 연결 또는 Direct Connect 연결을 사용하지 않고 VPC에 있는 것처럼 AWS Batch 에 액세스할 수 있습니다. VPC의 인스턴스에서 AWS Batch API에 액세스하는 데는 퍼블릭 IP 주소가 필요하지 않습니다.

AWS PrivateLink에서 제공되는 인터페이스 엔드포인트를 생성하여 이 프라이빗 연결을 설정합니다. 인터페이스 엔드포인트에 대해 사용 설정하는 각 서브넷에서 엔드포인트 네트워크 인터페이스를 생성합니다. 이는 AWS Batch로 향하는 트래픽의 진입점 역할을 하는 요청자 관리형 네트워크 인터페이스입니다.

자세한 내용은 AWS PrivateLink 사용 설명서의 인터페이스스 VPC 엔드포인트를 참조하세요.

에 대한 고려 사항 AWS Batch

에 대한 인터페이스 엔드포인트를 설정하기 전에 AWS PrivateLink 가이드의 인터페이스 엔드포인트 속성 및 제한 사항을 AWS Batch검토하세요.

AWS Batch 는 인터페이스 엔드포인트를 통해 모든 API 작업을 호출할 수 있도록 지원합니다.

에 대한 인터페이스 VPC 엔드포인트를 설정하기 전에 다음 고려 사항에 유의 AWS Batch하세요.

- Fargate 리소스 시작 유형을 사용하는 작업에는 Amazon ECS용 인터페이스 VPC 엔드포인트가 필요하지 않지만 다음 사항에 설명된 Amazon ECR AWS Batch, Secrets Manager 또는 Amazon CloudWatch Logs용 인터페이스 VPC 엔드포인트가 필요할 수 있습니다.
  - 작업을 실행하려면 Amazon ECS에 대한 인터페이스 VPC 엔드포인트를 생성해야 합니다. 자세한 내용은 Amazon Elastic Container Service 개발자 가이드의 인터페이스 VPC 엔드포인트(AWS PrivateLink)를 참조하세요.
  - 작업에서 Amazon ECR의 프라이빗 이미지를 가져오도록 허용하려면 Amazon ECR용 인터페이스 VPC 엔드포인트를 생성해야 합니다. 자세한 내용은 Amazon Elastic Container Registry 사용 설명서의 인터페이스 VPC 엔드포인트(AWS PrivateLink)를 참조하세요.
  - 작업에서 Secrets Manager에서 민감한 데이터를 가져오도록 허용하려면 Secrets Manager용 인터페이스 VPC 엔드포인트를 생성해야 합니다. 자세한 내용은 AWS Secrets Manager 사용 설명서의 VPC 엔드포인트와 함께 Secrets Manager 사용을 참조하세요.
  - VPC에 인터넷 게이트웨이가 없고 작업에서 `awslogs` 로그 드라이버를 사용하여 로그 정보를 CloudWatch Logs로 전송하는 경우에는 CloudWatch Logs용 인터페이스 VPC 엔드포인트를 생성해야 합니다. 자세한 내용은 Amazon CloudWatch Logs 사용 설명서의 인터페이스 VPC 엔드포인트에서 CloudWatch Logs 사용을 참조하세요.
- EC2 리소스를 사용하는 작업은 시작되는 컨테이너 인스턴스가 Amazon ECS 컨테이너 에이전트의 `1.25.1` 이상 버전으로 실행되어야 합니다. 자세한 내용을 알아보려면 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS Linux 컨테이너 에이전트 버전을 참조하세요.
- VPC 엔드포인트는 교차 리전 요청을 현재 지원하지 않습니다. API 호출을 AWS Batch(으)로 발행할 계획인 동일 리전에서 엔드포인트를 생성해야 합니다.
- VPC 엔드포인트는 Amazon Route 53을 통해 Amazon이 제공하는 DNS만 지원합니다. 자체 DNS를 사용하는 경우에는 조건부 DNS 전달을 사용할 수 있습니다. 자세한 내용은 Amazon VPC 사용 설명서의 DHCP 옵션 세트를 참조하세요.
- VPC 엔드포인트에 연결된 보안 그룹은 VPC의 프라이빗 서브넷에서 443 포트로 들어오는 연결을 허용해야 합니다.
- AWS Batch 는 AWS 리전다음에서 VPC 인터페이스 엔드포인트를 지원하지 않습니다.
  - 아시아 태평양(오사카)(`ap-northeast-3`)
  - 아시아 태평양(자카르타)(`ap-southeast-3`)

에 대한 인터페이스 엔드포인트 생성 AWS Batch

Amazon VPC 콘솔 또는 AWS Command Line Interface ()를 AWS Batch 사용하여 용 인터페이스 엔드포인트를 생성할 수 있습니다AWS CLI. 자세한 내용은 AWS PrivateLink 안내서의 인터페이스 엔드포인트 생성을 참조하세요.

다음 서비스 이름을 AWS Batch 사용하여 용 인터페이스 엔드포인트를 생성합니다.

- com.amazonaws.*region* .batch
- com.amazonaws.*region* .batch-fips(FIPS 준수 엔드포인트의 경우 AWS Batch 엔드포인트 및 할당량 참조)

예제:

```
com.amazonaws.us-east-2.batch
```

```
com.amazonaws.us-east-2.batch-fips
```

aws-cn 파티션에서는 형식이 다릅니다.

```
cn.com.amazonaws.region.batch
```

예제:

```
cn.com.amazonaws.cn-northwest-1.batch
```

AWS Batch 인터페이스 엔드포인트의 프라이빗 DNS 이름

인터페이스 엔드포인트에 대해 프라이빗 DNS를 활성화하는 경우 특정 DNS 이름을 사용하여 연결할 수 있습니다 AWS Batch. 다음 옵션을 제공합니다.

- batch.*region* .amazonaws.com
- batch.*region* .api.aws

FIPS 준수 엔드포인트의 경우:

- batch-fips.*region* .api.aws

- fips.batch.*region* .amazonaws.com 지원되지 않음

자세한 내용은 AWS PrivateLink 가이드의 인터페이스 엔드포인트를 통해 서비스 액세스를 참조하세요.

인터페이스 엔드포인트의 엔드포인트 정책을 생성

엔드포인트 정책은 인터페이스 엔드포인트에 연결할 수 있는 IAM 리소스입니다. 기본 엔드포인트 정책은 인터페이스 엔드포인트를 AWS Batch 통해에 대한 전체 액세스를 허용합니다. VPC에서 AWS Batch 에 허용되는 액세스를 제어하려면 사용자 지정 엔드포인트 정책을 인터페이스 엔드포인트에 연결합니다.

엔드포인트 정책은 다음 정보를 지정합니다.

- 작업을 수행할 수 있는 보안 주체(AWS 계정, 사용자자 및 IAM 역할) .
- 수행할 수 있는 작업
- 작업을 수행할 수 있는 리소스.

자세한 내용은 AWS PrivateLink 안내서의 엔드포인트 정책을 사용하여 서비스에 대한 액세스 제어를 참조하세요.

예: AWS Batch 작업에 대한 VPC 엔드포인트 정책

다음은 사용자 지정 엔드포인트 정책의 예입니다. 이 정책을 인터페이스 엔드포인트에 연결하면 모든 리소스의 모든 보안 주체에 대해 나열된 AWS Batch 작업에 대한 액세스 권한이 부여됩니다.

```
{
   "Statement": [
      {
         "Principal": "*",
         "Effect": "Allow",
         "Action": [
            "batch:SubmitJob",
            "batch:ListJobs",
            "batch:DescribeJobs"
         ],
         "Resource":"*"
      }
   ]
}
```

에 대한 규정 준수 검증 AWS Batch

AWS 서비스 가 특정 규정 준수 프로그램의 범위 내에 있는지 알아보려면 AWS 서비스 규정 준수 프로그램 범위 내를 참조하고 관심 있는 규정 준수 프로그램을 선택합니다. 일반 정보는 AWS 규정 준수 프로그램.

를 사용하여 타사 감사 보고서를 다운로드할 수 있습니다 AWS Artifact. 자세한 내용은 에서 보고서 다운로드 AWS Artifact에서 .

사용 시 규정 준수 책임은 데이터의 민감도, 회사의 규정 준수 목표 및 관련 법률과 규정에 따라 AWS 서비스 결정됩니다. 사용 시 규정 준수 책임에 대한 자세한 내용은 AWS 보안 설명서를 AWS 서비스참조하세요.

의 인프라 보안 AWS Batch

관리형 서비스인는 AWS 글로벌 네트워크 보안으로 보호 AWS Batch 됩니다. AWS 보안 서비스 및가 인프라를 AWS 보호하는 방법에 대한 자세한 내용은 AWS 클라우드 보안을 참조하세요. 인프라 보안 모범 사례를 사용하여 환경을 설계하려면 보안 원칙 AWS Well-Architected Framework의 인프라 보호를 참조하세요 AWS .

AWS 에서 게시한 API 호출을 사용하여 네트워크를 AWS Batch 통해에 액세스합니다. 클라이언트는 다음을 지원해야 합니다.

- Transport Layer Security(TLS). TLS 1.2는 필수이며 TLS 1.3을 권장합니다.
- DHE(Ephemeral Diffie-Hellman) 또는 ECDHE(Elliptic Curve Ephemeral Diffie-Hellman)와 같은 완전 전송 보안(PFS)이 포함된 암호 제품군. Java 7 이상의 최신 시스템은 대부분 이러한 모드를 지원합니다.

모든 네트워크 위치에서 이러한 API 작업을 호출할 수 있지만 AWS Batch 는 소스 IP 주소에 따른 제한을 포함할 수 있는 리소스 기반 액세스 정책을 지원합니다. AWS Batch 정책을 사용하여 특정 Amazon Virtual Private Cloud(Amazon VPC) 엔드포인트 또는 특정 VPCs. 이렇게 하면 네트워크 내의 특정 VPC에서만 지정된 AWS Batch 리소스에 대한 AWS 네트워크 액세스가 효과적으로 격리됩니다.

교차 서비스 혼동된 대리자 방지

혼동된 대리자 문제는 작업을 수행할 권한이 없는 엔터티가 권한이 더 많은 엔터티에 작업을 수행하도록 강요할 수 있는 보안 문제입니다. 에서 AWS교차 서비스 가장은 혼동된 대리자 문제를 초래할 수 있

습니다. 교차 서비스 가장은 한 서비스(직접 호출하는 서비스)가 다른 서비스(직접 호출되는 서비스)를 직접 호출할 때 발생할 수 있습니다. 직접 호출하는 서비스는 다른 고객의 리소스에 대해 액세스 권한이 없는 방식으로 작동하게 권한을 사용하도록 조작될 수 있습니다. 이를 방지하기 위해 AWS 에서는 계정의 리소스에 대한 액세스 권한이 부여된 서비스 위탁자를 사용하여 모든 서비스에 대한 데이터를 보호하는 데 도움이 되는 도구를 제공합니다.

리소스 정책에서 aws:SourceArn 및 aws:SourceAccount 전역 조건 컨텍스트 키를 사용하여 리소스에 다른 서비스를 AWS Batch 제공하는 권한을 제한하는 것이 좋습니다. 만약 aws:SourceArn 값에 Amazon S3 버킷 ARN과 같은 계정 ID가 포함되어 있지 않은 경우, 권한을 제한하려면 두 전역 조건 컨텍스트 키를 모두 사용해야 합니다. 두 전역 조건 컨텍스트 키와 계정을 포함한 aws:SourceArn 값을 모두 사용하는 경우, aws:SourceAccount 값 및 aws:SourceArn 값의 계정은 동일한 정책 명령문에서 사용할 경우 반드시 동일한 계정 ID를 사용해야 합니다. 하나의 리소스만 교차 서비스 액세스와 연결되도록 허용하려는 경우 aws:SourceArn을 사용하세요. 해당 계정의 모든 리소스가 교차 서비스 사용과 연결되도록 허용하려는 경우 aws:SourceAccount를 사용하세요.

의 값은가 AWS Batch 저장하는 리소스여야 aws:SourceArn 합니다.

혼동된 대리인 문제로부터 보호하는 가장 효과적인 방법은 리소스의 전체 ARN이 포함된 aws:SourceArn글로벌 조건 컨텍스트 키를 사용하는 것입니다. 리소스의 전체 ARN을 모르거나 여러 리소스를 지정하는 경우, ARN의 알 수 없는 부분에 대해 와일드카드 문자(*)를 포함한 aws:SourceArn 글로벌 조건 컨텍스트 키를 사용합니다. 예를 들어 arn:aws:*servicename*:*:*123456789012*:*입니다.

다음 예제에서는의 aws:SourceArn 및 aws:SourceAccount 전역 조건 컨텍스트 키를 사용하여 혼동된 대리자 문제를 AWS Batch 방지하는 방법을 보여줍니다.

예: 하나의 컴퓨팅 환경에만 액세스하는 역할

다음 역할은 하나의 컴퓨팅 환경에 액세스하는 데만 사용할 수 있습니다. 작업 대기열이 여러 컴퓨팅 환경과 연결될 수 있으므로 작업 이름을 *(으)로 지정해야 합니다.

JSON

```
{
  "Version":"2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Principal": {
```

```
        "Service": "batch.amazonaws.com"
      },
      "Action": "sts:AssumeRole",
      "Condition": {
        "StringEquals": {
          "aws:SourceAccount": "123456789012"
        },
        "ArnLike": {
          "aws:SourceArn": [
            "arn:aws:batch:us-east-1:123456789012:compute-environment/testCE",
            "arn:aws:batch:us-east-1:123456789012:job/*"
          ]
        }
      }
    }
  ]
}
```

예: 여러 컴퓨팅 환경에 액세스하는 역할

다음 역할은 여러 컴퓨팅 환경에 액세스하는 데 사용할 수 있습니다. 작업 대기열이 여러 컴퓨팅 환경과 연결될 수 있으므로 작업 이름을 *(으)로 지정해야 합니다.

JSON

```
{
  "Version":"2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Principal": {
        "Service": "batch.amazonaws.com"
      },
      "Action": "sts:AssumeRole",
      "Condition": {
        "StringEquals": {
          "aws:SourceAccount": "123456789012"
        },
        "ArnLike": {
          "aws:SourceArn": [
            "arn:aws:batch:us-east-1:123456789012:compute-environment/*",
```

```
            "arn:aws:batch:us-east-1:123456789012:job/*"
          ]
        }
      }
    }
  ]
}
```

를 사용하여 AWS Batch API 호출 로깅 AWS CloudTrail

AWS Batch 는 사용자 AWS CloudTrail, 역할 또는 서비스가 수행한 작업에 대한 레코드를 제공하는 AWS 서비스와 통합됩니다 AWS Batch. CloudTrail은에 대한 모든 API 호출을 이벤트 AWS Batch 로 캡처합니다. 캡처되는 호출에는 AWS Batch 콘솔의 호출과 AWS Batch API 작업에 대한 코드 호출이 포함됩니다. 추적을 생성하면 이벤트를 포함한 CloudTrail 이벤트를 지속적으로 Amazon S3 버킷에 배포할 수 있습니다 AWS Batch. 추적을 구성하지 않은 경우에도 CloudTrail 콘솔의 이벤트 기록에서 최신 이벤트를 볼 수 있습니다. CloudTrail에서 수집한 정보를 사용하여 AWS Batch에 수행된 요청, 요청이 수행된 IP 주소, 요청을 수행한 사람, 요청이 수행된 시간 및 추가 세부 정보를 확인할 수 있습니다.

CloudTrail에 대한 자세한 설명은 AWS CloudTrail 사용자 가이드를 참조하십시오.

주제

- AWS Batch CloudTrail의 정보
- 참조: AWS Batch 로그 파일 항목 이해

AWS Batch CloudTrail의 정보

AWS 계정을 생성할 때 계정에서 CloudTrail이 활성화됩니다. 활동이 발생하면 AWS Batch해당 활동이 이벤트 기록의 다른 AWS 서비스 이벤트와 함께 CloudTrail 이벤트에 기록됩니다. AWS 계정에서 최근 이벤트를 보고 검색하고 다운로드할 수 있습니다. 자세한 설명은 CloudTrail 이벤트 기록으로 이벤트 보기를 참조하세요.

에 대한 이벤트를 포함하여 AWS 계정에 이벤트를 지속적으로 기록하려면 추적을 AWS Batch생성합니다. CloudTrail은 추적을 사용하여 Amazon S3 버킷으로 로그 파일을 전송할 수 있습니다. 기본적으로 콘솔에서 추적을 생성하면 추적이 모든 AWS 리전에 적용됩니다. 추적은 AWS 파티션의 모든 리전에서 이벤트를 로깅하고 지정한 Amazon S3 버킷으로 로그 파일을 전송합니다. 또한 CloudTrail 로그에서 수집된 이벤트 데이터를 추가로 분석하고 조치를 취하도록 다른 AWS 서비스를 구성할 수 있습니다. 자세한 내용은 다음 자료를 참조하세요.

- 추적 생성 개요
- CloudTrail 지원 서비스 및 통합
- CloudTrail에서 Amazon SNS 알림 구성
- 여러 리전으로부터 CloudTrail 로그 파일 받기 및 여러 계정으로부터 CloudTrail 로그 파일 받기

모든 AWS Batch 작업은 CloudTrail에서 로깅되며 https://docs.aws.amazon.com/batch/latest/APIReference/ 문서화됩니다. 예를 들어 `SubmitJob`, `ListJobs`, `DescribeJobs` 섹션을 호출하면 CloudTrail 로그 파일에 항목이 생성됩니다.

모든 이벤트 또는 로그 항목에는 요청을 생성했던 사용자에 관한 정보가 포함됩니다. ID 정보를 이용하면 다음을 쉽게 판단할 수 있습니다.

- 요청을 루트로 했는지 아니면 IAM 사용자 자격 증명 정보로 했는지 여부.
- 역할 또는 페더레이션 사용자의 임시 자격 증명을 사용하여 요청이 생성되었는지 여부.
- 요청이 다른 AWS 서비스에서 이루어졌는지 여부입니다.

자세한 설명은 CloudTrail userIdentity 요소를 참조하세요.

참조: AWS Batch 로그 파일 항목 이해

트레일이란 지정한 S3 버킷에 이벤트를 로그 파일로 입력할 수 있게 하는 구성입니다. CloudTrail 로그 파일에는 하나 이상의 로그 항목이 포함될 수 있습니다. 이벤트는 모든 소스로부터의 단일 요청을 나타내며 요청 작업, 작업 날짜와 시간, 요청 파라미터 등에 대한 정보가 들어 있습니다. CloudTrail 로그 파일은 퍼블릭 API 직접 호출의 주문 스택 트레이스가 아니므로 특정 순서로 표시되지 않습니다.

다음 예제는 `CreateComputeEnvironment` 작업을 보여주는 CloudTrail 로그 항목이 나타냅니다.

```
{
  "eventVersion": "1.05",
  "userIdentity": {
    "type": "AssumedRole",
    "principalId": "AIDACKCEVSQ6C2EXAMPLE:admin",
    "arn": "arn:aws:sts::012345678910:assumed-role/Admin/admin",
    "accountId": "012345678910",
    "accessKeyId": "AKIAIOSFODNN7EXAMPLE",
    "sessionContext": {
      "attributes": {
        "mfaAuthenticated": "false",
        "creationDate": "2017-12-20T00:48:46Z"
```

```
      },
      "sessionIssuer": {
        "type": "Role",
        "principalId": "AIDACKCEVSQ6C2EXAMPLE",
        "arn": "arn:aws:iam::012345678910:role/Admin",
        "accountId": "012345678910",
        "userName": "Admin"
      }
    }
  },
  "eventTime": "2017-12-20T00:48:46Z",
  "eventSource": "batch.amazonaws.com",
  "eventName": "CreateComputeEnvironment",
  "awsRegion": "us-east-1",
  "sourceIPAddress": "203.0.113.1",
  "userAgent": "aws-cli/1.11.167 Python/2.7.10 Darwin/16.7.0 botocore/1.7.25",
  "requestParameters": {
    "computeResources": {
      "subnets": [
        "subnet-5eda8e04"
      ],
      "tags": {
        "testBatchTags": "CLI testing CE"
      },
      "desiredvCpus": 0,
      "minvCpus": 0,
      "instanceTypes": [
        "optimal"
      ],
      "securityGroupIds": [
        "sg-aba9e8db"
      ],
      "instanceRole": "ecsInstanceRole",
      "maxvCpus": 128,
      "type": "EC2"
    },
    "state": "ENABLED",
    "type": "MANAGED",
    "computeEnvironmentName": "Test"
  },
  "responseElements": {
    "computeEnvironmentName": "Test",
    "computeEnvironmentArn": "arn:aws:batch:us-east-1:012345678910:compute-environment/
Test"
```

```
  },
  "requestID": "890b8639-e51f-11e7-b038-EXAMPLE",
  "eventID": "874f89fa-70fc-4798-bc00-EXAMPLE",
  "readOnly": false,
  "eventType": "AwsApiCall",
  "recipientAccountId": "012345678910"
}
```

AWS Batch IAM 문제 해결

다음 정보를 사용하여 및 IAM으로 작업할 때 발생할 수 있는 일반적인 문제를 진단 AWS Batch 하고 수정할 수 있습니다.

주제

- 에서 작업을 수행할 권한이 없음 AWS Batch
- iam:PassRole을 수행하도록 인증되지 않음
- 내 AWS 계정 외부의 사람이 내 AWS Batch 리소스에 액세스하도록 허용하고 싶습니다.

에서 작업을 수행할 권한이 없음 AWS Batch

에서 작업을 수행할 권한이 없다는 AWS Management Console 메시지가 표시되면 관리자에게 문의하여 지원을 받아야 합니다. 관리자는 사용자 이름과 암호를 제공한 사람입니다.

다음 예제 오류는 `mateojackson` 사용자가 콘솔을 사용하여 가상 *`my-example-widget`* 리소스에 대한 세부 정보를 보려고 하지만 가상 `batch:`*`GetWidget`* 권한이 없을 때 발생합니다.

```
User: arn:aws:iam::123456789012:user/mateojackson is not authorized to perform:
 batch:GetWidget on resource: my-example-widget
```

이 경우 Mateo는 *`my-example-widget`* 작업을 사용하여 `batch:`*`GetWidget`* 리소스에 액세스하도록 허용하는 정책을 업데이트하라고 관리자에게 요청합니다. 역할을 전달할 수 있는 권한 부여에 대한 자세한 내용은 AWS 서비스에 역할을 전달할 수 있는 사용자 권한 부여를 참조하세요.

iam:PassRole을 수행하도록 인증되지 않음

`iam:PassRole` 작업을 수행할 수 있는 권한이 없다는 오류가 수신되면 AWS Batch에 역할을 전달할 수 있도록 정책을 업데이트해야 합니다.

일부 AWS 서비스 에서는 새 서비스 역할 또는 서비스 연결 역할을 생성하는 대신 기존 역할을 해당 서비스에 전달할 수 있습니다. 이렇게 하려면 역할을 서비스에 전달할 권한이 있어야 합니다.

다음 예 오류는 `marymajor`라는 IAM 사용자가 콘솔을 사용하여 AWS Batch에서 작업을 수행하려고 하는 경우에 발생합니다. 하지만 작업을 수행하려면 서비스 역할이 부여한 권한이 서비스에 있어야 합니다. Mary는 서비스에 역할을 전달할 권한이 없습니다.

```
User: arn:aws:iam::123456789012:user/marymajor is not authorized to perform:
 iam:PassRole
```

이 경우, Mary가 `iam:PassRole`작업을 수행할 수 있도록 Mary의 정책을 업데이트해야 합니다.

도움이 필요한 경우 AWS 관리자에게 문의하세요. 관리자는 로그인 자격 증명을 제공한 사람입니다.

내 AWS 계정 외부의 사람이 내 AWS Batch 리소스에 액세스하도록 허용하고 싶습니다.

다른 계정의 사용자 또는 조직 외부의 사람이 리소스에 액세스할 때 사용할 수 있는 역할을 생성할 수 있습니다. 역할을 수임할 신뢰할 수 있는 사람을 지정할 수 있습니다. 리소스 기반 정책 또는 액세스 제어 목록(ACL)을 지원하는 서비스의 경우, 이러한 정책을 사용하여 다른 사람에게 리소스에 대한 액세스 권한을 부여할 수 있습니다.

자세한 내용은 다음을 참조하세요.

- 에서 이러한 기능을 AWS Batch 지원하는지 여부를 알아보려면 섹션을 참조하세요AWS Batch 에서 IAM을 사용하는 방법.
- 소유 AWS 계정 한의 리소스에 대한 액세스 권한을 제공하는 방법을 알아보려면 IAM 사용 설명서의 소유한 다른의 IAM 사용자에게 액세스 권한 제공을 참조 AWS 계정 하세요.
- 타사에 리소스에 대한 액세스 권한을 제공하는 방법을 알아보려면 IAM 사용 설명서의 타사가 AWS 계정 소유한에 대한 액세스 권한 제공을 AWS 계정참조하세요.
- ID 페더레이션을 통해 액세스 권한을 제공하는 방법을 알아보려면 IAM 사용 설명서의 외부에서 인증된 사용자에게 액세스 권한 제공(ID 페더레이션)을 참조하세요.
- 크로스 계정 액세스에 대한 역할과 리소스 기반 정책 사용의 차이점을 알아보려면 IAM 사용 설명서의 IAM의 크로스 계정 리소스 액세스를 참조하세요.

AWS Step Functions

AWS Batch 콘솔을 사용하여 Step Functions 상태 머신과 해당 머신에서 사용하는 함수에 대한 세부 정보를 볼 수 있습니다.

섹션

- 자습서: 상태 머신 세부 정보 보기
- 자습서: 상태 머신 편집
- 자습서: 상태 머신 실행

자습서: 상태 머신 세부 정보 보기

AWS Batch 콘솔에는 현재 AWS 리전에서 AWS Batch 작업을 제출하는 워크플로 단계가 하나 이상 포함된 상태 머신이 목록으로 표시됩니다.

상태 머신을 선택하여 워크플로의 그래픽 표현을 확인합니다. 파란색으로 강조 표시된 단계는 AWS Batch 작업을 나타냅니다. 그래프 컨트롤을 사용하여 그래프를 확대, 축소 및 가운데에 표시할 수 있습니다.

> **Note**
>
> AWS Batch 작업이 상태 머신 정의에서 JsonPath와 동적으로 참조되면 함수 세부 정보를 AWS Batch 콘솔에 표시할 수 없습니다. 대신 함수 이름이 동적 참조로 나열되고 그래프의 해당 단계가 회색으로 표시됩니다.

상태 머신 세부 정보를 보려면

1. Step Functions 페이지에 의해 구동되는 AWS Batch 콘솔 워크플로 오케스트레이션 페이지를 엽니다.
2. 상태 머신을 선택합니다.
   <result>

   AWS Batch 콘솔에 세부 정보 페이지가 열립니다.
   </result>

자세한 내용은 AWS Step Functions 개발자 안내서의 Step Functions를 참조하세요.

자습서: 상태 머신 편집

상태 머신을 편집하려는 경우 AWS Batch에 Step Functions 콘솔의 정의 편집 페이지가 열립니다.

상태 머신을 편집하려면

1. Step Functions 페이지에 의해 구동되는 AWS Batch 콘솔 워크플로 오케스트레이션 페이지를 엽니다.
2. 상태 머신을 선택합니다.
3. [Edit]를 선택합니다.

   Step Functions 콘솔에 정의 편집 페이지가 열립니다.
4. 상태 머신을 편집하고 저장을 선택합니다.

상태 머신 편집에 대한 자세한 내용은 AWS Step Functions 개발자 안내서의 Step Functions 상태 머신 언어를 참조하세요.

자습서: 상태 머신 실행

상태 머신을 실행하려는 경우 AWS Batch에 Step Functions 콘솔의 새 실행 페이지가 열립니다.

상태 머신을 실행하려면

1. Step Functions 페이지에 의해 구동되는 AWS Batch 콘솔 워크플로 오케스트레이션 페이지를 엽니다.
2. 상태 머신을 선택합니다.
3. 실행을 선택합니다.

   Step Functions 콘솔에 새 실행 페이지가 열립니다.
4. (선택 사항) 상태 머신을 편집하고 실행 시작을 선택합니다.

상태 머신 실행에 대한 자세한 내용은 AWS Step Functions 개발자 안내서의 Step Functions 상태 머신 실행 개념을 참조하세요.

Amazon EventBridge에 대한 AWS Batch 이벤트 스트림

Amazon EventBridge용 AWS Batch 이벤트 스트림을 사용하여 작업 대기열 내 작업의 현재 상태에 관한 실시간에 가까운 알림을 받을 수 있습니다.

EventBridge를 사용하면 AWS Batch 서비스에 대한 추가 통찰력을 얻을 수 있습니다. 보다 구체적으로는, 이를 사용하여 작업 진행 상황을 확인하고, AWS Batch 사용자 지정 워크플로를 구축하며, 사용 보고서 또는 지표를 생성하거나 자체 대시보드를 구축할 수 있습니다. AWS Batch 및 EventBridge가 있으면 작업 상태 변경에 대해 AWS Batch를 지속적으로 폴링하는 코드를 예약하거나 모니터링할 필요가 없습니다. 대신 다양한 Amazon EventBridge 대상을 사용하여 AWS Batch 작업 상태 변경을 비동기적으로 처리할 수 있습니다. 이는 AWS Lambda, Amazon Simple Queue Service, Amazon Simple Notification Service 또는 Amazon Kinesis Data Streams를 포함할 수 있습니다.

AWS Batch 이벤트 스트림의 이벤트는 최소 한 번 이상 전달됩니다. 중복 이벤트가 전송되는 경우, 해당 이벤트는 중복을 식별하기에 충분한 정보를 제공합니다. 이렇게 하면 이벤트의 타임스탬프와 작업 상태를 비교할 수 있습니다.

AWS Batch 작업은 EventBridge 대상으로 사용할 수 있습니다. 단순한 규칙을 사용하여 사용자는 이벤트를 매치하고 이에 대한 응답으로 AWS Batch 작업을 제출할 수 있습니다. 자세한 내용은 Amazon EventBridge 사용 설명서의 EventBridge란?을 참조하세요. 또한 cron 또는 rate 표현식을 사용하여 특정 시간에 자체 트리거되는 자동 작업을 EventBridge로 예약할 수 있습니다. 자세한 내용은 Amazon EventBridge 사용 설명서의 예약 운영 EventBridge 규칙 생성을 참조하세요. 예제 연습은 EventBridge 대상으로 사용되는 AWS Batch작업 섹션을 참조하세요. EventBridge Scheduler 사용에 대한 자세한 내용은 Amazon EventBridge 사용 설명서의 Amazon EventBridge Scheduler 설정을 참조하세요.

주제

- AWS Batch 이벤트
- 자습서:에서 AWS 사용자 알림 사용 AWS Batch
- EventBridge 대상으로 사용되는 AWS Batch작업
- 자습서: EventBridge를 사용하여 AWS Batch 작업 이벤트 수신
- 자습서: 실패한 작업 이벤트에 대한 Amazon Simple Notification Service 알림 보내기

AWS Batch 이벤트

AWS Batch 는 작업 상태 변경 이벤트를 EventBridge로 전송합니다.는 작업 상태를 AWS Batch 추적합니다. 이전에 제출한 작업의 상태가 변경되면 이벤트가 간접적으로 호출됩니다. `RUNNING` 상태에 있던

작업이 FAILED로 이동하는 경우를 예로 들 수 있습니다. 이러한 이벤트는 작업 상태 변경 이벤트로 분류됩니다.

> **Note**
>
> AWS Batch 는 향후 다른 이벤트 유형, 소스 및 세부 정보를 추가할 수 있습니다. 프로그래밍 방식으로 이벤트 JSON 데이터를 역직렬화하는 경우, 알 수 없는 속성을 처리할 수 있도록 애플리케이션이 준비되어야 합니다. 이는 이러한 추가 속성이 추가되는 경우 및 추가될 때 문제가 발생하지 않도록 하기 위한 것입니다.

작업 상태 변경 이벤트

기존(이전에 제출한) 작업의 상태가 변경될 때마다 이벤트가 생성됩니다. AWS Batch 작업 상태에 대한 자세한 내용은 작업 상태 섹션을 참조하세요.

> **Note**
>
> 첫 작업 제출에 대한 이벤트는 생성되지 않습니다.

Example작업 상태 변경 이벤트

작업 상태 변경 이벤트는 다음 형식으로 제공됩니다. detail 섹션은 AWS Batch API 참조의 DescribeJobs API 작업에서 반환되는 JobDetail 객체와 유사합니다. EventBridge 파라미터에 대한 자세한 정보는 Amazon EventBridge 사용 설명서의 이벤트 및 이벤트 패턴을 참조하세요.

```
{
    "version": "0",
    "id": "c8f9c4b5-76e5-d76a-f980-7011e206042b",
    "detail-type": "Batch Job State Change",
    "source": "aws.batch",
    "account": "123456789012",
    "time": "2022-01-11T23:36:40Z",
    "region": "us-east-1",
    "resources": [
        "arn:aws:batch:us-east-1:123456789012:job/4c7599ae-0a82-49aa-ba5a-4727fcce14a8"
    ],
    "detail": {
        "jobArn": "arn:aws:batch:us-east-1:123456789012:job/4c7599ae-0a82-49aa-
ba5a-4727fcce14a8",
```

```
        "jobName": "event-test",
        "jobId": "4c7599ae-0a82-49aa-ba5a-4727fcce14a8",
        "jobQueue": "arn:aws:batch:us-east-1:123456789012:job-queue/
PexjEHappyPathCanary2JobQueue",
        "status": "RUNNABLE",
        "attempts": [],
        "createdAt": 1641944200058,
        "retryStrategy": {
            "attempts": 2,
            "evaluateOnExit": []
        },
        "dependsOn": [],
        "jobDefinition": "arn:aws:batch:us-east-1:123456789012:job-definition/first-
run-job-definition:1",
        "parameters": {},
        "container": {
            "image": "137112412989.dkr.ecr.us-east-1.amazonaws.com/amazonlinux:latest",
            "command": [
                "sleep",
                "600"
            ],
            "volumes": [],
            "environment": [],
            "mountPoints": [],
            "ulimits": [],
            "networkInterfaces": [],
            "resourceRequirements": [
                {
                    "value": "2",
                    "type": "VCPU"
                }, {
                    "value": "256",
                    "type": "MEMORY"
                }
            ],
            "secrets": []
        },
        "propagateTags": false,
        "platformCapabilities": []
    }
}
```

작업 대기열 차단 이벤트

가 RUNNABLE 상태에서 작업을 AWS Batch 감지하여 대기열을 차단할 때마다 Amazon CloudWatch Events에 이벤트가 생성됩니다. 지원되는 차단된 대기열 원인에 대한 자세한 내용은 RUNNABLE 상태에서 정체된 작업 섹션을 참조하세요. DescribeJobs API 작업의 statusReason 필드에서도 동일한 이유를 사용할 수 있습니다.

Example작업 대기열 차단 이벤트

작업 대기열 차단 이벤트는 다음과 같은 형식으로 제공됩니다. detail 섹션은 AWS Batch API 참조의 DescribeJobs API 작업에서 반환되는 JobDetail 객체와 유사합니다. EventBridge 파라미터에 대한 자세한 정보는 Amazon EventBridge 사용 설명서의 이벤트 및 이벤트 패턴을 참조하세요.

```
{
    "version": "0",
    "id": "c8f9c4b5-76e5-d76a-f980-7011e206042b",
    "detail-type": "Batch Job Queue Blocked",
    "source": "aws.batch",
    "account": "123456789012",
    "time": "2022-01-11T23:36:40Z",
    "region": "us-east-1",
    "resources": [
        "arn:aws:batch:us-east-1:123456789012:job/4c7599ae-0a82-49aa-
ba5a-4727fcce14a8",
        "arn:aws:batch:us-east-1:123456789012:job-queue/PexjEHappyPathCanary2JobQueue"
    ],
    "detail": {
        "jobArn": "arn:aws:batch:us-east-1:123456789012:job/4c7599ae-0a82-49aa-
ba5a-4727fcce14a8",
        "jobName": "event-test",
        "jobId": "4c7599ae-0a82-49aa-ba5a-4727fcce14a8",
        "jobQueue": "arn:aws:batch:us-east-1:123456789012:job-queue/
PexjEHappyPathCanary2JobQueue",
        "status": "RUNNABLE",
        "statusReason": "blocked-reason",
        "attempts": [],
        "createdAt": 1641944200058,
        "retryStrategy": {
            "attempts": 2,
            "evaluateOnExit": []
        },
        "dependsOn": [],
```

```
        "jobDefinition": "arn:aws:batch:us-east-1:123456789012:job-definition/first-
run-job-definition:1",
        "parameters": {},
        "container": {
            "image": "137112412989.dkr.ecr.us-east-1.amazonaws.com/amazonlinux:latest",
            "command": [
                "sleep",
                "600"
            ],
            "volumes": [],
            "environment": [],
            "mountPoints": [],
            "ulimits": [],
            "networkInterfaces": [],
            "resourceRequirements": [
                {
                    "value": "2",
                    "type": "VCPU"
                }, {
                    "value": "256",
                    "type": "MEMORY"
                }
            ],
            "secrets": []
        },
        "propagateTags": false,
        "platformCapabilities": []
    }
}
```


서비스 작업 상태 변경 이벤트

기존 서비스 작업의 상태가 변경될 때마다 이벤트가 생성됩니다. 서비스 작업 상태에 대한 자세한 내용은 AWS Batch 서비스 작업 상태를 SageMaker AI 상태로 매핑 섹션을 참조하세요.

> **Note**
>
> 첫 작업 제출에 대한 이벤트는 생성되지 않습니다.

Example서비스 작업 상태 변경 이벤트

작업 상태 변경 이벤트는 다음 형식으로 제공됩니다. `detail` 섹션은 AWS Batch API 레퍼런스의 DescribeServiceJob API 작업에서 반환되는 응답과 유사합니다. EventBridge 파라미터에 대한 자세한 정보는 Amazon EventBridge 사용 설명서의 이벤트 및 이벤트 패턴을 참조하세요.

> **Note**
>
> `tags` 및 `serviceRequestPayload` 필드는 이벤트 `detail`에 포함되지 않습니다.

```
{
  "version": "0",
  "id": "c8f9c4b5-76e5-d76a-f980-7011e206042b",
  "detail-type": "Batch Service Job State Change",
  "source": "aws.batch",
  "account": "123456789012",
  "time": "2022-01-11T23:36:40Z",
  "region": "us-east-1",
  "resources": [
    "arn:aws:batch:us-east-1:123456789012:service-job/4c7599ae-0a82-49aa-
ba5a-4727fcce14a8"
  ],
  "detail": {
    "attempts": [
      {
        "serviceResourceId": {
          "name": "TrainingJobArn",
          "value": "arn:aws:sagemaker:us-east-1:123456789012:training-job/AWSBatchmy-
training-job88b610a69aa8380ca5b0a7aba3f81cb8"
        },
        "startedAt": 1641944300058,
        "stoppedAt": 1641944400058,
        "statusReason": "Received status from SageMaker: Training job completed"
      }
    ],
    "createdAt": 1641944200058,
    "jobArn": "arn:aws:batch:us-east-1:123456789012:service-job/4c7599ae-0a82-49aa-
ba5a-4727fcce14a8",
    "jobId": "0bb17543-ece6-4480-b1a7-a556d344746b",
    "jobName": "event-test",
    "jobQueue": "arn:aws:batch:us-east-1:123456789012:job-queue/HappyPathJobQueue",
```

```
    "latestAttempt": {
      "serviceResourceId": {
        "name": "TrainingJobArn",
        "value": "arn:aws:sagemaker:us-east-1:123456789012:training-job/AWSBatchmy-
training-job88b610a69aa8380ca5b0a7aba3f81cb8"
      }
    },
    "serviceJobType": "SAGEMAKER_TRAINING",
    "startedAt": 1641944300058,
    "status": "SUCCEEDED",
    "statusReason": "Received status from SageMaker: Training job completed",
    "stoppedAt": 1641944400058,
    "timeoutConfig": {
      "attemptDurationSeconds": 60
    }
  }
}
```

서비스 작업 대기열 차단 이벤트

가 차단된 대기열을 AWS Batch 감지하면 Amazon CloudWatch Events에 이벤트가 생성됩니다. 차단된 대기열의 사유는 `DescribeServiceJob` API 작업의 `statusReason` 필드에서 찾을 수 있습니다.

Example서비스 작업 대기열 차단 이벤트

서비스 작업 대기열 차단 이벤트는 다음과 같은 형식으로 제공됩니다. `detail` 섹션은 AWS Batch API 레퍼런스의 DescribeServiceJob API 작업에서 반환되는 응답과 유사합니다. EventBridge 파라미터에 대한 자세한 정보는 Amazon EventBridge 사용 설명서의 이벤트 및 이벤트 패턴을 참조하세요.

> **Note**
>
> `tags` 및 `serviceRequestPayload` 필드는 이벤트 `detail`에 포함되지 않습니다.

```
{
  "version": "0",
  "id": "c8f9c4b5-76e5-d76a-f980-7011e206042b",
  "detail-type": "Batch Service Job Queue Blocked",
  "source": "aws.batch",
  "account": "123456789012",
  "time": "2022-01-11T23:36:40Z",
  "region": "us-east-1",
```

```
  "resources": [
    "arn:aws:batch:us-east-1:123456789012:service-job/4c7599ae-0a82-49aa-
ba5a-4727fcce14a8"
  ],
  "detail": {
    "attempts": [],
    "createdAt": 1641944200058,
    "jobArn": "arn:aws:batch:us-east-1:123456789012:service-job/4c7599ae-0a82-49aa-
ba5a-4727fcce14a8",
    "jobId": "6271dfdf-d8a7-41b1-a4d2-55a2224f5375",
    "jobName": "event-test",
    "jobQueue": "arn:aws:batch:us-east-1:123456789012:job-queue/HappyPathJobQueue",
    "serviceJobType": "SAGEMAKER_TRAINING",
    "status": "RUNNABLE",
    "statusReason": "blocked-reason",
    "timeoutConfig": {
      "attemptDurationSeconds": 60
    }
  }
}
```

서비스 작업 선점 이벤트

선점 이벤트는 선점 시퀀스가 시작되거나 완료될 때마다 게시됩니다. 시작된 이벤트는 대여된 용량을 회수하기 위해 선점이 처음 트리거될 때 생성됩니다. 선점 시퀀스가 완료되고 선점된 시도의 요약이 포함되면 완료된 이벤트가 생성됩니다.

Example서비스 작업 선점 시작 이벤트

서비스 작업 선점 시작 이벤트는 다음 형식으로 전달됩니다. EventBridge 파라미터에 대한 자세한 정보는 Amazon EventBridge 사용 설명서의 이벤트 및 이벤트 패턴을 참조하세요.

```
{
  "version": "0",
  "id": "1b4a511e-2737-226a-a1e7-fc97f1cd9681",
  "detail-type": "Batch Service Job Preemption Started",
  "source": "aws.batch",
  "account": "123456789012",
  "time": "2026-03-16T19:57:27Z",
  "region": "us-east-1",
  "resources": [
    "arn:aws:batch:us-east-1:123456789012:service-job/51f245a9-2995-4a53-
bced-7b3c00028f84"
```

```
  ],
  "detail": {
    "statusReason": "PREEMPTION_IN_PROGRESS: Cross-share preemption triggered to
 reclaim borrowed capacity",
    "preemptedJobArn": "arn:aws:batch:us-east-1:123456789012:service-
job/51f245a9-2995-4a53-bced-7b3c00028f84",
    "status": "RUNNING"
  }
}
```

Example서비스 작업 선점 완료 이벤트

서비스 작업 선점 완료 이벤트는 다음 형식으로 전달됩니다. preemptionSummary 필드에는 개수 및 가장 최근의 선점된 시도 정보를 포함하여 선점된 시도에 대한 세부 정보가 포함됩니다.

```
{
  "version": "0",
  "id": "2b1c6151-c166-edf5-822c-211b3bfd46b2",
  "detail-type": "Batch Service Job Preemption Completed",
  "source": "aws.batch",
  "account": "123456789012",
  "time": "2026-03-16T19:57:47Z",
  "region": "us-east-1",
  "resources": [
    "arn:aws:batch:us-east-1:123456789012:service-job/51f245a9-2995-4a53-
bced-7b3c00028f84"
  ],
  "detail": {
    "preemptedJobArn": "arn:aws:batch:us-east-1:123456789012:service-
job/51f245a9-2995-4a53-bced-7b3c00028f84",
    "preemptionSummary": {
      "preemptedAttemptCount": 1,
      "recentPreemptedAttempts": [
        {
          "serviceResourceId": {
            "name": "TrainingJobArn",
            "value": "arn:aws:sagemaker:us-east-1:123456789012:training-job/AWSBatchqm-
training-job9b2f08f911cf3dd794c1b3e72ae7ca5f"
          },
          "startedAt": 1773690923359,
          "stoppedAt": 1773691064669,
          "statusReason": "Cross-share preemption triggered to reclaim borrowed
 capacity"
```

```
        }
      ]
    },
    "status": "RUNNABLE"
  }
}
```

자습서:에서 AWS 사용자 알림 사용 AWS Batch

AWS 사용자 알림을 사용하여 AWS Batch 이벤트에 대한 알림을 받을 전송 채널을 설정할 수 있습니다. 이벤트가 지정한 규칙과 일치하면 알림을 받습니다. 이메일, 채팅 애플리케이션의 Amazon Q Developer 채팅 알림 또는 AWS Console Mobile Application 푸시 알림을 비롯한 여러 채널을 통해 이벤트에 대한 알림을 받을 수 있습니다. 콘솔 알림 센터에서도 알림을 볼 수 있습니다. 사용자 알림은 집계를 지원하므로 특정 이벤트 중에 받는 알림 수를 줄일 수 있습니다.

에서 사용자 알림을 구성하려면 AWS Batch:

1. AWS Batch 콘솔을 엽니다.
2. 대시보드를 선택합니다.
3. 알림 구성을 선택합니다.
4. AWS 사용자 알림에서 알림 구성 생성을 선택합니다.

사용자 알림을 구성하고 보는 방법에 대한 자세한 내용은 AWS 사용자 알림 시작하기를 참조하세요.

EventBridge 대상으로 사용되는 AWS Batch작업

Amazon EventBridge는 거의 실시간으로 Amazon Web Services 리소스의 변경 사항을 제공합니다. 일반적으로 아Amazon Elastic Container Service, Amazon Elastic Kubernetes Service의 AWS Batch과 AWS Fargate 작업은 EventBridge 대상이 될 수 있습니다. 단순한 규칙을 사용하여 사용자는 이벤트를 매치하고 이에 대한 응답으로 AWS Batch 작업을 제출할 수 있습니다. 자세한 내용은 Amazon EventBridge 사용 설명서의 EventBridge란?을 참조하세요.

또한 EventBridge로 cron 또는 rate 표현식을 사용하여 특정 시간에 간접 호출되는 자동화 작업을 예약할 수 있습니다. 자세한 내용은 Amazon EventBridge 사용 설명서의 예약 운영 EventBridge 규칙 생성을 참조하세요.

이벤트가 이벤트 패턴과 일치하는 경우에 운영되는 규칙 생성에 대한 자세한 내용은 Amazon EventBridge 사용 설명서 이벤트 대응 Amazon EventBridge 규칙 생성하기를 참조하세요.

EventBridge 대상 AWS Batch 작업의 일반 사용 사례에는 다음과 같은 사용 사례가 포함됩니다.

- 예약된 작업은 일정한 시간 간격으로 발생합니다. 예를 들어 Amazon EC2 스팟 인스턴스 비용이 더 저렴한 사용량이 적은 시간에만 cron 작업이 발생합니다.
- AWS Batch 작업은 CloudTrail에 기록된 API 작업에 대응하여 실행됩니다. 예를 들어, 객체가 지정된 Amazon S3 버킷에 업로드될 때마다 작업이 제출됩니다. 이러한 상황이 발생할 때마다 EventBridge 입력 트랜스포머는 객체의 버킷과 키 이름을 AWS Batch 파라미터에 전달합니다.

> **Note**
>
> 이 시나리오에서 모든 관련 AWS 리소스는 동일한 리전에 있어야 합니다. 여기에는 Amazon S3 버킷, EventBridge 규칙, CloudTrail 로그와 같은 리소스가 포함됩니다.

EventBridge 규칙 및 대상과 함께 AWS Batch 작업을 제출하기 전에 EventBridge 서비스는 AWS Batch 작업을 실행할 수 있는 몇 가지 권한이 필요합니다. EventBridge 콘솔에서 AWS Batch 작업을 대상으로 지정하는 규칙을 생성할 때 사용자는 그 역할도 생성할 수 있습니다. 이 역할에 필요한 서비스 보안 주체 및 IAM 권한에 대한 자세한 내용은 EventBridge IAM 역할 섹션을 참조하세요.

주제

- 자습서: 예약된 AWS Batch 작업 생성
- 자습서: 이벤트 패턴이 있는 규칙 생성
- 자습서: EventBridge 입력 변환기를 사용하여 일정에 따라 AWS Batch 대상에 이벤트 정보 전달

자습서: 예약된 AWS Batch 작업 생성

다음은 예약된 AWS Batch 작업과 필요한 EventBridge IAM 역할을 생성하는 방법을 소개합니다.

EventBridge를 사용하여 예약된 AWS Batch 작업을 생성하려면

> **Note**
>
> 이 절차는 Amazon ECS, Amazon EKS 및 AWS Fargate의 모든 AWS Batch 작업에 적용됩니다.

1. Amazon EventBridge 콘솔(https://console.aws.amazon.com/events/)을 엽니다.

2. 탐색 모음에서 사용할 AWS 리전(을)를 선택합니다.
3. 탐색 창에서 규칙를 선택합니다.
4. 규칙 생성을 선택합니다.
5. 이름에 해당 컴퓨팅 환경의 고유한 이름을 지정합니다. 이름은 최대 64자를 포함할 수 있습니다. 대문자 및 소문자, 숫자, 하이픈(-) 및 밑줄(_)을 포함할 수 있습니다.

> **Note**
>
> 규칙은 동일한 지역과 동일한 이벤트 버스의 다른 규칙과 동일한 이름을 가질 수 없습니다.

6. (선택 사항) 설명에서 규칙에 대한 설명을 입력합니다.
7. 이벤트 버스에서 이 규칙과 연결할 이벤트 버스를 선택합니다. 이 규칙이 자신의 계정에서 발생하는 이벤트와 일치하도록 하려면 기본을 선택합니다 사용자 계정의 AWS 서비스가 이벤트를 내보내면 그 이벤트는 계정의 기본 이벤트 버스로 이동합니다.
8. (선택 사항) 규칙을 즉시 실행하지 않으려면 선택한 버스의 규칙을 해제하십시오.
9. 규칙 유형에서 스케줄을 선택합니다.
10. 계속해서 규칙 생성하기 또는 다음을 선택합니다.
11. 스케줄 패턴에서 다음을 수행합니다.

    - 오전 8시와 같이 특정 시간에 실행되는 세분화된 일정을 선택합니다. 매월 첫째 월요일의 PST와 cron 표현식을 입력합니다. 자세한 내용은 Amazon EventBridge 사용 설명서의 Cron 표현식을 참조하세요.
    - 10분마다와 같이 일반 속도로 실행되는 일정을 선택하고 rate 표현식을 입력합니다.

12. 다음을 선택합니다.
13. 대상 유형에서 AWS 서비스를 선택합니다.
14. 대상 선택에서 배치 작업 대기열을 선택합니다. 이후 다음 항목을 구성합니다.

    - Job queue(작업 대기열): 작업을 예약할 작업 대기열의 Amazon 리소스 이름(ARN)을 입력합니다.
    - Job definition(작업 정의): 작업에 사용할 작업 정의의 이름과 개정 또는 전체 ARN을 입력합니다.
    - Job name(작업 이름): 작업의 이름을 입력합니다.
    - Array size(배열 크기): (선택 사항) 두 개 이상의 복사본을 실행할 수 있도록 작업의 배열 크기를 입력합니다. 자세한 내용은 배열 작업 섹션을 참조하세요.

- Job attempts(작업 시도): (선택 사항) 작업이 실패할 경우 작업을 다시 시도할 최대 횟수를 입력합니다. 자세한 내용은 작업 자동 재시도 섹션을 참조하세요.

15. Batch job queue(배치 작업 대기열) 대상 유형에 대해 EventBridge에서는 대상에 이벤트를 보낼 권한이 필요합니다. EventBridge는 규칙 실행에 필요한 IAM 역할을 생성할 수 있습니다. 다음 중 하나를 수행하세요.

    - IAM 역할을 자동으로 생성하려면 이 특정 리소스에 대해 새 역할 생성을 선택합니다.
    - 이미 생성한 IAM 역할을 사용하려면 기존 역할 사용을 선택합니다.

16. (선택 사항) 추가 설정(Additional settings)을 확장합니다.

    a. 대상 입력 구성의 경우 이벤트의 텍스트가 대상으로 전달되기 전에 처리되는 방식을 선택합니다.

    b. 최대 이벤트 수명에서 처리되지 않은 이벤트가 보관되는 시간 간격을 지정합니다.

    c. 재시도 횟수에는 이벤트 재시도 횟수를 입력합니다.

    d. DLQ(Dead Letter Queue)에서 처리되지 않은 이벤트 처리 방법에 대한 옵션을 선택합니다. 필요한 경우 DLQ(Dead Letter Queue) 대기열로 사용할 Amazon SQS 대기열을 지정합니다.

17. (선택 사항) 이 규칙에 다른 대상을 추가하려면 다른 대상 추가를 선택합니다.
18. 다음을 선택합니다.
19. (선택 사항) 태그에서 새 태그 추가를 선택하여 규칙의 리소스 레이블을 추가합니다. 자세한 내용은 Amazon EventBridge 태그를 참조하세요.
20. 다음을 선택합니다.
21. 검토 및 생성에서 구성 단계를 검토하십시오. 변경해야 하는 경우 편집을 선택합니다 작업을 마쳤으면 규칙 생성을 선택합니다.

이벤트 규칙을 생성하는 방법에 대한 자세한 내용은 Amazon EventBridge 사용 설명서의 예약 운영 EventBridge 규칙 생성을 참조하세요.

자습서: 이벤트 패턴이 있는 규칙 생성

다음 절차는 이벤트 패턴이 있는 규칙을 생성하는 방법입니다.

이벤트가 정의된 패턴과 일치할 때 이벤트를 대상으로 보내는 규칙을 생성하려면

> **Note**
>
> 이 절차는 Amazon ECS, Amazon EKS 및 AWS Fargate의 모든 AWS Batch 작업에 적용됩니다.

1. Amazon EventBridge 콘솔(https://console.aws.amazon.com/events/)을 엽니다.
2. 탐색 모음에서 사용할 AWS 리전(을)를 선택합니다.
3. 탐색 창에서 규칙를 선택합니다.
4. 규칙 생성을 선택합니다.
5. 이름에 해당 컴퓨팅 환경의 고유한 이름을 지정합니다. 이름은 최대 64자를 포함할 수 있습니다. 대문자 및 소문자, 숫자, 하이픈(-) 및 밑줄(_)을 포함할 수 있습니다.

   > **Note**
   >
   > 규칙은 동일한 지역과 동일한 이벤트 버스의 다른 규칙과 동일한 이름을 가질 수 없습니다.

6. (선택 사항) 설명에서 규칙에 대한 설명을 입력합니다.
7. 이벤트 버스에서 이 규칙과 연결할 이벤트 버스를 선택합니다. 이 규칙이 자신의 계정에서 발생하는 이벤트와 일치하도록 하려면 기본을 선택합니다 사용자 계정의 AWS 서비스가 이벤트를 내보내면 그 이벤트는 계정의 기본 이벤트 버스로 이동합니다.
8. (선택 사항) 규칙을 즉시 실행하지 않으려면 선택한 버스의 규칙을 해제하십시오.
9. 규칙 유형(Rule type)에서 이벤트 패턴이 있는 규칙(Rule with an event pattern)을 생성합니다.
10. [Next]를 선택합니다.
11. 이벤트 소스의 경우, AWS 이벤트 또는 EventBridge 파트너 이벤트를 선택합니다.
12. (선택 사항) 샘플 이벤트의 경우:

    a. 샘플 이벤트 유형에서 AWS 이벤트를 선택합니다.

    b. 샘플 이벤트에서 배치 작업 상태 변경을 선택합니다.

13. 생성 방법에서 패턴 양식 사용을 선택합니다.
14. 이벤트 패턴의 경우:

    a. 이벤트 소스에서 AWS 서비스를 선택합니다.

b. AWS 서비스에서 배치를 선택합니다.

c. 이벤트 유형에서 배치 잡 상태 변경을 선택합니다.

15. 다음을 선택합니다.
16. 대상 유형에서 AWS 서비스를 선택합니다.
17. 대상 유형 선택에서 대상 유형을 선택합니다. 예를 들어, 배치 작업 대기열을 선택합니다. 다음 사항을 지정합니다.

    - Job queue(작업 대기열): 작업을 예약할 작업 대기열의 Amazon 리소스 이름(ARN)을 입력합니다.
    - Job definition(작업 정의): 작업에 사용할 작업 정의의 이름과 개정 또는 전체 ARN을 입력합니다.
    - Job name(작업 이름): 작업의 이름을 입력합니다.
    - Array size(배열 크기): (선택 사항) 두 개 이상의 복사본을 실행할 수 있도록 작업의 배열 크기를 입력합니다. 자세한 내용은 배열 작업 섹션을 참조하세요.
    - Job attempts(작업 시도): (선택 사항) 작업이 실패할 경우 작업을 다시 시도할 최대 횟수를 입력합니다. 자세한 내용은 작업 자동 재시도 섹션을 참조하세요.

18. Batch job queue(배치 작업 대기열) 대상 유형에 대해 EventBridge에서는 대상에 이벤트를 보낼 권한이 필요합니다. EventBridge는 규칙 실행에 필요한 IAM 역할을 생성할 수 있습니다. 다음 중 하나를 수행하세요.

    - IAM 역할을 자동으로 생성하려면 이 특정 리소스에 대해 새 역할 생성을 선택합니다.
    - 이전에 생성한 IAM 역할을 사용하려면 기존 역할 사용을 선택합니다.

19. (선택 사항) 추가 설정(Additional settings)을 확장합니다.

    a. 대상 입력 구성에서 이벤트의 텍스트 처리 방법을 선택합니다.

    b. 최대 이벤트 수명에서 처리되지 않은 이벤트가 보관되는 시간 간격을 지정합니다.

    c. 재시도 횟수에는 이벤트 재시도 횟수를 입력합니다.

    d. DLQ(Dead Letter Queue)에서 처리되지 않은 이벤트 처리 방법에 대한 옵션을 선택합니다. 필요한 경우 DLQ(Dead Letter Queue) 대기열로 사용할 Amazon SQS 대기열을 지정합니다.

20. (선택 사항) 다른 대상을 추가하려면 다른 대상 추가를 선택합니다.
21. 다음을 선택합니다.
22. (선택 사항) 태그에서 새 태그 추가를 선택하여 리소스 레이블을 추가합니다. 자세한 내용은 Amazon EventBridge 사용 설명서의 Amazon EventBridge 태그를 참조하세요.
23. 다음을 선택합니다.

24. 검토 및 생성에서 구성 단계를 검토하십시오. 변경해야 하는 경우 편집을 선택합니다 작업을 마친 다음 규칙 생성을 선택합니다.

    규칙 생성에 관한 자세한 내용은 Amazon EventBridge 사용 설명서의 이벤트 대응 Amazon EventBridge 규칙 생성을 참조하세요.

자습서: EventBridge 입력 변환기를 사용하여 일정에 따라 AWS Batch 대상에 이벤트 정보 전달

EventBridge 입력 변환기를 사용하여 작업 제출 AWS Batch 시에 이벤트 정보를 전달할 수 있습니다. 이는 다른 AWS 이벤트 정보의 결과로 작업을 호출하는 경우 특히 유용할 수 있습니다. Amazon S3 버킷에 객체를 업로드하는 경우를 예로 들 수 있습니다. 컨테이너 명령에 파라미터 대체 값이 포함된 작업 정의를 사용할 수도 있습니다. EventBridge 입력 변환기는 이벤트 데이터를 기반으로 파라미터 값을 제공할 수 있습니다.

그런 다음 AWS Batch 이벤트에서 정보를 구문 분석하여 `parameters` 객체로 변환하는 이벤트 대상을 생성합니다. 작업이 실행될 때 트리거 이벤트의 파라미터는 작업 컨테이너의 명령으로 전달됩니다.

> **Note**
>
> 이 시나리오에서는 모든 AWS 리소스(예: Amazon S3 버킷, EventBridge 규칙 및 CloudTrail 로그)가 동일한 리전에 있어야 합니다.

입력 변환기를 사용하는 AWS Batch 대상을 생성하려면

1. Amazon EventBridge 콘솔(https://console.aws.amazon.com/events/)을 엽니다.
2. 탐색 모음에서 사용할 AWS 리전 를 선택합니다.
3. 탐색 창에서 규칙을 선택합니다.
4. 규칙 생성을 선택합니다.
5. 이름에 해당 컴퓨팅 환경의 고유한 이름을 지정합니다. 이름은 최대 64자를 포함할 수 있습니다. 대문자 및 소문자, 숫자, 하이픈(-) 및 밑줄(_)을 포함할 수 있습니다.

   > **Note**
   >
   > 규칙은 동일한 및 동일한 이벤트 버스의 다른 규칙 AWS 리전 과 동일한 이름을 가질 수 없습니다.

6. (선택 사항) 설명에서 규칙에 대한 설명을 입력합니다.
7. 이벤트 버스에서 이 규칙과 연결할 이벤트 버스를 선택합니다. 이 규칙이 자신의 계정에서 발생하는 이벤트와 일치하도록 하려면 기본을 선택합니다. 계정 AWS 서비스 의가 이벤트를 내보내면 항상 계정의 기본 이벤트 버스로 이동합니다.
8. (선택 사항) 규칙을 즉시 실행하지 않으려면 선택한 버스의 규칙을 해제하십시오.
9. 규칙 유형에서 스케줄을 선택합니다.
10. 계속해서 규칙 생성하기 또는 다음을 선택합니다.
11. 스케줄 패턴에서 다음을 수행합니다.

    - 오전 8시와 같이 특정 시간에 실행되는 세분화된 일정을 선택합니다. 매월 첫째 월요일의 PST와 cron 표현식을 입력합니다. 자세한 내용은 Amazon EventBridge 사용 설명서의 Cron 표현식을 참조하세요.
    - 10분마다와 같이 일반 속도로 실행되는 일정을 선택하고 rate 표현식을 입력합니다.

12. 다음을 선택합니다.
13. 대상 유형에서 AWS 서비스를 선택합니다.
14. 대상 선택에서 배치 작업 대기열을 선택합니다. 이후 다음 항목을 구성합니다.

    - Job queue(작업 대기열): 작업을 예약할 작업 대기열의 Amazon 리소스 이름(ARN)을 입력합니다.
    - Job definition(작업 정의): 작업에 사용할 작업 정의의 이름과 개정 또는 전체 ARN을 입력합니다.
    - Job name(작업 이름): 작업의 이름을 입력합니다.
    - Array size(배열 크기): (선택 사항) 두 개 이상의 복사본을 실행할 수 있도록 작업의 배열 크기를 입력합니다. 자세한 내용은 배열 작업 단원을 참조하십시오.
    - Job attempts(작업 시도): (선택 사항) 작업이 실패할 경우 작업을 다시 시도할 최대 횟수를 입력합니다. 자세한 내용은 작업 자동 재시도 단원을 참조하십시오.

15. Batch job queue(배치 작업 대기열) 대상 유형에 대해 EventBridge에서는 대상에 이벤트를 보낼 권한이 필요합니다. EventBridge는 규칙 실행에 필요한 IAM 역할을 생성할 수 있습니다. 다음 중 하나를 수행하세요.

    - IAM 역할을 자동으로 생성하려면 이 특정 리소스에 대해 새 역할 생성을 선택합니다.
    - 이미 생성한 IAM 역할을 사용하려면 기존 역할 사용을 선택합니다.

16. (선택 사항) 추가 설정(Additional settings)을 확장합니다.

17. Additional settings(추가 설정) 섹션의 Configure target input(대상 입력 구성)에서 Input Transformer(입력 변환기)를 선택합니다.
18. Configure input transformer(입력 구성 변환기)를 선택합니다.
19. (선택 사항) 샘플 이벤트의 경우:
    a. 샘플 이벤트 유형에서 AWS 이벤트를 선택합니다.
    b. 샘플 이벤트에서 배치 작업 상태 변경을 선택합니다.
20. 대상 입력 변환기 섹션의 입력 경로에서 트리거 이벤트의 구문 분석 값을 지정합니다. 예를 들어, 배치 작업 상태 변경 이벤트를 파싱하려면 다음의 JSON 형식을 사용합니다.

```
{
    "instance": "$.detail.jobId",
    "state": "$.detail.status"
}
```

21. 템플릿에는 다음 사항을 입력합니다.

```
{
    "instance": <jobId> ,
    "status": <status>
}
```

22. 확인을 선택합니다.
23. 최대 이벤트 수명에서 처리되지 않은 이벤트가 보관되는 시간 간격을 지정합니다.
24. 재시도 횟수에는 이벤트 재시도 횟수를 입력합니다.
25. DLQ(Dead Letter Queue)에서 처리되지 않은 이벤트 처리 방법에 대한 옵션을 선택합니다. 필요한 경우 DLQ(Dead Letter Queue) 대기열로 사용할 Amazon SQS 대기열을 지정합니다.
26. (선택 사항) 다른 대상을 추가하려면 다른 대상 추가를 선택합니다.
27. 다음을 선택합니다.
28. (선택 사항) 태그에서 새 태그 추가를 선택하여 리소스 레이블을 추가합니다. 자세한 내용은 Amazon EventBridge 사용 설명서의 Amazon EventBridge 태그를 참조하세요.
29. 다음을 선택합니다.
30. 검토 및 생성에서 구성 단계를 검토하십시오. 변경해야 하는 경우 편집을 선택합니다 작업을 마친 다음 규칙 생성을 선택합니다.

자습서: EventBridge를 사용하여 AWS Batch 작업 이벤트 수신

이 자습서에서는 AWS Batch 작업 이벤트를 수신 대기하고 CloudWatch Logs 로그 스트림으로 출력하는 간단한 AWS Lambda 함수를 설정합니다.

사전 조건

이 자습서에서는 작업을 수락할 준비가 된 작업 중인 컴퓨팅 환경과 작업 대기열이 있다고 가정합니다. 이벤트를 캡처할 실행 중인 컴퓨팅 환경 및 작업 대기열이 없는 경우 AWS Batch 자습서 시작하기의 단계에 따라 하나를 생성합니다. 이 자습서를 마치고 나면 이 작업 대기열에 작업을 제출하여 Lambda 함수가 올바로 구성되었는지 테스트할 수 있습니다.

주제

- 자습서: Lambda 함수 생성
- 자습서: 이벤트 규칙 등록
- 자습서: 구성 테스트

자습서: Lambda 함수 생성

이 절차에서는 AWS Batch 이벤트 스트림 메시지의 대상으로 사용할 간단한 Lambda 함수를 생성합니다.

대상 Lambda 함수를 생성하려면

1. https://console.aws.amazon.com/lambda/에서 AWS Lambda 콘솔을 엽니다.
2. 함수 생성과 새로 작성을 차례로 선택합니다.
3. 함수 이름에 batch-event-stream-handler를 입력합니다.
4. 런타임에서 Python 3.8을 선택합니다.
5. 함수 생성(Create function)을 선택합니다.
6. 함수 코드 섹션에서 다음 예제와 일치하도록 샘플 코드를 수정합니다.

```
import json


def lambda_handler(event, _context):
    # _context is not used
    del _context
```

```
    if event["source"] != "aws.batch":
        raise ValueError("Function only supports input from events with a source
type of: aws.batch")

    print(json.dumps(event))
```

다음은 AWS Batch에서 전송하는 이벤트를 인쇄하는 간단한 Python 3.8 함수입니다. 모든 설정이 올바르게 구성되면 이 자습서가 끝날 때 이 Lambda 함수와 연결된 CloudWatch Logs 로그 스트림에 이벤트 세부 정보가 표시됩니다.

7. 배포(Deploy)를 선택합니다.

자습서: 이벤트 규칙 등록

이 섹션에서 AWS Batch 리소스에서 나온 작업 이벤트를 캡처하는 EventBridge 이벤트 규칙을 생성합니다. 이 규칙은 규칙이 정의된 계정 내의 AWS Batch에서 나온 모든 이벤트를 캡처합니다. 작업 메시지 자체에 작업이 제출된 작업 대기열과 같은 이벤트 소스에 대한 정보가 포함됩니다. 이 정보를 사용하여 프로그래밍 방식으로 이벤트를 필터링하고 정렬할 수 있습니다.

> **Note**
>
> AWS Management Console(을)를 사용하여 이벤트 규칙을 만들 경우 콘솔이 Lambda 함수를 호출할 EventBridge 권한을 부여하는 데 필요한 IAM 권한을 자동으로 추가합니다. AWS CLI(을)를 사용하여 이벤트 규칙을 생성하는 경우 권한을 명시적으로 부여해야 합니다. 자세한 내용을 알아보려면 Amazon EventBridge 사용 설명서의 이벤트 및 이벤트 패턴을 참조하세요.

EventBridge 규칙을 생성하려면

1. https://console.aws.amazon.com/events/에서 Amazon EventBridge 콘솔을 엽니다.
2. 탐색 창에서 규칙을 선택합니다.
3. 규칙 생성을 선택합니다.
4. 규칙에 대해 이름과 설명을 입력하세요.

   규칙은 동일한 지역과 동일한 이벤트 버스의 다른 규칙과 동일한 이름을 가질 수 없습니다.
5. 이벤트 버스에서 이 규칙과 연결할 이벤트 버스를 선택합니다. 이 규칙이 자신의 계정에서 발생하는 이벤트와 일치하도록 하려면 AWS 기본 이벤트 버스(default event bus)를 선택합니다. 계정의 AWS 서비스가 이벤트를 출력하면 항상 계정의 기본 이벤트 버스로 이동합니다.

6. 규칙 유형(Rule type)에서 이벤트 패턴이 있는 규칙(Rule with an event pattern)을 생성합니다.
7. 다음을 선택합니다.
8. 이벤트 소스에서 기타를 선택합니다.
9. 이벤트 패턴에서 사용자 지정 패턴(JSON 편집기)을 선택합니다.
10. 다음 이벤트 패턴을 텍스트 영역에 붙여 넣습니다.

```
{
  "source": [
    "aws.batch"
  ]
}
```

이 규칙은 모든 AWS Batch 그룹과 모든 AWS Batch 이벤트에 적용됩니다. 또는 더 한정적인 규칙을 만들어 일부 결과를 필터링할 수 있습니다.

11. 다음을 선택합니다.
12. 대상 유형에서 AWS서비스를 선택합니다.
13. 대상 선택에서 Lambda 함수를 선택하고 Lambda 함수를 선택합니다.
14. (선택 사항)추가 설정에서 다음을 수행합니다.

    a. 최대 이벤트 기간(Maximum age of event)에 1분(00:01)에서 24시간(24:00) 사이의 값을 입력합니다.

    b. 재시도(Retry attempts)에 0에서 185 사이의 숫자를 입력합니다.

    c. 배달 못한 편지 대기열(Dead-letter queue)에서 표준 Amazon SQS 대기열을 배달 못한 편지 대기열로 사용할지를 선택합니다. 이벤트가 대상에 성공적으로 전달되지 않은 경우 EventBridge는 이 규칙과 일치하는 이벤트를 배달 못한 편지 대기열로 보냅니다. 다음 중 하나를 수행합니다.

    - 배달 못한 편지 대기열을 사용하지 않으려면 없음(None)을 선택합니다.
    - 현재 AWS 계정에서 DLQ(Dead Letter Queue)로 사용할 Amazon SQS 대기열 선택(Select an Amazon SQS queue in the current account to use as the dead-letter queue)을 선택하고 드롭다운에서 사용할 대기열을 선택합니다.
    - 다른 AWS 계정에서 배달 못한 편지 대기열로 사용할 Amazon SQS 대기열 선택을 선택한 다음, 사용할 대기열의 ARN을 입력합니다. 메시지를 보낼 수 있는 EventBridge 권한을 부여하는 리소스 기반 정책을 대기열에 연결해야 합니다. 자세한 정보는 Amazon EventBridge 사용 설명서의 DLQ(Dead Letter Queue)에 대한 권한 부여를 참조하세요.

15. 다음을 선택합니다.

16. (선택 사항)규칙에 대해 하나 이상의 태그를 입력하세요. 자세한 정보는 Amazon EventBridge 사용 설명서의 Amazon EventBridge 태그를 참조하세요.
17. 다음을 선택합니다.
18. 규칙의 세부 정보를 검토하고 규칙 생성을 선택합니다.

자습서: 구성 테스트

작업 대기열에 작업을 제출하여 EventBridge 구성을 테스트할 수 있습니다. 모두 제대로 구성된 경우 Lambda 함수가 트리거되고 해당 함수에 대한 CloudWatch Logs 로그 스트림에 이벤트 데이터를 씁니다.

구성을 테스트하려면

1. https://console.aws.amazon.com/batch/에서 AWS Batch 콘솔을 엽니다.
2. 새 AWS Batch 작업을 제출합니다. 자세한 내용은 자습서: 작업 제출 섹션을 참조하세요.
3. https://console.aws.amazon.com/cloudwatch/에서 CloudWatch 콘솔을 엽니다.
4. 탐색 창에서 로그를 선택하고 Lambda 함수의 로그 그룹을 선택합니다(예: /aws/lambda/*my-function*).
5. 이벤트 데이터를 보려면 로그 스트림을 선택합니다.

자습서: 실패한 작업 이벤트에 대한 Amazon Simple Notification Service 알림 보내기

이 자습서에서는 작업이 `FAILED` 상태로 변경된 작업 이벤트만 캡처하는 Amazon EventBridge 이벤트 규칙을 구성합니다. 이 자습서를 마치면 선택적으로 이 작업 대기열에 작업을 제출할 수도 있습니다. 이는 Amazon SNS 알림을 올바르게 구성했는지 테스트합니다.

사전 조건

이 자습서에서는 작업을 수락할 준비가 된 작업 중인 컴퓨팅 환경과 작업 대기열이 있다고 가정합니다. 이벤트를 캡처할 실행 중인 컴퓨팅 환경 및 작업 대기열이 없는 경우 AWS Batch 자습서 시작하기의 단계에 따라 하나를 생성합니다.

주제

- 자습서: Amazon SNS 주제 생성 및 구독

- 자습서: 이벤트 규칙 등록
- 자습서: 규칙 테스트
- 대체 규칙: 배치 작업 대기열 차단됨

자습서: Amazon SNS 주제 생성 및 구독

본 자습서를 위해 새 이벤트 규칙의 이벤트 대상으로 사용할 Amazon SNS 주제를 구성합니다.

Amazon SNS 주제를 생성하려면

1. https://console.aws.amazon.com/sns/v3/home에서 Amazon SNS 콘솔을 엽니다.
2. 주제(Topics), 주제 생성(Create topic)을 차례로 선택합니다.
3. 유형에서 표준을 선택합니다.
4. 주제 이름에 **`JobFailedAlert`**(을)를 입력하고 주제 생성을 선택합니다.
5. JobFailedAlert 화면에서 구독 생성을 선택합니다.
6. 프로토콜(Protocol)에서 이메일(Email)을 선택합니다.
7. 엔드포인트(Endpoint)에 현재 액세스 권한이 있는 이메일 주소를 입력하고 구독 생성(Create subscription)을 선택합니다.
8. 이메일 계정을 확인하고 구독 확인 이메일 메시지를 기다립니다. 메시지를 수신하면 구독 확인(Confirm subscription)을 선택합니다.

자습서: 이벤트 규칙 등록

다음으로 작업 실패 이벤트만 캡처하는 이벤트 규칙을 등록합니다.

EventBridge 규칙을 등록하려면

1. https://console.aws.amazon.com/events/에서 Amazon EventBridge 콘솔을 엽니다.
2. 탐색 창에서 규칙을 선택합니다.
3. 규칙 생성을 선택합니다.
4. 규칙에 대해 이름과 설명을 입력하세요.

   규칙은 동일한 지역과 동일한 이벤트 버스의 다른 규칙과 동일한 이름을 가질 수 없습니다.

5. 이벤트 버스에서 이 규칙과 연결할 이벤트 버스를 선택합니다. 이 규칙이 자신의 계정에서 발생하는 이벤트와 일치하도록 하려면 AWS 기본 이벤트 버스(default event bus)를 선택합니다. 계정의 AWS 서비스가 이벤트를 출력하면 항상 계정의 기본 이벤트 버스로 이동합니다.
6. 규칙 유형(Rule type)에서 이벤트 패턴이 있는 규칙(Rule with an event pattern)을 생성합니다.
7. 다음을 선택합니다.
8. 이벤트 소스에서 기타를 선택합니다.
9. 이벤트 패턴에서 사용자 지정 패턴(JSON 편집기)을 선택합니다.
10. 다음 이벤트 패턴을 텍스트 영역에 붙여 넣습니다.

```
{
  "detail-type": [
    "Batch Job State Change"
  ],
  "source": [
    "aws.batch"
  ],
  "detail": {
    "status": [
      "FAILED"
    ]
  }
}
```

이 코드는 작업 상태가 `FAILED`인 이벤트를 일치시키는 EventBridge 규칙을 정의합니다. 이벤트 패턴에 대한 자세한 내용은 Amazon EventBridge 사용 설명서에서 이벤트 및 이벤트 패턴을 참조하세요.

11. 다음을 선택합니다.
12. 대상 유형에서 AWS서비스를 선택합니다.
13. 대상 선택에서 SNS 주제를 선택하고, 주제에 대해 JobFailedAlert를 선택합니다.
14. (선택 사항)추가 설정에서 다음을 수행합니다.

    a. 최대 이벤트 기간(Maximum age of event)에 1분(00:01)에서 24시간(24:00) 사이의 값을 입력합니다.

    b. 재시도(Retry attempts)에 0에서 185 사이의 숫자를 입력합니다.

    c. 배달 못한 편지 대기열(Dead-letter queue)에서 표준 Amazon SQS 대기열을 배달 못한 편지 대기열로 사용할지를 선택합니다. 이벤트가 대상에 성공적으로 전달되지 않은 경우

EventBridge는 이 규칙과 일치하는 이벤트를 배달 못한 편지 대기열로 보냅니다. 다음 중 하나를 수행합니다.

- 배달 못한 편지 대기열을 사용하지 않으려면 없음(None)을 선택합니다.
- 현재 AWS 계정에서 DLQ(Dead Letter Queue)로 사용할 Amazon SQS 대기열 선택(Select an Amazon SQS queue in the current account to use as the dead-letter queue)을 선택하고 드롭다운에서 사용할 대기열을 선택합니다.
- 다른 AWS 계정에서 배달 못한 편지 대기열로 사용할 Amazon SQS 대기열 선택을 선택한 다음, 사용할 대기열의 ARN을 입력합니다. 메시지를 보낼 수 있는 EventBridge 권한을 부여하는 리소스 기반 정책을 대기열에 연결해야 합니다. 자세한 정보는 Amazon EventBridge 사용 설명서의 DLQ(Dead Letter Queue)에 대한 권한 부여를 참조하세요.

15. 다음을 선택합니다.
16. (선택 사항)규칙에 대해 하나 이상의 태그를 입력하세요. 자세한 정보는 Amazon EventBridge 사용 설명서의 Amazon EventBridge 태그를 참조하세요.
17. 다음을 선택합니다.
18. 규칙의 세부 정보를 검토하고 규칙 생성을 선택합니다.

자습서: 규칙 테스트

규칙을 테스트하려면 0이 아닌 종료 코드로 시작한 직후에 종료되는 작업을 제출합니다. 이벤트 규칙이 올바로 구성되었다면 몇 분 후에 이벤트 텍스트가 포함된 이메일 메시지를 수신할 것입니다.

규칙을 테스트하려면

1. https://console.aws.amazon.com/batch/에서 AWS Batch 콘솔을 엽니다.
2. 새 AWS Batch 작업을 제출합니다. 자세한 내용은 자습서: 작업 제출 섹션을 참조하세요. 작업 명령의 경우 이 명령을 대체하여 종료 코드가 1인 컨테이너를 종료합니다.

```
/bin/sh, -c, 'exit 1'
```

3. 이메일에서 작업 실패 알림에 대한 이메일 알림이 수신되었는지 확인합니다.

대체 규칙: 배치 작업 대기열 차단됨

배치 작업 대기열 차단됨을 모니터링하는 이벤트 규칙을 생성하려면 다음 대체 방법을 사용하여 이 자습서를 반복합니다.

1. 자습서: Amazon SNS 주제 생성 및 구독에서 주제 이름으로 *BlockedJobQueue*를 사용합니다.
2. 자습서: 이벤트 규칙 등록에서 JSON 편집기에서 다음 패턴을 사용합니다.

```
{
   "detail-type": [
     "Batch Job Queue Blocked"
   ],
   "source": [
     "aws.batch"
   ]
}
```

Elastic Fabric Adapter

Elastic Fabric Adapter(EFA)는 HPC(고성능 컴퓨팅) 애플리케이션을 가속화하는 네트워크 디바이스입니다. AWS Batch는 다음 조건이 충족될 경우 EFA를 사용하는 애플리케이션을 지원합니다.

- EFA를 지원하는 인스턴스 유형 목록은 Amazon EC2 사용 설명서의 지원되는 인스턴스 유형을 참조하세요.

  > **Tip**
  >
  > AWS 리전에서 EFA를 지원하는 인스턴스 유형 목록을 보려면 다음 명령을 실행합니다. 그런 다음 반환된 목록을 AWS Batch 콘솔의 사용 가능한 인스턴스 유형 목록과 상호 참조합니다.
  >
  > ```
  > $ aws ec2 describe-instance-types  --region us-east-1  --filters Name=network-info.efa-supported,Values=true  --query "InstanceTypes[*].[InstanceType]"  --output text | sort
  > ```

- EFA를 지원하는 운영 체제 목록은 지원되는 운영 체제를 참조하세요.
- AMI에는 EFA 드라이버가 로드되어 있습니다.
- EFA의 보안 그룹은 보안 그룹 자체 내의 모든 인바운드 및 아웃바운드 트래픽을 허용해야 합니다.
- EFA를 사용하는 모든 인스턴스는 동일한 클러스터 배치 그룹에 있어야 합니다.
- 작업 정의는 EFA 디바이스가 컨테이너로 전달될 수 있도록 `hostPath`가 `/dev/infiniband/uverbs0`으로 설정된 `devices` 멤버를 포함해야 합니다. `containerPath`가 지정된 경우에도 `/dev/infiniband/uverbs0`으로 설정되어야 합니다. `permissions`가 설정된 경우 `READ` | `WRITE` | `MKNOD`로 설정되어야 합니다.

  다중 노드 병렬 작업과 단일 노드 컨테이너 작업에 대한 LinuxParameters 멤버의 위치는 서로 다릅니다. 아래 예는 차이를 보여주지만 필수 값은 누락되었습니다.

  Example다중 노드 병렬 작업의 예

  ```
  {
    "jobDefinitionName": "EFA-MNP-JobDef",
    "type": "multinode",
    "nodeProperties": {
      ...
  ```

```
    "nodeRangeProperties": [
      {
        ...
        "container": {
          ...
          "linuxParameters": {
            "devices": [
              {
                "hostPath": "/dev/infiniband/uverbs0",
                "containerPath": "/dev/infiniband/uverbs0",
                "permissions": [
                    "READ", "WRITE", "MKNOD"
                ]
              },
            ],
          },
        },
      },
    ],
  },
}
```

Example단일 노드 컨테이너 작업의 예

```
{
  "jobDefinitionName": "EFA-Container-JobDef",
  "type": "container",
  ...
  "containerProperties": {
    ...
    "linuxParameters": {
      "devices": [
        {
          "hostPath": "/dev/infiniband/uverbs0",
        },
      ],
    },
  },
}
```

EFA에 대한 자세한 정보는 Amazon EC2 사용 설명서의 Elastic Fabric Adapter를 참조하세요.

모니터링 AWS Batch

모니터링은 AWS Batch 및 AWS 솔루션의 안정성, 가용성 및 성능을 유지하는 데 중요한 부분입니다.

다중 지점 장애가 발생할 경우 더 쉽게 디버깅할 수 있도록 AWS 솔루션의 모든 부분에서 모니터링 데이터를 수집하는 것이 좋습니다. 먼저 다음 질문에 답하는 모니터링 계획을 수립합니다. 어떻게 답해야 할지 잘 모르는 경우에도 계속해서 Amazon CloudWatch Logs을 사용하여 성능 기준을 설정할 수 있습니다.

- 모니터링의 목표
- 모니터링할 리소스
- 이러한 리소스를 모니터링하는 빈도
- 사용할 모니터링 도구
- 모니터링 작업을 수행할 사람
- 문제 발생 시 알려야 할 대상

다음 단계는 다양한 시간과 다양한 로드 조건에서 AWS Batch 성능을 측정하여 환경에서 정상 성능의 기준을 설정하는 것입니다. 모니터링할 때 현재 성능 데이터와 비교할 수 있도록 모니터링 데이터를 기록 AWS Batch해 둡니다. 이를 통해 정상적인 성능 패턴과 성능 이상을 식별하고 문제 해결 방법을 고안할 수 있습니다.

이 섹션의 주제는 AWS Batch로깅 및 모니터링을 시작하는 데 도움이 될 수 있습니다.

주제

- 에서 CloudWatch Logs 사용 AWS Batch
- AWS Batch CloudWatch Container Insights
- CloudWatch Logs를 사용하여 Amazon EKS 작업 AWS Batch 에서 모니터링
- AWS 계획된 상태 수명 주기 이벤트

에서 CloudWatch Logs 사용 AWS Batch

EC2 리소스에서 AWS Batch 작업을 구성하여 자세한 로그 정보 및 지표를 CloudWatch Logs로 보낼 수 있습니다. 이렇게 하면 한 곳에서 작업의 다양한 로그를 편리하게 볼 수 있습니다. CloudWatch

Logs에 대한 자세한 내용은 Amazon CloudWatch 사용 설명서의 Amazon CloudWatch Logs란 무엇입니까?를 참조하세요.

> **Note**
>
> 기본적으로 AWS Fargate 컨테이너에 대해 CloudWatch Logs가 켜져 있습니다.

CloudWatch Logs 로깅을 활성화하고 사용자 지정하려면 다음과 같은 일회성 구성 작업을 검토합니다.

- EC2 리소스를 기반으로 하는 AWS Batch 컴퓨팅 환경의 경우 `ecsInstanceRole` 역할에 IAM 정책을 추가합니다. 자세한 내용은 the section called “자습서: CloudWatch Logs IAM 정책 추가” 단원을 참조하십시오.
- 자세한 CloudWatch 모니터링이 포함된 Amazon EC2 시작 템플릿을 생성한 다음 AWS Batch 컴퓨팅 환경을 생성할 때 템플릿을 지정합니다. 기존 이미지에 CloudWatch 에이전트를 설치한 다음 AWS Batch 처음 실행 마법사에서 이미지를 지정할 수도 있습니다.
- (선택 사항) awslogs 드라이버를 구성합니다. EC2 및 Fargate 리소스 모두에서 기본 동작을 변경하는 파라미터를 추가할 수 있습니다. 자세한 내용은 the section called “awslogs 로그 드라이버 사용” 단원을 참조하십시오.

주제

- 자습서: CloudWatch Logs IAM 정책 추가
- CloudWatch 에이전트 설치 및 구성
- 자습서: CloudWatch Logs 보기

자습서: CloudWatch Logs IAM 정책 추가

작업이 로그 데이터 및 세부 지표를 CloudWatch Logs에 전송하려면 먼저 CloudWatch Logs API를 사용하는 IAM 정책을 생성해야 합니다. IAM 정책을 생성한 후 `ecsInstanceRole` 역할에 정책을 연결할 수 있습니다.

> **Note**
>
> `ECS-CloudWatchLogs` 정책이 `ecsInstanceRole` 역할에 연결되지 않은 경우에도 기본 지표를 CloudWatch Logs로 전송할 수 있습니다. 하지만 기본 지표에는 로그 데이터 또는 사용 가능한 디스크 공간과 같은 세부 지표가 포함되지 않습니다.

AWS Batch 컴퓨팅 환경은 Amazon EC2 리소스를 사용합니다. AWS Batch 처음 실행 마법사를 사용하여 컴퓨팅 환경을 생성할 때는 `ecsInstanceRole` 역할을 AWS Batch 생성하고 이를 사용하여 환경을 구성합니다.

처음 실행 마법사를 사용하지 않는 경우 AWS Command Line Interface 또는 AWS Batch API에서 컴퓨팅 환경을 생성할 때 `ecsInstanceRole` 역할을 지정할 수 있습니다. 자세한 내용은 AWS CLI 명령 레퍼런스 또는 AWS Batch API 참조를 참조하세요.

`ECS-CloudWatchLogs` IAM; 정책을 생성하려면

1. https://console.aws.amazon.com/iam/에서 IAM 콘솔을 엽니다.
2. 탐색 창에서 Policies를 선택합니다.
3. 정책 생성을 선택합니다.
4. JSON을 선택하고 다음 정책을 입력합니다.

   JSON

```
{
    "Version":"2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "logs:CreateLogGroup",
                "logs:CreateLogStream",
                "logs:PutLogEvents",
                "logs:DescribeLogStreams"
            ],
            "Resource": [
                "arn:aws:logs:*:*:*"
            ]
        }
    ]
}
```

5. 다음: 태그를 선택합니다.
6. (선택 사항) 태그 추가에서 태그 추가를 선택하여 정책에 태그를 추가합니다.
7. 다음: 검토를 선택합니다.

8. 정책 검토 페이지에서 이름에 **ECS-CloudWatchLogs**를 입력하고, 선택 사항인 설명을 입력합니다.
9. 정책 생성을 선택합니다.

ECS-CloudWatchLogs 정책을 **ecsInstanceRole**에 연결하려면

1. IAM 콘솔(https://console.aws.amazon.com/iam/)을 엽니다.
2. 탐색 창에서 역할을 선택합니다.
3. ecsInstanceRole을 선택합니다. 역할이 존재하지 않을 경우, Amazon ECS 인스턴스 역할의 절차를 따라 역할을 생성합니다.
4. 권한 추가를 선택하고 정책 연결을 선택합니다.
5. ECS-CloudWatchLogs 정책을 선택하고 정책 연결을 선택합니다.

CloudWatch 에이전트 설치 및 구성

CloudWatch 모니터링이 포함된 Amazon EC2 시작 템플릿을 생성할 수 있습니다. 자세한 내용은 Amazon EC2 사용 설명서의 시작 템플릿에서 인스턴스 시작 및 고급 세부 정보를 참조하세요.

기존 Amazon EC2 AMI에 CloudWatch 에이전트를 설치한 다음 AWS Batch 처음 실행 마법사에서 이미지를 지정할 수도 있습니다. 자세한 내용은 CloudWatch 에이전트 설치 및 AWS Batch 자습서 시작하기 섹션을 참조하세요.

> **Note**
>
> 시작 템플릿은 AWS Fargate 리소스에서 지원되지 않습니다.

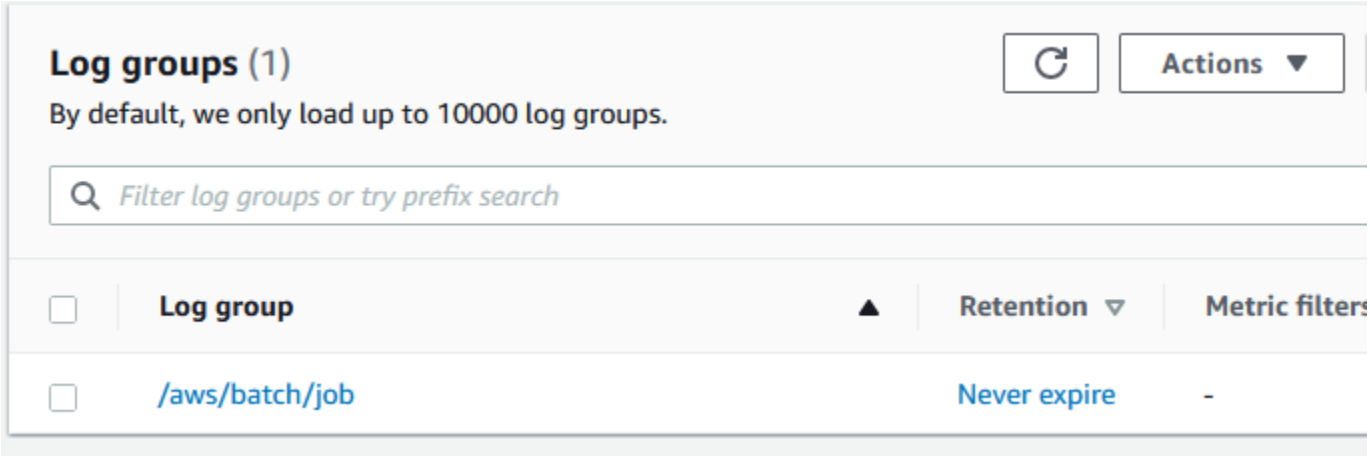
자습서: CloudWatch Logs 보기

에서 CloudWatch Logs 로그를 보고 검색할 수 있습니다 AWS Management Console.

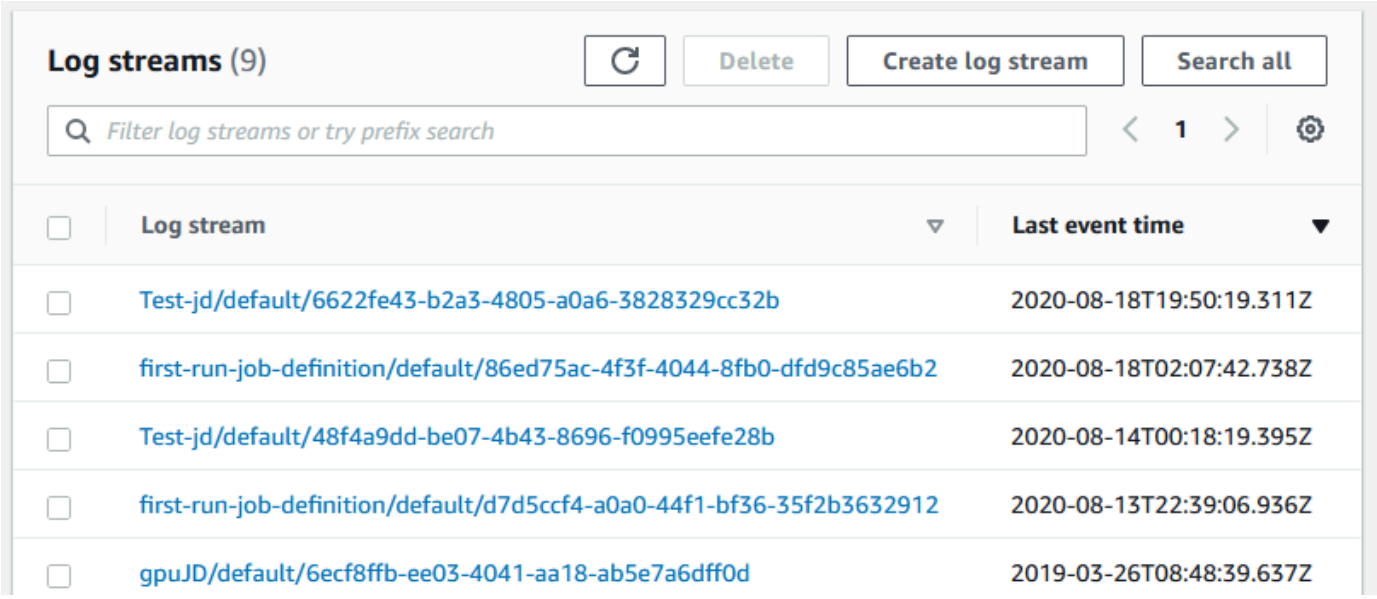
> **Note**
>
> CloudWatch Logs에 데이터가 표시되는 데 몇 분 정도 걸릴 수 있습니다.

CloudWatch Logs 데이터를 보려면

1. https://console.aws.amazon.com/cloudwatch/에서 CloudWatch 콘솔을 엽니다.
2. 왼쪽 탐색 창에서 로그를 선택한 다음, 로그 그룹을 선택합니다.

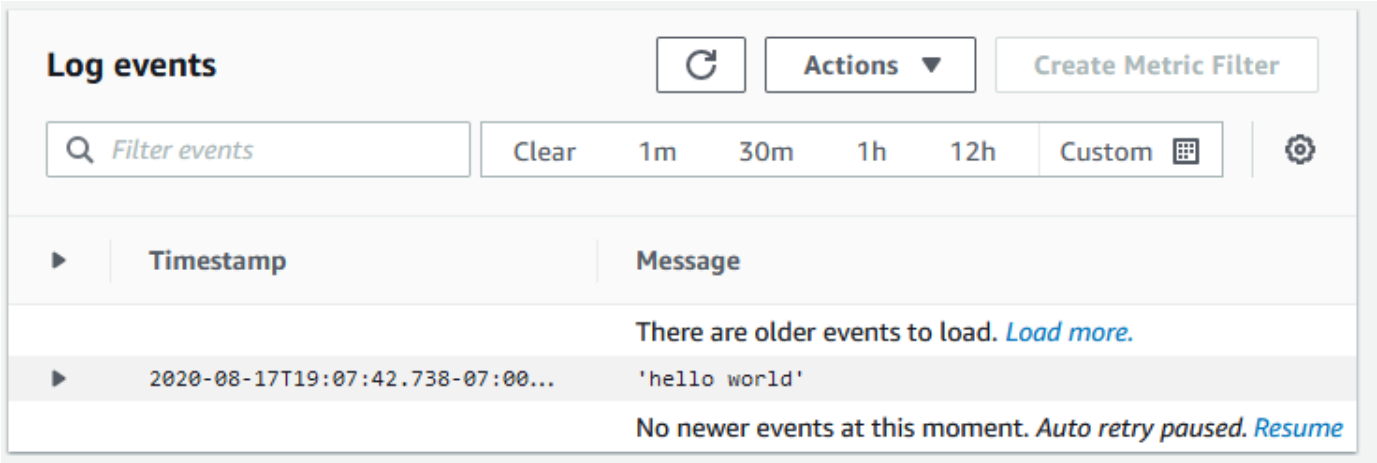


3. 보려는 로그 그룹을 선택합니다.



4. 보려는 로그 스트림을 선택합니다. 기본적으로 스트림은 작업 이름의 첫 200자와 Amazon ECS 작업 ID로 식별됩니다.

> **Tip**
>
> 로그 스트림 데이터를 다운로드하려면 작업을 선택합니다.



AWS Batch CloudWatch Container Insights

CloudWatch Container Insights는 AWS Batch 컴퓨팅 환경 및 작업에서 지표 및 로그를 수집하고, 종합하며, 요약합니다. 이 지표에는 CPU, 메모리, 디스크, 네트워크 사용률이 포함되어 있습니다. CloudWatch 대시보드에 이러한 지표를 추가할 수 있습니다.

운영 데이터는 성능 로그 이벤트로 수집됩니다. 이들은 정형 JSON 스키마를 사용하는 항목이기 때문에 카디널리티가 높은 데이터를 적정 규모로 수집 및 저장할 수 있습니다. 이러한 데이터를 토대로 CloudWatch는 CloudWatch 지표로서 컴퓨팅 환경 및 작업 수준에서 상위 수준의 집계 지표를 생성합니다. 자세한 내용은 Amazon CloudWatch 사용 설명서의 Amazon ECS용 Container Insights 구조화된 로그를 참조하세요.

> **Important**
>
> CloudWatch Container Insights는 사용자 지정 지표로 청구됩니다. 자세한 내용은 Amazon CloudWatch Events 요금을 참조하세요.

주제

- Container Insights를 설정합니다.

Container Insights를 설정합니다.

AWS Batch 컴퓨팅 환경에서 Container Insights를 활성화하려면 다음 단계를 완료합니다.

1. AWS Batch 콘솔을 엽니다.
2. 환경을 선택합니다.
3. 원하는 컴퓨터 환경을 선택합니다.
4. Container Insights 탭에서 컴퓨팅 환경에 대한 Container Insights를 켭니다.

> **Tip**
> 기본 간격을 선택하여 지표를 집계하거나 사용자 지정 간격을 만들 수 있습니다.

기본적으로 다음 지표가 표시됩니다. Amazon ECS Container Insights 지표의 전체 목록은 Amazon CloudWatch 사용 설명서의 Amazon ECS 컨테이너 지표를 참조하세요.

- **`JobCount`** - 컴퓨팅 환경에서 실행되는 작업 수입니다.
- **`ContainerInstanceCount`** - Amazon ECS 에이전트를 실행하고 컴퓨팅 환경에 등록된 Amazon Elastic Compute 클라우드 인스턴스의 수입니다.
- **`MemoryReserved`** - 컴퓨팅 환경 작업에 예약된 메모리입니다. 이 지표는 작업 정의에 정의된 메모리 예약이 있는 작업에 대해서만 수집됩니다.
- **`MemoryUtilized`** - 컴퓨팅 환경 작업에 사용되는 메모리입니다. 이 지표는 작업 정의에 정의된 메모리 예약이 있는 작업에 대해서만 수집됩니다.
- **`CpuReserved`** - 컴퓨팅 환경 작업에 예약된 CPU 유닛입니다. 이 지표는 작업 정의에 정의된 CPU 예약이 있는 작업에 대해서만 수집됩니다.
- **`CpuUtilized`** - 컴퓨팅 환경의 작업에 사용되는 CPU 유닛입니다. 이 지표는 작업 정의에 정의된 CPU 예약이 있는 작업에 대해서만 수집됩니다.
- **`NetworkRxBytes`** - 수신된 바이트 수입니다. 이 지표는 `awsvpc` 또는 브리지 네트워크 모드를 사용하는 작업의 컨테이너에 대해서만 제공됩니다.
- **`NetworkTxBytes`** - 전송된 바이트 수입니다. 이 지표는 `awsvpc` 또는 브리지 네트워크 모드를 사용하는 작업의 컨테이너에 대해서만 제공됩니다.
- **`StorageReadBytes`** - 스토리지에서 읽은 바이트 수입니다.
- **`StorageWriteBytes`** - 스토리지에 기록된 바이트 수입니다.

CloudWatch Logs를 사용하여 Amazon EKS 작업 AWS Batch 에서 모니터링

사용자는 Amazon CloudWatch Logs을 사용하여 한 위치에서 모든 로그 파일을 모니터링하고 저장 및 검색할 수 있습니다. CloudWatch Logs를 사용하면 여러 소스의 로그 데이터를 검색, 필터링 및 분석할 수 있습니다.

CloudWatch Logs AWS Batch 의 Amazon EKS 작업에서 모니터링할 플러그인이 포함된 AWS for Fluent Bit image를 다운로드할 수 있습니다. Fluent Bit는 Docker와 Kubernetes 호환되는 오픈 소스 로그 프로세서 및 전달자입니다. Fluentd보다 리소스 사용량이 적기 때문에 로그 라우터로 Fluent Bit를 사용하는 것이 좋습니다. 자세한 내용은 Amazon CloudWatch Observability EKS 추가 기능 또는 헬름 차트를 사용하여 CloudWatch 에이전트 설치를 참조하세요.

사전 조건

- `CloudWatchAgentServerPolicy` 작업자 노드의 AWS Identity and Access Management 정책에 정책을 연결합니다. 자세한 내용은 사전 조건 확인을 참조하세요.

추가 기능 설치

용 Fluent Bit를 설치하고 CloudWatch 그룹을 생성하는 방법에 AWS 대한 지침은 Amazon CloudWatch Observability EKS 추가 기능을 사용하여 CloudWatch 에이전트 설치 또는 차트 Helm을 Amazon CloudWatch 참조하세요.

추가 기능을 설치할 때는 다음과 같은 추가 구성 데이터를 제공해야 합니다.

- 를 사용하여 추가 기능을 설치하는 경우 구성 값에 다음 허용치를 제공해야 AWS Management Console 합니다.

```
{
  "tolerations": [
    {
      "key": "batch.amazonaws.com/batch-node",
      "operator": "Exists"
    }
  ]
}
```

- 를 사용하여 추가 기능을 설치하는 경우 다음 인수를 AWS CLI 추가합니다.

```
--configuration-values '{"tolerations":[{"key":"batch.amazonaws.com/batch-
node","operator":"Exists"}]}'
```

> **Tip**
>
> Fluent Bit는 AWS Batch 노드의 0.5 CPU와 100MB의 메모리를 사용한다는 점을 기억하세요. 이렇게 하면 AWS Batch 작업에 사용할 수 있는 총 용량이 줄어듭니다. 작업 규모를 조정할 때 이 점을 고려하세요.

AWS 계획된 상태 수명 주기 이벤트

AWS 상태는 예정된 변경 사항이 리소스에 AWS Batch 영향을 미칠 때 사전 알림을 제공합니다. 계획된 수명 주기 이벤트라고 하는 이러한 알림은 AMI 사용 중단, 운영 체제 end-of-support 및 조치가 필요한 인프라 업데이트와 같은 변경 사항을 알려줍니다. AWS Batch 는 AWS Health를 활용하여 마이그레이션을 계획하고 배치 처리 워크로드 중단을 방지할 수 있도록 이러한 변경 사항을 조기에 파악할 수 있도록 합니다.

AWS 상태에 대한 일반적인 정보는 AWS 상태란 무엇입니까?를 참조하십시오. AWS 상태 사용 설명서의 .

계획된 수명 주기 이벤트는 무엇입니까 AWS Batch?

AWS AMI end-of-support 및 조치를 취해야 하는 기타 인프라 변경과 같이 예정된 변경 사항이 AWS Batch 리소스에 영향을 미칠 때 상태는 계획된 수명 주기 이벤트 알림을 보냅니다.

계획된 수명 주기 이벤트의 특성은 다음과 같습니다.

- 유형 범주 - `scheduledChange`
- 이벤트 유형 코드 - 패턴을 따릅니다. `AWS_BATCH_PLANNED_LIFECYCLE_EVENT`
- 리드 타임 - 가능한 경우 주요 변경의 경우 최소 180일, 사소한 변경의 경우 90일입니다.
- 이벤트 시작 시간 - 리소스가 변경의 영향을 받을 수 있는 가장 빠른 날짜입니다.
- 동적 리소스 추적 - 영향을 받는 리소스는 `PENDING` 상태로 나열됩니다. 필요한 작업을 완료하거나 리소스를 삭제하면 상태가 로 업데이트됩니다`RESOLVED`.
- 범위 - 리소스에 영향을 미친 AWS 리전별로 그룹화된 계획된 각 수명 주기 이벤트에 대해 단일 이벤트 ARN을 받습니다.

> **Note**
>
> 리소스 상태 업데이트는 비동기식이며 현재 상태를 반영하는 데 최대 72시간이 걸릴 수 있습니다. end-of-support 이전에 영향을 받는 모든 리소스를 해결하면 AWS 상태 이벤트 상태가 로 변경됩니다`Closed`.

예: Amazon ECS Amazon Linux 2 AMI 사용 중단

Amazon ECS Amazon Linux 2 AMI 사용 중단은에 대해 계획된 수명 주기 이벤트의 예입니다 AWS Batch. Amazon Linux 2가 지원 종료 AWS 에 도달했으며 2026년 1월부터 새 Amazon ECS 컴퓨팅 환경의 기본 AMI를 Amazon Linux 2에서 Amazon Linux 2023으로 AWS Batch 변경합니다.

고객은 영향을 받는 컴퓨팅 환경을 식별한 AWS 상태 계획 수명 주기 이벤트 알림을 받았습니다. 영향을 받는 각 컴퓨팅 환경은 `PENDING` 상태로 나열되었습니다. 컴퓨팅 환경이 Amazon Linux 2023으로 마이그레이션된 후 상태가 로 업데이트되었습니다`RESOLVED`. 이를 통해 팀은 컴퓨팅 환경 플릿에서 마이그레이션 진행 상황을 추적할 수 있었습니다.

이 사용 중단에 대한 자세한 내용은 섹션을 참조하세요Amazon ECS Amazon Linux 2 AMI 사용 중단. 마이그레이션 단계는 섹션을 참조하세요ECS AL2에서 ECS AL2023으로 마이그레이션하는 방법.

계획된 수명 주기 이벤트 보기

AWS 상태 대시보드 AWS Batch 에서에 대해 계획된 수명 주기 이벤트를 볼 수 있습니다.

계획된 수명 주기 이벤트를 보려면

1. https://health.aws.amazon.com/health/home AWS 상태 대시보드를 엽니다.
2. 탐색 창에서 예약된 변경 사항을 선택합니다.
3. 배치 계획 수명 주기 이벤트를 찾습니다. 서비스별로 필터링하거나 일정 보기를 사용하여 월별 타임라인의 이벤트를 볼 수 있습니다.
4. 이벤트를 선택하여 세부 정보 및 영향을 받는 리소스 탭을 봅니다.

영향을 받는 리소스 탭에는 영향을 받는 각 리소스가 현재 상태와 함께 나열됩니다.

- 보류 중 - 리소스에 작업이 필요합니다.
- 해결됨 - 필수 작업이 완료되었거나 리소스가 삭제되었습니다.
- 알 수 없음 - 상태를 확인할 수 없습니다.

영향을 받는 리소스 목록을 CSV 또는 JSON 형식으로 다운로드할 수 있습니다. 계정이 AWS 조직의 일부인 경우 조직 보기에는 모든 멤버 계정의 영향을 받는 리소스가 표시됩니다.

일정 보기 프로젝트는 과거 최대 3개월, 향후 1년까지 변경되므로 그에 따라 유지 관리 기간을 계획할 수 있습니다.

Amazon EventBridge를 사용하여 계획된 수명 주기 이벤트 모니터링

Amazon EventBridge 규칙을 생성하여에 대해 계획된 AWS 상태 수명 주기 이벤트를 자동으로 감지하고 대응할 수 있습니다 AWS Batch. AWS Health는 이벤트를 EventBridge에 전송하고, 이러한 이벤트와 일치하는 규칙을 생성하여 AWS Lambda 함수, Amazon Simple Notification Service 주제, Amazon Simple Queue Service 대기열과 같은 대상으로 라우팅할 수 있습니다.

다음은 AWS 상태 계획 수명 주기 이벤트와 일치하는 이벤트 패턴의 예입니다 AWS Batch.

```
{
  "source": ["aws.health"],
  "detail-type": ["AWS Health Event"],
  "detail": {
    "service": ["BATCH"],
    "eventTypeCategory": ["scheduledChange"]
  }
}
```

계획된 수명 주기 이벤트가 있는 EventBridge 규칙의 일반적인 사용 사례는 다음과 같습니다.

- Amazon Simple Notification Service를 통해 팀 채팅 채널에 알림을 보냅니다.
- 새 이벤트가 감지되면 운영 티켓을 자동으로 생성합니다.
- AWS Lambda 함수를 호출하여 영향을 받는 리소스를 평가하고 마이그레이션 워크플로를 시작합니다.

AWS Health에 대한 EventBridge 규칙 구성에 대한 자세한 내용은 Health 사용 설명서의 Amazon EventBridge를 사용하여 Health 이벤트 모니터링을 AWS 참조하세요. AWS 샘플 자동화는 GitHub의 AWS Health Tools 리포지토리를 참조하세요.

AWS Batch 리소스 태깅

AWS Batch 리소스 관리를 돕기 위해 태그 형식으로 각 리소스에 고유한 메타데이터를 할당할 수 있습니다. 이 주제에서는 태그를 설명하고 태그를 생성하는 방법을 보여줍니다.

주제

태그 기본 사항

태그는 AWS 리소스에 할당하는 레이블입니다. 각 태그는 사용자가 정의하는 키와 선택적 값으로 구성됩니다.

태그를 사용하면 AWS 리소스를 용도, 소유자, 환경과 같은 다양한 기준으로 분류할 수 있습니다. 동일한 유형의 리소스가 많은 경우 할당한 태그에 따라 특정 리소스를 빠르게 식별할 수 있습니다. 예를 들어 AWS Batch 서비스에 태그 집합을 정의하면 각 서비스의 소유자 및 스택 수준을 추적하는 데 도움이 됩니다. 각 리소스 유형에 대해 일관된 태그 키 집합을 고안하는 것이 좋습니다.

태그가 리소스에 자동으로 할당되는 것은 아닙니다. 태그를 추가한 후에는 언제든지 태그 키와 값을 편집하거나 리소스에서 태그를 제거할 수 있습니다. 리소스를 삭제하면 리소스 태그도 삭제됩니다.

태그는 AWS Batch에는 의미가 없으며 엄격하게 문자열로 해석됩니다. 태그의 값을 빈 문자열로 설정할 수 있지만 태그의 값을 Null로 설정할 수는 없습니다. 해당 리소스에 대해 키가 기존 태그와 동일한 태그를 추가하는 경우 새 값이 이전 값을 덮어씁니다.

AWS Management Console, AWS CLI, AWS Batch API를 사용하여 태그 관련 작업을 수행할 수 있습니다.

AWS Identity and Access Management(IAM)를 사용하는 경우 AWS 계정에서 태그를 생성, 편집 또는 삭제할 수 있는 권한이 있는 사용자를 제어할 수 있습니다.

리소스에 태그 지정

신규 또는 기존 AWS Batch 컴퓨팅 환경, 작업, 작업 정의, 작업 대기열 및 예약 정책에 태그를 지정할 수 있습니다.

AWS Batch 콘솔을 사용하는 경우 새 리소스가 생성될 때 태그를 적용하거나 관련 리소스 페이지의 태그 탭을 사용하여 언제든지 기존 리소스에 태그를 적용할 수 있습니다.

AWS Batch API AWS CLI, 또는 AWS SDK를 사용하는 경우 관련 API 작업의 `tags` 파라미터를 사용하여 새 리소스에 태그를 적용하거나 API `TagResource` 작업을 사용하여 기존 리소스에 태그를 적용할 수 있습니다. 자세한 내용은 TagResource를 참조하세요.

일부 리소스 생성 작업에서는 리소스 생성 시 리소스에 태그를 지정할 수 있습니다. 리소스 생성 중에 태그를 적용할 수 없는 경우 리소스 생성 프로세스는 실패합니다. 이로써 생성 중에 태그를 지정하려는 리소스는 지정된 태그와 함께 생성되거나 전혀 생성되지 않습니다. 생성 시 리소스에 태그를 지정하면 리소스 생성 후 사용자 지정 태그 지정 스크립트를 실행할 필요가 없습니다.

다음 표에서는 태그를 지정할 수 있는 AWS Batch 리소스와 생성 시 태그를 지정할 수 있는 리소스를 설명합니다.

AWS Batch 리소스에 대한 태그 지원

Resource	태그 지원	태그 전달 지원	생성 시 태그 지정 지원(AWS Batch API, AWS CLI, AWS SDK)
AWS Batch 컴퓨팅 환경	예	아니요. 컴퓨팅 환경 태그는 다른 리소스로 전파되지 않습니다. 리소스 태그는 CreateComputeEnvironment API 작업에서 전달된 ComputeResources 객체의 태그 멤버에 지정됩니다.	예
AWS Batch 작업	예	예	예
AWS Batch 작업 정의	예	아니요	예

Resource	태그 지원	태그 전달 지원	생성 시 태그 지정 지원(AWS Batch API, AWS CLI, AWS SDK)
AWS Batch 작업 대기열	예	아니요	예
AWS Batch 예약 정책	예	아니요	예
AWS Batch 서비스 환경	예	아니요	예
AWS Batch 사용 가능한 리소스	예	아니요	예
AWS Batch 할당량 공유	예	아니요	예

태그 제한

태그에 적용되는 기본 제한은 다음과 같습니다.

- 리소스당 최대 태그 수 – 50개
- 각 리소스에 대해 각 태그 키는 고유하며 하나의 값만 가질 수 있습니다.
- 최대 키 길이 - UTF-8 형식의 유니코드 문자 128자
- 최대 값 길이 - UTF-8 형식의 유니코드 문자 256자
- 태그 지정 스키마를 여러 AWS 서비스와 리소스에서 사용하는 경우 다른 서비스에서 허용되는 문자에 제한이 있을 수 있음에 유의하세요. 일반적으로 허용되는 문자는 UTF-8로 표시할 수 있는 문자, 숫자 및 공백과 특수 문자 + - = . _ : / @입니다.
- 태그 키와 값은 대소문자를 구분합니다.
- AWS 용도로 예약된 키 또는 값에는 `aws:`, `AWS:` 또는 이러한 접두사의 대문자 또는 소문자 조합을 사용하지 않습니다. 이 접두사가 지정된 태그 키나 값은 편집하거나 삭제할 수 없습니다. 이 접두사가 포함된 태그는 리소스당 태그 수 제한에 포함되지 않습니다.

자습서: 콘솔을 사용하여 태그 관리

AWS Batch 콘솔을 사용하여 신규 또는 기존 컴퓨팅 환경, 작업, 작업 정의 및 작업 대기열과 연결된 태그를 관리할 수 있습니다.

생성 중 개별 리소스에서 태그 추가

AWS Batch 컴퓨팅 환경, 작업, 작업 정의, 작업 대기열 및 예약 정책을 생성할 때 태그를 추가할 수 있습니다.

개별 리소스의 태그 추가 및 삭제

AWS Batch 를 사용하면 리소스 페이지에서 클러스터와 연결된 태그를 직접 추가하거나 삭제할 수 있습니다.

개별 리소스에서 태그를 추가하거나 삭제하려면

1. https://console.aws.amazon.com/batch/ AWS Batch 콘솔을 엽니다.
2. 탐색 모음에서 사용할 리전을 선택합니다.
3. 탐색 창에서 리소스 유형을 선택합니다(예: 작업 대기열).
4. 특정 리소스를 선택한 다음, 태그 편집을 선택합니다.
5. 필요에 따라 태그를 추가하거나 삭제합니다.
   - 태그를 추가하려면 목록 끝에 있는 빈 텍스트 상자에 키와 값을 지정합니다.
   - 태그를 삭제하려면 태그 옆의

     Delete icon
     버튼을 선택합니다.
6. 추가하거나 삭제하려는 각 태그에 대해 이 프로세스를 반복한 다음 태그 편집을 선택하여 작업을 마칩니다.

CLI 또는 API를 사용하여 태그 관리

다음 AWS CLI 명령 또는 AWS Batch API 작업을 사용하여 리소스에 대한 태그를 추가, 업데이트, 나열 및 삭제합니다.

AWS Batch 리소스에 대한 태그 지원

Task	API 작업	AWS CLI	AWS Tools for Windows PowerShell
하나 이상의 태그를 추가하거나 덮어씁니다.	TagResource	tag-resource	Add-BATResourceTag
하나 이상의 태그를 삭제합니다.	UntagResource	untag-resource	Remove-BATResourceTag
리소스에 대한 태그 나열	ListTagsForResource	list-tags-for-resource	Get-BATResourceTag

다음 예제는 AWS CLI를 사용하여 리소스에 태그를 지정하거나 태그를 제거하는 방법을 보여줍니다.

예제 1: 기존 리소스에 태그 지정

다음 명령은 기존 리소스에 태그를 지정합니다.

```
aws batch tag-resource --resource-arn resource_ARN --tags team=devs
```

예제 2: 기존 리소스에서 태그 제거

다음 명령은 기존 리소스에서 태그를 삭제합니다.

```
aws batch untag-resource --resource-arn resource_ARN --tag-keys tag_key
```

예제 3: 리소스의 태그 나열

다음 명령은 기존 리소스와 연결된 태그를 나열합니다.

```
aws batch list-tags-for-resource --resource-arn resource_ARN
```

일부 리소스 생성 작업에서는 리소스를 생성할 때 태그를 지정할 수 있습니다. 다음 태스크는 생성 시 태그 지정을 지원합니다.

작업	API 작업	AWS CLI	AWS Tools for Windows PowerShell
컴퓨팅 환경 생성	CreateComputeEnvironment	create-compute-environment	New-BATComputeEnvironment
작업 대기열 생성	CreateJobQueue	create-job-queue	New-BATJobQueue
예약 정책 생성	CreateSchedulingPolicy	create-scheduling-policy	New-BATSchedulingPolicy
작업 정의 등록	RegisterJobDefinition	register-task-definition	Register-BATJobDefinition
작업 제출	SubmitJob	submit-job	Submit-BATJob
소모성 리소스 생성	CreateConsumableResource	create-consumable-resource	New-BATConsumableResource
서비스 환경 생성	CreateServiceEnvironment	create-service-environment	New-BATServiceEnvironment
할당량 공유 생성	CreateQuotaShare	create-quota-share	New-BATQuotaShare

에 대한 모범 사례 AWS Batch

복잡한 아키텍처 AWS Batch 를 관리하지 않고도를 사용하여 까다로운 다양한 컴퓨팅 워크로드를 대규모로 실행할 수 있습니다. AWS Batch 작업은 역학, 게임, 기계 학습과 같은 다양한 사용 사례에서 사용할 수 있습니다.

이 주제에서는 사용 시 고려해야 할 모범 사례 AWS Batch 와 사용 시 워크로드를 실행하고 최적화하는 방법에 대한 지침을 다룹니다 AWS Batch.

주제

- 사용 시기 AWS Batch
- 대규모 실행을 위한 체크리스트
- 컨테이너 및 AMI 최적화
- 적절한 컴퓨팅 환경 리소스를 선택하세요.
- Amazon EC2 온디맨드 또는 Amazon EC2 스팟
- 에 대한 Amazon EC2 스팟 모범 사례 사용 AWS Batch
- 오류 및 문제 해결

사용 시기 AWS Batch

AWS Batch 는 대규모 및 저비용으로 작업을 실행하고 대기열 서비스와 비용 최적화 규모 조정을 제공합니다. 그러나 모든 워크로드가를 사용하여 실행하기에 적합한 것은 아닙니다 AWS Batch.

- 짧은 작업 - 작업이 몇 초 동안만 실행되면 배치 작업을 예약하는 데 드는 오버헤드가 작업 자체의 런타임보다 오래 걸릴 수 있습니다. 해결 방법으로 binpack 작업을 제출하기 전에 작업을 함께 수행합니다 AWS Batch. 그런 다음 AWS Batch 작업을 반복하도록 작업을 구성합니다. 예를 들어 개별 태스크 인수를 Amazon DynamoDB 테이블 또는 Amazon S3 버킷의 파일로 스테이징합니다. 작업이 각각 3~5분씩 실행되도록 태스크를 그룹화하는 것을 고려해 봅니다. binpack 작업이 끝나면 AWS Batch 작업 내에서 작업 그룹을 반복하십시오.
- 즉시 실행해야 하는 작업 -는 작업을 빠르게 처리할 AWS Batch 수 있습니다. 그러나 AWS Batch 는 스케줄러이며 비용 성능, 작업 우선 순위 및 처리량을 최적화합니다. 요청을 처리하는 데 시간이 필요할 AWS Batch 수 있습니다. 몇 초 이내에 응답이 필요한 경우 Amazon ECS 또는 Amazon EKS를 사용하는 서비스 기반 접근 방식이 더 적합합니다.

대규모 실행을 위한 체크리스트

50,000개 이상의 vCPU에서 대규모 워크로드를 실행하기 전에 다음 체크리스트를 고려해 보세요.

> **Note**
>
> 백만 vCPUs에서 대규모 워크로드를 실행하려는 경우 또는 대규모로 실행되는 지침이 필요한 경우 AWS 팀에 문의하세요.

- Amazon EC2 할당량 확인하기 — AWS Management Console의 Service Quotas 패널에서 Amazon EC2 할당량(한도라고도 함)을 확인하세요. 필요한 경우 최대 Amazon EC2 인스턴스 수에 대한 할당량 증가를 요청하세요. Amazon EC2 스팟 인스턴스와 Amazon 온디맨드 인스턴스에는 별도의 할당량이 있다는 점을 기억하세요. 자세한 내용은 Service Quotas 시작하기를 참조하세요.
- 각 리전의 Amazon Elastic Block Store 할당량 확인하기 - 각 인스턴스는 운영 체제용 GP2 또는 GP3 볼륨을 사용합니다. 기본적으로 각 AWS 리전 할당량은 300TiB입니다. 하지만 각 인스턴스는 이 할당량의 일부로 건수를 사용합니다. 따라서 각 리전의 Amazon Elastic Block Store 할당량을 확인할 때 이 점을 고려해야 합니다. 할당량에 도달하면 인스턴스를 더 생성할 수 없습니다. 자세한 내용은 일반 참조에서 Amazon Elastic Block Store 엔드포인트 및 할당량을 참조하세요.
- 스토리지용 Amazon S3 사용하기 — Amazon S3는 높은 처리량을 제공하며 각 가용 영역의 작업 및 인스턴스 수를 기반으로 프로비저닝할 스토리지의 양을 추측할 필요가 없도록 지원합니다. 자세한 내용은 모범 사례 설계 패턴: Amazon S3 성능 최적화를 참조하세요.
- 점진적으로 확장하여 조기에 병목 현상 식별하기 - 백만 개 이상의 vCPU에서 실행되는 작업의 경우 병목 현상을 조기에 식별할 수 있도록 낮게 시작하여 점진적으로 늘립니다. 예를 들어 50,000개의 vCPU에서 실행하는 것으로 시작하세요. 그런 다음 개수를 20만 개의 vCPU로 늘리고 그 다음에는 50만 개의 vCPU 등으로 늘립니다. 즉, 원하는 vCPU 수에 도달할 때까지 vCPU 수를 계속 늘립니다.
- 모니터링하여 조기에 잠재적 문제를 식별하기 - 대규모 실행 시 잠재적인 중단과 문제를 방지하려면 애플리케이션과 아키텍처를 모두 모니터링해야 합니다. 1,000개에서 5,000개까지 vCPU를 확장할 때도 중단이 발생할 수 있습니다. Amazon CloudWatch Logs를 사용하여 로그 데이터를 검토하거나 클라이언트 라이브러리를 사용하여 CloudWatch Embedded Metrics를 사용할 수 있습니다. 자세한 내용은 CloudWatch Logs 에이전트 참조와 `aws-embedded-metrics` 섹션을 참조하세요.

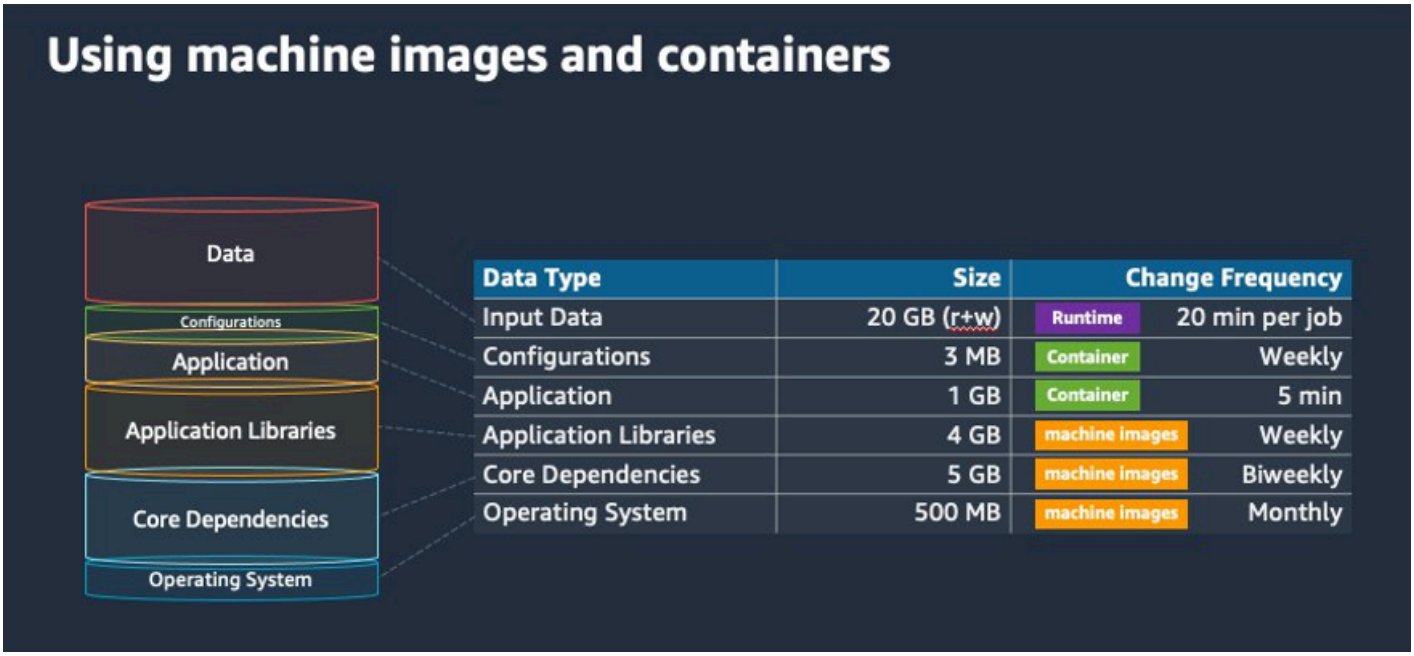
컨테이너 및 AMI 최적화

처음 실행하는 작업 세트에서 컨테이너 크기와 구조는 중요합니다. 컨테이너가 4GB보다 큰 경우 특히 그렇습니다. 컨테이너 이미지는 레이어로 빌드됩니다. 레이어는 Docker가 세 개의 동시 스레드를 사용

하여 병렬로 검색합니다. `max-concurrent-downloads` 파라미터를 사용하여 동시 스레드 수를 늘릴 수 있습니다. 자세한 내용은 Dockerd 설명서를 참조하세요.

더 큰 컨테이너를 사용할 수 있지만 시작 시간을 단축하려면 컨테이너 구조와 크기를 최적화하는 것이 좋습니다.

- 컨테이너가 작을수록 더 빨리 가져올 수 있습니다 - 컨테이너가 작을수록 애플리케이션 시작 시간이 빨라질 수 있습니다. 컨테이너 크기를 줄이려면 자주 업데이트되지 않는 라이브러리 또는 파일을 Amazon Machine Image(AMI)로 오프로드합니다. 바인드 탑재를 사용하여 컨테이너에 대한 액세스 권한을 부여할 수도 있습니다. 자세한 정보는 바인드 탑재 사용 섹션을 참조하세요.
- 크기가 균일한 레이어를 만들고 큰 레이어를 분할합니다 - 각 레이어는 스레드 하나로 검색됩니다. 따라서 레이어가 크면 작업 시작 시간에 큰 영향을 미칠 수 있습니다. 더 큰 컨테이너 크기와 더 빠른 시작 시간 사이의 균형을 맞추려면 최대 레이어 크기를 2GB로 설정하는 것이 좋습니다. `docker history your_image_id` 명령을 실행하여 컨테이너 이미지 구조와 레이어 크기를 확인할 수 있습니다. 자세한 내용은 Docker 설명서를 참조하세요.
- Amazon Elastic Container Registry를 컨테이너 리포지토리로 사용합니다 - 수천 개의 작업을 병렬로 실행하면 자체 관리형 리포지토리에 장애가 발생하거나 처리량이 제한될 수 있습니다. Amazon ECR은 대규모로 작동하며, 최대 백만 개 이상의 vCPU로 워크로드를 처리할 수 있습니다.



적절한 컴퓨팅 환경 리소스를 선택하세요.

AWS Fargate 는 Amazon EC2보다 초기 설정 및 구성이 덜 필요하며, 특히 처음 사용하는 경우 사용하기가 더 쉬울 수 있습니다. Fargate를 사용하면 서버를 관리하거나, 용량 계획을 처리하거나, 보안을 위해 컨테이너 워크로드를 격리할 필요가 없습니다.

다음과 같은 요구 사항이 있는 경우 Fargate 인스턴스를 사용하는 것이 좋습니다.

- 작업은 빠르게 시작해야 합니다.특히 30초 이내에 시작해야 합니다.
- 작업의 요구 사항은 vCPU 16개 이하, GPU 없음, 메모리 120GiB 이하입니다.

자세한 내용은 Fargate를 사용하는 경우 단원을 참조하십시오.

다음 요구 사항이 있는 경우 Amazon EC2 인스턴스를 사용하는 것이 좋습니다.

- 인스턴스 선택에 대한 제어력을 높이거나 특정 인스턴스 유형을 사용해야 합니다.
- 작업에는 GPUs, 더 많은 메모리, 사용자 지정 AMI 또는 Amazon Elastic Fabric Adapter와 같이 제공할 AWS Fargate 수 없는 리소스가 필요합니다.
- 높은 수준의 처리량 또는 동시성이 필요합니다.
- AMI, Amazon EC2 시작 템플릿을 사용자 지정하거나 특수 Linux 파라미터에 대한 액세스를 사용자 지정해야 합니다.

Amazon EC2를 사용하면 특정 요구 사항에 맞게 워크로드를 더 세밀하게 조정하고 필요한 경우 대규모로 실행할 수 있습니다.

Amazon EC2 온디맨드 또는 Amazon EC2 스팟

대부분의 AWS Batch 고객은 온디맨드 인스턴스에 대한 비용 절감으로 인해 Amazon EC2 스팟 인스턴스를 사용합니다. 하지만 워크로드가 여러 시간 동안 실행되고 중단할 수 없는 경우에는 온디맨드 인스턴스가 더 적합할 수 있습니다. 언제든지 스팟 인스턴스를 먼저 사용해 보고 필요한 경우 온디맨드로 전환할 수 있습니다.

다음과 같은 요구 사항 및 기대치가 있는 경우 Amazon EC2 온디맨드 인스턴스를 사용하세요.

- 작업 런타임이 1시간 이상이므로 워크로드가 중단되는 것을 용납할 수 없습니다.
- 전체 워크로드에 대한 엄격한 SLO(서비스 수준 목표)가 있으며 컴퓨팅 시간을 늘릴 수는 없습니다.

- 필요한 인스턴스에 중단이 발생할 가능성이 높습니다.

다음과 같은 요구 사항 및 기대치가 있는 경우 Amazon EC2 스팟 인스턴스를 사용하세요.

- 작업 런타임이 일반적으로 30분 이하입니다.
- 워크로드에서 잠재적인 중단과 작업 일정 재조정을 용납할 수 있습니다. 자세한 정보는 스팟 인스턴스 어드바이저를 참조하세요.
- 장기 실행 작업이 중단된 경우 체크포인트에서 다시 시작할 수 있습니다.

먼저 스팟 인스턴스에 제출한 다음 온디맨드 인스턴스를 대체 옵션으로 사용하여 두 구매 모델을 혼합하여 사용할 수 있습니다. 예를 들어 Amazon EC2 스팟 인스턴스에서 실행되는 컴퓨팅 환경에 연결된 대기열에 작업을 제출하세요. 작업이 중단되는 경우 Amazon EventBridge에서 이벤트를 포착하여 스팟 인스턴스 재확보와 연관시킵니다. 그런 다음 AWS Lambda 함수 또는를 사용하여 온디맨드 대기열에 작업을 다시 제출합니다 AWS Step Functions. 자세한 내용은 자습서: 실패한 작업 이벤트에 대한 Amazon Simple Notification Service 알림 보내기, Amazon EC2 스팟 인스턴스 중단 처리 모범 사례 및 Step Functions AWS Batch 로 관리를 참조하세요.

> **Important**
>
> 온디맨드 컴퓨팅 환경에 다양한 인스턴스 유형, 크기 및 가용 영역을 사용하여 Amazon EC2 스팟 인스턴스 풀의 가용성을 유지하고 중단률을 줄이세요.

에 대한 Amazon EC2 스팟 모범 사례 사용 AWS Batch

Amazon Elastic Compute Cloud(EC2) 스팟 인스턴스를 선택하면 워크플로를 최적화하여 비용을 절감할 수 있으며, 때로는 크게 절감할 수 있습니다. 자세한 내용은 Amazon EC2 Spot 보안 모범 사례를 참조하세요.

워크플로를 최적화하여 비용을 절감하려면 다음과 같은 AWS Batch에 대한 Amazon EC2 스팟 모범 사례를 참조하세요.

- **`SPOT_CAPACITY_OPTIMIZED`** 할당 전략을 선택합니다 - AWS Batch 는 가장 깊은 Amazon EC2 스팟 용량 풀에서 Amazon EC2 인스턴스를 선택합니다. 중단이 걱정된다면 이 방법을 선택하는 것이 좋습니다. 자세한 내용은 에 대한 인스턴스 유형 할당 전략 AWS Batch 단원을 참조하십시오.
- 인스턴스 유형 다각화 - 인스턴스 유형을 다각화하려면 호환되는 크기와 패밀리를 고려한 다음 가격 또는 가용성에 따라를 AWS Batch 선택합니다. 예를 들어, `c5.24xlarge`를 `c5.12xlarge` 또는

c5a, c5n, c5d, m5, m5d 패밀리의 대안으로 고려할 수 있습니다. 자세한 내용은 인스턴스 유형 및 가용 영역에 대한 유연성 유지를 참고하세요

- 작업 런타임 또는 체크포인트를 줄입니다 - Amazon EC2 스팟 인스턴스를 사용할 때 중단이 발생하지 않도록 1시간 이상 걸리는 작업은 실행하지 않는 것이 좋습니다. 작업을 30분 이하로 구성된 더 작은 부분으로 나누거나 체크포인트를 지정하면 중단 가능성을 크게 줄일 수 있습니다.
- 자동 재시도 사용 - AWS Batch 작업 중단을 방지하려면 작업에 대한 자동 재시도를 설정합니다. 0이 아닌 종료 코드가 반환되거나, 서비스 오류가 발생하거나, 인스턴스 재확보가 발생하는 등의 이유로 배치 작업이 중단될 수 있습니다. 자동 재시도는 최대 10회까지 설정할 수 있습니다. 시작하려면 최소 1~3회의 자동 재시도를 설정하는 것이 좋습니다. Amazon EC2 스팟 중단 추적에 대한 자세한 내용은 스팟 중단 대시보드를 참조하세요.

  의 AWS Batch경우 재시도 파라미터를 설정하면 작업이 작업 대기열 앞에 배치됩니다. 즉, 작업에 우선 순위가 부여됩니다. 작업 정의를 생성하거나에서 작업을 제출할 때 재시도 전략을 구성할 AWS CLI수 있습니다. 자세한 내용은 작업 이벤트를 참조하세요

```
$ aws batch submit-job --job-name MyJob \
    --job-queue MyJQ \
    --job-definition MyJD \
    --retry-strategy attempts=2
```

- 사용자 지정 재시도를 사용합니다 - 특정 애플리케이션 종료 코드 또는 인스턴스 재확보에 대한 작업 재시도 전략을 구성할 수 있습니다. 다음 예제에서는 호스트가 장애를 일으킨 경우 작업을 최대 5회까지 재시도할 수 있습니다. 하지만 다른 이유로 작업이 실패하면 작업이 종료되고 상태가 FAILED로 설정됩니다.

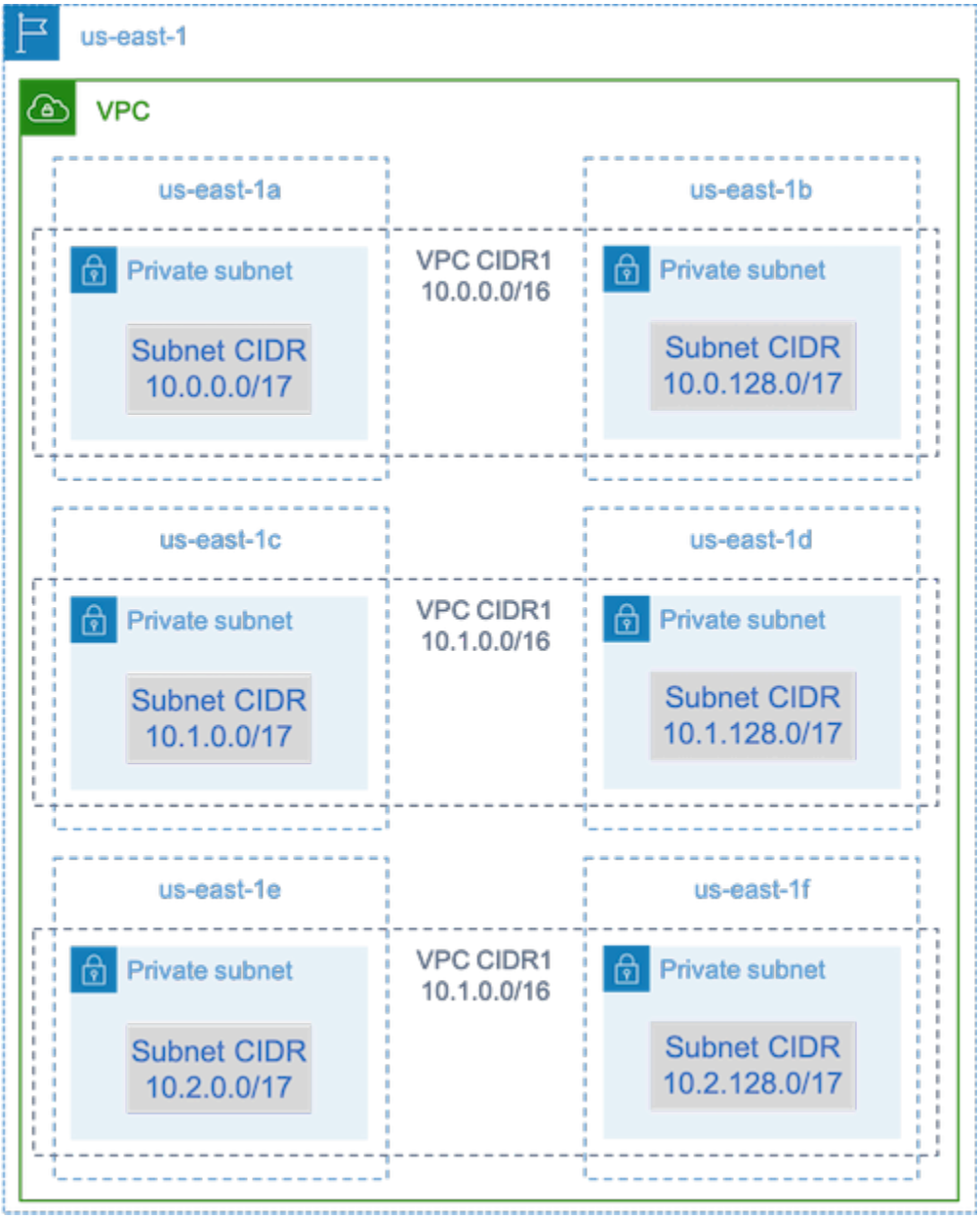
```
"retryStrategy": {
    "attempts": 5,
    "evaluateOnExit":
    [{
        "onStatusReason" :"Host EC2*",
        "action": "RETRY"
    },{
      "onReason" : "*",
        "action": "EXIT"
    }]
}
```

- 스팟 중단 대시보드를 사용합니다 - 스팟 중단 대시보드를 사용하여 스팟 중단을 추적할 수 있습니다. 애플리케이션은 재확보된 Amazon EC2 스팟 인스턴스와 해당 스팟 인스턴스가 속한 가용 영역에 대한 지표를 제공합니다. 자세한 정보는 스팟 인스턴스 중단을 참조하세요.

오류 및 문제 해결

의 오류는 AWS Batch 종종 애플리케이션 수준에서 발생하거나 특정 작업 요구 사항을 충족하지 않는 인스턴스 구성으로 인해 발생합니다. 다른 문제로는 작업이 RUNNABLE 상태에서 멈추거나 컴퓨팅 환경이 INVALID 상태에서 멈추는 경우가 있습니다. 작업이 RUNNABLE 상태에서 멈추는 문제 해결에 대한 자세한 내용은 RUNNABLE 상태에서 정체된 작업 섹션을 참조하세요. INVALID 상태의 컴퓨팅 환경 문제 해결에 대한 자세한 내용은 INVALID 컴퓨팅 환경 섹션을 참조하세요.

- Amazon EC2 스팟 vCPU 할당량을 확인합니다 - 현재 Service Quotas이 작업 요구 사항을 충족하는지 확인합니다. 예를 들어 현재 Service Quotas이 vCPU 256개이고 작업에 10,000개의 vCPU가 필요하다고 가정해 보겠습니다. 그러면 Service Quotas이 작업 요구 사항을 충족하지 못합니다. 자세한 내용 및 문제 해결 지침은 Amazon EC2 Service Quotas 및 Amazon EC2 리소스의 서비스 할당량을 늘리려면 어떻게 해야 하나요? 섹션을 참조하세요.
- 애플리케이션 실행 전 작업 실패 - DockerTimeoutError 오류나 CannotPullContainerError 오류로 인해 일부 작업이 실패할 수 있습니다. 문제 해결 정보는 의 "DockerTimeoutError" 오류를 해결하려면 어떻게 해야 합니까 AWS Batch?를 참조하세요.
- 불충분한 IP 주소 - VPC와 서브넷의 IP 주소 수에 따라 생성할 수 있는 인스턴스 수가 제한될 수 있습니다. Classless Inter-Domain Routing(CIDR)를 사용하여 워크로드 실행에 필요한 것보다 더 많은 IP 주소를 제공합니다. 필요한 경우 주소 공간이 큰 전용 VPC를 빌드할 수도 있습니다. 예를 들어, 10.x.0.0/16에 여러 개의 CIDR 있는 VPC를 만들고, 모든 가용 영역에 CIDR이 10.x.y.0/17인 서브넷을 만들 수 있습니다. 이 예제에서 x는 1~4 사이이고 y는 0 또는 128입니다. 이 구성은 모든 서브넷에서 36,000개의 IP 주소를 제공합니다.



- 인스턴스가 Amazon EC2에 등록되어 있는지 확인합니다 - Amazon EC2 콘솔에 인스턴스가 표시되지만 Amazon ECS 클러스터에 Amazon Elastic Container Service 컨테이너 인스턴스가 없는 경우 Amazon ECS 에이전트가 Amazon Machine Image(AMI)에 설치되지 않은 것일 수 있습니다 Amazon ECS 에이전트, AMI의 Amazon EC2 데이터 또는 시작 템플릿도 올바르게 구성되지 않을 수 있습니다. 근본 원인을 분리하려면 별도의 Amazon EC2 인스턴스를 만들거나 SSH를 사용하여 기존 인스턴스에 연결합니다. 자세한 내용은 Amazon ECS 컨테이너 에이전트 구성, Amazon ECS 로그 파일 위치, 컴퓨팅 리소스 AMI 섹션을 참조하세요.
- AWS 대시보드 검토 - AWS 대시보드를 검토하여 예상 작업 상태와 컴퓨팅 환경이 예상대로 확장되는지 확인합니다. CloudWatch에서 작업 로그를 검토할 수도 있습니다.

- 인스턴스 생성을 확인합니다 - 인스턴스가 생성되면 컴퓨팅 환경이 예상대로 확장되었음을 의미합니다. 인스턴스가 생성되지 않은 경우 컴퓨팅 환경에서 변경할 관련 서브넷을 찾아보세요. 자세한 내용을 알아보려면 Auto Scaling 그룹에 대한 크기 조정 활동 확인을 참조하세요.

  인스턴스가 관련 작업 요구 사항을 충족할 수 있는지도 확인하는 것이 좋습니다. 예를 들어 작업에 1TiB의 메모리가 필요할 수 있지만, 컴퓨팅 환경에서는 192GB로 제한된 C5 인스턴스 유형을 사용합니다.
- 에서 인스턴스를 요청하고 있는지 확인 AWS Batch - Auto Scaling 그룹 기록을 확인하여에서 인스턴스를 요청하고 있는지 확인합니다 AWS Batch. 이는 Amazon EC2가 인스턴스 획득을 시도하는 방식을 나타냅니다. Amazon EC2 스팟이 특정 가용 영역의 인스턴스를 확보할 수 없다는 오류 메시지가 표시되면, 가용 영역이 특정 인스턴스 패밀리를 제공하지 않기 때문일 수 있습니다.
- 인스턴스가 Amazon ECS에 등록되었는지 확인합니다 - Amazon EC2 콘솔에 인스턴스가 표시되지만 Amazon ECS 클러스터에 Amazon ECS 컨테이너 인스턴스가 없는 경우 Amazon ECS 에이전트가 Amazon Machine Image(AMI)에 설치되지 않은 것일 수 있습니다. 또한 Amazon ECS 에이전트, AMI의 Amazon EC2 데이터 또는 시작 템플릿이 올바르게 구성되지 않았을 수 있습니다. 근본 원인을 분리하려면 별도의 Amazon EC2 인스턴스를 만들거나 SSH를 사용하여 기존 인스턴스에 연결합니다. 자세한 내용은 CloudWatch 에이전트 구성 파일: 로그 섹션, Amazon ECS 로그 파일 위치 및 컴퓨팅 리소스 AMI(을)를 참조하세요.
- 지원 티켓을 엽니다 - 일부 문제 해결 후에도 여전히 문제가 발생하고 지원 계획이 있는 경우 지원 티켓을 엽니다. 지원 티켓에는 문제, 워크로드 세부 사항, 구성 및 테스트 결과에 대한 정보를 포함해야 합니다. 자세한 내용은 지원 플랜 비교를 참조하세요.
- AWS Batch 및 HPC 포럼 검토 - 자세한 내용은 AWS Batch 및 HPC 포럼을 참조하세요.
- AWS Batch 런타임 모니터링 대시보드 검토 -이 대시보드는 서버리스 아키텍처를 사용하여 Amazon ECS의 이벤트를 캡처 AWS Batch하고 Amazon EC2를 사용하여 작업 및 인스턴스에 대한 인사이트를 제공합니다. 자세한 내용은 AWS Batch Runtime Monitoring Dashboards Solution 섹션을 참조하세요.

문제 해결 AWS Batch

컴퓨팅 환경, 작업 대기열, 작업 정의, 작업 등과 관련한 문제를 해결해야 하는 경우도 있습니다. 이 장에서는 AWS Batch 환경에서 이러한 문제를 해결하고 해결하는 방법을 설명합니다.

AWS Batch 는 IAM 정책, 역할 및 권한을 사용하며 Amazon EC2, Amazon ECS AWS Fargate및 Amazon Elastic Kubernetes Service 인프라에서 실행됩니다. 이러한 서비스와 관련된 문제를 해결하려면 다음을 참조하세요.

- IAM 사용 설명서의 IAM 문제 해결
- Amazon Elastic Container Service 개발자 가이드의 Amazon ECS 문제 해결
- Amazon EKS 사용 설명서의 Amazon EKS 문제 해결
- Amazon EC2 사용 설명서의 EC2 인스턴스 문제 해결

목차

    - 지원되지 않는 Kubernetes 버전
    - 인스턴스 프로파일이 존재하지 않음
    - 유효하지 않은 Kubernetes 네임스페이스
    - 삭제된 컴퓨팅 환경
    - 노드가 Amazon EKS 클러스터에 조인하지 않음
- AWS Batch Amazon EKS 작업의 상태가 멈춤 RUNNABLE
- AWS Batch Amazon EKS 작업의 상태가 멈춤 STARTING
    - 시나리오: 영구 볼륨 클레임 연결 또는 탑재 실패
- aws-auth ConfigMap 필드가 제대로 구성되었는지 확인
- RBAC 권한 또는 바인딩이 제대로 구성되지 않음

AWS Batch

다음 주제를 검토하여 AWS Batch를 사용할 때 발생할 수 있는 일반적인 문제에 대한 검토 프로세스와 잠재적 솔루션을 찾습니다.

주제

- 자동 인스턴스 패밀리 업데이트를 수신하기 위한 최적의 인스턴스 유형 구성
- INVALID 컴퓨팅 환경
- RUNNABLE 상태에서 정체된 작업
- 생성 시 태그가 지정되지 않은 스팟 인스턴스
- 스팟 인스턴스가 스케일 다운되지 않음
- Secrets Manager 암호를 검색할 수 없음
- 작업 정의 리소스 요구 사항을 재정의할 수 없음
- desiredvCpus 설정을 업데이트할 때 나타나는 오류 메시지

자동 인스턴스 패밀리 업데이트를 수신하기 위한 최적의 인스턴스 유형 구성

AWS Batch 는 작업 대기열의 수요와 일치하도록 `optimal`에 대한 instanceTypes에서 단일 옵션을 지원했습니다. 이제 `default_x86_64` 및 `default_arm64`라는 두 가지 새로운 인스턴스 유형 옵션이 도입되었습니다. 인스턴스 유형을 선택하지 않는 경우 `default_x86_64`가 사용됩니다. 이러한 새 옵션은 작업 대기열 요구 사항을 기반으로 다양한 패밀리 및 세대에 걸쳐 비용 효율적인 인스턴스 유형을 자동 선택하여 워크로드를 빠르게 실행할 수 있게 해 줍니다.

이제 optimal 옵션은 리전 가용성에 따라 최신 m, c 및 r 인스턴스 패밀리에서 인스턴스 유형을 선택합니다. AWS Batch 는 이러한 패밀리 내의 최신 세대로 풀을 정기적으로 업데이트합니다. 를 사용하는 경우 별도의 조치가 필요하지 optimal않습니다.

그러나 ENABLED 및 VALID 컴퓨팅 환경(CEs)만 새 인스턴스 유형으로 업데이트됩니다. DISABLED 또는 INVALID CE가 있는 경우, 해당 CE가 다시 활성화되고 VALID 상태로 설정되면 업데이트를 받게 됩니다.

INVALID 컴퓨팅 환경

관리형 컴퓨팅 환경을 잘못 구성했을 수 있습니다. 잘못 구성한 경우 컴퓨팅 환경이 INVALID 상태가 되어 배치 작업을 수락할 수 없습니다. 다음 섹션에서는 발생 가능한 원인과 원인에 따른 문제 해결 방법을 설명합니다.

> **Important**
>
> AWS Batch 는 Amazon EC2 시작 템플릿, Amazon EC2 Auto Scaling 그룹, Amazon EC2 스팟 플릿, Amazon Amazon EC2 클러스터를 포함하여 사용자를 대신하여 계정 내에서 여러 AWS 리소스를 생성하고 관리합니다. 이러한 관리형 리소스는 최적의 AWS Batch 작동을 보장하도록 특별히 구성됩니다. AWS Batch 설명서에 명시되어 있지 않는 한 이러한 배치 관리형 리소스를 수동으로 수정하면 INVALID 컴퓨팅 환경에 예상치 못한 동작이 발생하여 최적화되지 않은 인스턴스 규모 조정 동작, 워크로드 프로세싱 지연 또는 예상치 못한 비용을 유발할 수 있습니다. 이러한 수동 수정은 AWS Batch 서비스에서 결정론적으로 지원할 수 없습니다. 항상 지원되는 배치 API 또는 배치 콘솔을 사용하여 컴퓨팅 환경을 관리하세요.

부정확한 역할 이름 또는 ARN

컴퓨팅 환경이 INVALID 상태로 전환되는 가장 일반적인 원인은 AWS Batch 서비스 역할 또는 Amazon EC2 스팟 플릿 역할에 잘못된 이름 또는 Amazon 리소스 이름(ARN)이 있기 때문입니다. 이는 AWS CLI 또는 AWS SDKs. 에서 컴퓨팅 환경을 생성할 때 올바른 서비스 또는 스팟 플릿 역할을 선택할 수 AWS Management Console AWS Batch 있습니다. 그러나 이름 또는 ARN을 수동으로 입력하지만 잘못 입력한다고 가정해 보겠습니다. 그러면 컴퓨팅 환경 결과도 마찬가지로 INVALID입니다.

그러나 AWS CLI 명령 또는 SDK 코드에 IAM 리소스의 이름 또는 ARN을 수동으로 입력한다고 가정해 보겠습니다. 이 경우는 문자열을 검증할 AWS Batch 수 없습니다. 대신 AWS Batch 는 잘못된 값을 수락하고 환경 생성을 시도해야 합니다. 가 환경을 생성 AWS Batch 하지 못하면 환경이 INVALID 상태로 전환되고 다음 오류가 표시됩니다.

유효하지 않은 서비스 역할인 경우:

```
CLIENT_ERROR - Not authorized to perform sts:AssumeRole (Service:
AWSSecurityTokenService; Status Code: 403; Error Code: AccessDenied;
Request ID: dc0e2d28-2e99-11e7-b372-7fcc6fb65fe7)
```

유효하지 않은 스팟 집합 역할인 경우:

```
CLIENT_ERROR - Parameter: SpotFleetRequestConfig.IamFleetRole
is invalid. (Service: AmazonEC2; Status Code: 400; Error Code:
InvalidSpotFleetRequestConfig; Request ID: 331205f0-5ae3-4cea-
bac4-897769639f8d) Parameter: SpotFleetRequestConfig.IamFleetRole is
invalid
```

이 문제가 발생하는 한 가지 일반적인 원인은 다음 시나리오입니다. AWS CLI 또는 AWS SDKs를 사용할 때 전체 Amazon 리소스 이름(ARN) 대신 IAM 역할의 이름만 지정합니다. 역할을 생성한 방식에 따라 ARN에 `aws-service-role` 경로 접두사가 포함될 수도 있습니다. 예를 들어 에 대한 서비스 연결 역할 사용 AWS Batch의 절차를 사용하여 AWS Batch 서비스 역할을 수동으로 생성하는 경우 서비스 역할 ARN은 다음과 같을 수 있습니다.

```
arn:aws:iam::123456789012:role/AWSBatchServiceRole
```

하지만 오늘 콘솔을 처음 실행할 때 서비스 역할을 만든 경우에는 서비스 역할 ARN은 다음과 같을 수 있습니다.

```
arn:aws:iam::123456789012:role/aws-service-role/AWSBatchServiceRole
```

이 문제는 AWS Batch 서비스 수준 정책(`AWSBatchServiceRole`)을 비서비스 역할에 연결하는 경우에도 발생할 수 있습니다. 예를 들어, 이 시나리오에서는 다음과 유사한 오류 메시지가 표시될 수 있습니다.

```
CLIENT_ERROR - User: arn:aws:sts::account_number:assumed-role/batch-replacement-role/
aws-batch is not
   authorized to perform: action on resource ...
```

이 문제를 해결하려면 다음 중 한 가지를 사용합니다.

- AWS Batch 컴퓨팅 환경을 생성할 때 서비스 역할에 빈 문자열을 사용합니다.
- `arn:aws:iam::account_number:role/aws-service-role/batch.amazonaws.com/AWSServiceRoleForBatch` 형식으로 서비스 역할을 지정합니다.

AWS CLI 또는 AWS SDKs는 ARN이 `aws-service-role` 경로 접두사를 사용하지 않는다고 AWS Batch 가정합니다. 이러한 이유로, 컴퓨팅 환경을 생성할 때는 IAM 역할의 전체 ARN을 지정하는 것이 좋습니다.

이와 같이 잘못 구성된 컴퓨팅 환경을 복원하려면 `INVALID` 컴퓨팅 환경 복원 섹션을 참조하세요.

`INVALID` 컴퓨팅 환경 복원

`INVALID` 상태의 컴퓨팅 환경이 있는 경우에는 업데이트를 통해 유효하지 않은 파라미터를 복원합니다. 부정확한 역할 이름 또는 ARN의 경우에는 올바른 서비스 역할을 사용하여 컴퓨팅 환경을 업데이트합니다.

잘못 구성된 컴퓨팅 환경을 복원하려면

1. https://console.aws.amazon.com/batch/ AWS Batch 콘솔을 엽니다.
2. 탐색 모음에서 사용할 AWS 리전 를 선택합니다.
3. 탐색 창에서 컴퓨팅 환경을 선택합니다.
4. 컴퓨팅 환경 페이지에서 편집할 컴퓨팅 환경 옆의 라디오 버튼을 선택한 다음 편집을 선택합니다 .
5. 컴퓨팅 환경 업데이트 페이지의 서비스 역할에서 컴퓨팅 환경에 사용할 IAM 역할을 선택합니다. AWS Batch 콘솔에는 컴퓨팅 환에 대해 올바른 신뢰 관계를 가지고 있는 역할만 표시됩니다.

   > **Tip**
   > 서비스 연결 역할을 생성하는 방법에 대한 지침은 에 역할 사용 AWS Batch 섹션을 참조하세요.

6. 저장을 선택하여 컴퓨팅 환경을 업데이트합니다.

`RUNNABLE` 상태에서 정체된 작업

컴퓨팅 환경에 컴퓨팅 리소스가 포함되어 있지만 작업이 `RUNNABLE` 상태 이상으로 진행되지 않는다고 가정해 보겠습니다. 이 경우 무언가가 작업을 컴퓨팅 리소스에 배치하지 못하게 하여 작업 대기열이 차단되고 있을 가능성이 높습니다. 다음은 작업이 차례를 기다리고 있는지 아니면 대기열을 차단하고 있는지 확인하는 방법입니다.

가 헤드에 `RUNNABLE` 작업이 있고 대기열을 차단했음을 AWS Batch 감지하면 Amazon CloudWatch Events에서 사유와 함께 작업 대기열 차단 이벤트 이벤트를 수신합니다. 동일한 이유가 `ListJobs` 및 `DescribeJobs` API 호출의 일부로 `statusReason` 필드에도 업데이트됩니다.

선택적으로 CreateJobQueue 및 UpdateJobQueue API 작업을 통해 jobStateTimeLimitActions 파라미터를 구성할 수 있습니다.

> **Note**
>
> 현재 Amazon ECS, Amazon EKS 또는 Fargate 컴퓨팅 환경에 연결된 작업 대기열의 경우와 함께 사용할 수 있는 유일한 작업은 작업을 취소하는 jobStateLimitActions.action 것입니다.

jobStateTimeLimitActions 파라미터는가 특정 상태의 작업에 대해 AWS Batch 수행하는 작업 세트를 지정하는 데 사용됩니다. maxTimeSeconds 필드를 통해 초 단위로 시간 임곗값을 설정할 수 있습니다.

작업이 정의된 RUNNABLE 상태인 경우는가 경과maxTimeSeconds한 후 지정된 작업을 statusReason AWS Batch 수행합니다.

예를 들어, 충분한 용량을 사용할 수 있을 때까지 대기 중인 RUNNABLE 상태의 모든 작업이 최대 4시간 동안 대기하도록 jobStateTimeLimitActions 파라미터를 설정할 수 있습니다. 작업을 취소CAPACITY:INSUFFICIENT_INSTANCE_CAPACITY하고 다음 작업이 작업 대기열의 헤드statusReason로 이동하도록 허용하기 전에 maxTimeSeconds를 로 설정하고를 14400으로 설정하여이 작업을 수행할 수 있습니다.

작업 대기열이 차단된 것을 감지하면가 AWS Batch 제공하는 이유는 다음과 같습니다. 이 목록은 ListJobs 및 DescribeJobs API 작업에서 반환되는 메시지를 제공합니다. 또한 동일한 값을 jobStateLimitActions.statusReason 파라미터에 대해 정의할 수 있습니다.

1. 이유: 연결된 모든 컴퓨팅 환경에 용량 부족 오류가 있습니다. 요청되면 용량 부족 오류가 발생하는 Amazon EC2 인스턴스를 AWS Batch 감지합니다. 작업을 수동으로 취소하면 후속 작업이 대기열의 헤드로 이동할 수 있습니다.
   - 작업이 멈춘 동안 **statusReason** 메시지: `CAPACITY:INSUFFICIENT_INSTANCE_CAPACITY - Service cannot fulfill the capacity requested for instance type [instanceTypeName]`
   - **jobStateTimeLimitActions**에 사용되는 **reason**: `CAPACITY:INSUFFICIENT_INSTANCE_CAPACITY`
   - 작업이 취소된 후 **statusReason** 메시지: `Canceled by JobStateTimeLimit action due to reason: CAPACITY:INSUFFICIENT_INSTANCE_CAPACITY`

참고:

a. 이 감지가 작동하려면 AWS Batch 서비스 역할에 `autoscaling:DescribeScalingActivities` 권한이 필요합니다. 에 대한 서비스 연결 역할 사용 AWS Batch 서비스 연결 역할(SLR) 또는 AWS 관리형 정책: AWSBatchServiceRole 정책 관리형 정책을 사용하는 경우 권한 정책이 업데이트되므로 다른 조치를 취할 필요가 없습니다.

b. SLR 또는 관리형 정책을 사용하는 경우 RUNNABLE 상태일 때 차단된 작업 대기열 이벤트와 업데이트된 작업 상태를 수신할 수 있도록 `autoscaling:DescribeScalingActivities` 및 `ec2:DescribeSpotFleetRequestHistory` 권한을 추가해야 합니다. 또한 AWS Batch 는 작업 대기열에 구성된 경우에도 `jobStateTimeLimitActions` 파라미터를 통해 `cancellation` 작업을 수행할 수 있도록 이러한 권한이 필요합니다.

c. 다중 노드 병렬(MNP) 작업의 경우 연결된 우선 순위가 높은 Amazon EC2 컴퓨팅 환경에 `insufficient capacity` 오류가 발생하면 우선 순위가 낮은 컴퓨팅 환경에 이 오류가 발생하더라도 대기열이 차단됩니다.

2. 이유: 모든 컴퓨팅 환경에는 작업 요구 사항보다 작은 `maxvCpus` 파라미터가 있습니다. 수동으로 또는 `statusReason`에 `jobStateTimeLimitActions` 파라미터를 설정하여 작업을 취소하면 후속 작업이 대기열의 상단으로 이동할 수 있습니다. 선택적으로, 차단된 작업의 요구 사항을 충족하도록 기본 컴퓨팅 환경의 `maxvCpus` 파라미터를 늘릴 수 있습니다.

   - 작업이 멈춘 동안 **`statusReason`** 메시지: `MISCONFIGURATION:COMPUTE_ENVIRONMENT_MAX_RESOURCE - CE(s) associated with the job queue cannot meet the CPU requirement of the job.`
   - **`jobStateTimeLimitActions`**에 사용되는 **`reason`**: `MISCONFIGURATION:COMPUTE_ENVIRONMENT_MAX_RESOURCE`
   - 작업이 취소된 후 **`statusReason`** 메시지: `Canceled by JobStateTimeLimit action due to reason: MISCONFIGURATION:COMPUTE_ENVIRONMENT_MAX_RESOURCE`

3. 이유: 작업 요구 사항을 충족하는 인스턴스가 있는 컴퓨팅 환경이 없습니다. 작업이 리소스를 요청하면는 연결된 컴퓨팅 환경이 수신 작업을 수용할 수 없음을 AWS Batch 감지합니다. 수동으로 또는 `statusReason`에 `jobStateTimeLimitActions` 파라미터를 설정하여 작업을 취소하면 후속 작업이 대기열의 상단으로 이동할 수 있습니다. 선택적으로, 컴퓨팅 환경의 허용되는 인스턴스 유형을 재정의하여 필요한 작업 리소스를 추가할 수 있습니다.

   - 작업이 멈춘 동안 **`statusReason`** 메시지: `MISCONFIGURATION:JOB_RESOURCE_REQUIREMENT - The job resource requirement (vCPU/memory/GPU) is higher than that can be met by the CE(s) attached to the job queue.`

- **jobStateTimeLimitActions**에 사용되는 **reason**: `MISCONFIGURATION:JOB_RESOURCE_REQUIREMENT`
- 작업이 취소된 후 **statusReason** 메시지: `Canceled by JobStateTimeLimit action due to reason: MISCONFIGURATION:JOB_RESOURCE_REQUIREMENT`

4. 이유: 모든 컴퓨팅 환경에 서비스 역할 문제가 있습니다. 이 문제를 해결하려면 서비스 역할 권한을 AWS 에 대한 관리형 정책 AWS Batch 섹션과 비교하고 모든 차이를 해결합니다. 참고: 이 오류를 해결하기 위해 `jobStateTimeLimitActions` 파라미터를 통해 프로그래밍 가능한 작업을 구성할 수 없습니다.

   유사한 오류를 방지하기 위해 에 대한 서비스 연결 역할 사용 AWS Batch 섹션을 사용하는 것이 좋습니다.

   수동으로 또는 statusReason에 `jobStateTimeLimitActions` 파라미터를 설정하여 작업을 취소하면 후속 작업이 대기열의 상단으로 이동할 수 있습니다. 서비스 역할 문제를 해결하지 않으면 다음 작업도 차단될 수 있습니다. 이 문제를 수동으로 조사하고 해결하는 것이 좋습니다.

   - 작업이 멈춘 동안 **statusReason** 메시지: `MISCONFIGURATION:SERVICE_ROLE_PERMISSIONS - Batch service role has a permission issue.`

5. 이유: 컴퓨팅 환경에 지원되지 않는 인스턴스 유형 구성이 있습니다. 이는 선택한 가용 영역에서 인스턴스 유형을 사용할 수 없거나 시작 템플릿 또는 시작 구성에 지정된 인스턴스 유형과 호환되지 않는 설정이 포함되어 있는 경우 발생할 수 있습니다. 이 문제를 해결하려면 지정된 AWS 리전 및 가용 영역에서 인스턴스 유형이 지원되는지 확인하고, 시작 템플릿 설정이 인스턴스 유형과 호환되는지 확인하고, 최신 세대 인스턴스 유형으로 업데이트를 고려하세요. 지원되는 인스턴스 유형을 찾는 방법에 대한 자세한 내용은 Amazon EC2 사용 설명서의 Amazon EC2 인스턴스 유형 찾기를 참조하세요.

   - 작업이 멈춘 동안 **statusReason** 메시지: `MISCONFIGURATION:EC2_INSTANCE_CONFIGURATION_UNSUPPORTED - Your compute environment associated with this job queue has an unsupported instance type configuration.`

6. 이유: 모든 컴퓨팅 환경이 유효하지 않습니다. 자세한 내용은 INVALID 컴퓨팅 환경 단원을 참조하십시오. 참고: 이 오류를 해결하기 위해 `jobStateTimeLimitActions` 파라미터를 통해 프로그래밍 가능한 작업을 구성할 수 없습니다.

   - 작업이 멈춘 동안 **statusReason** 메시지: `ACTION_REQUIRED - CE(s) associated with the job queue are invalid.`

7. Reason: AWS Batch 가 차단된 대기열을 감지했지만 이유를 확인할 수 없습니다. 참고: 이 오류를 해결하기 위해 `jobStateTimeLimitActions` 파라미터를 통해 프로그래밍 가능한 작업을 구성할 수 없습니다. 문제 해결에 대한 자세한 내용은 re:Post의 AWS에서 RUNNABLE 상태로 내 AWS Batch 작업이 중단된 이유를 참조하세요.
   - 작업이 멈춘 동안 **`statusReason`** 메시지: `UNDETERMINED - Batch job is blocked, root cause is undetermined.`

CloudWatch Events로부터 이벤트를 받지 못했거나 이유를 알 수 없는 이벤트를 받은 경우 이 문제의 몇 가지 일반적인 원인은 다음과 같습니다.

컴퓨팅 리소스에 **`awslogs`** 로그 드라이버가 구성되어 있지 않음

AWS Batch 작업은 로그 정보를 CloudWatch Logs로 전송합니다. 이를 위해서는 `awslogs` 로그 드라이버를 사용할 수 있도록 컴퓨팅 리소스를 구성해야 합니다. Amazon ECS에 최적화된 AMI(또는 Amazon Linux)를 기반으로 컴퓨팅 리소스 AMI를 구축한다고 가정해 보겠습니다. 그러면 이 드라이버가 `ecs-init` 패키지에 기본적으로 등록됩니다. 이제 다른 기본 AMI를 사용한다고 가정해 보겠습니다. 그러면 Amazon ECS 컨테이너 에이전트가 시작될 때 `ECS_AVAILABLE_LOGGING_DRIVERS` 환경 변수를 사용하여 `awslogs` 로그 드라이버를 사용 가능한 로그 드라이버로 지정해야 합니다. 자세한 내용은 컴퓨팅 리소스 AMI 사양 및 자습서: 컴퓨팅 리소스 AMI 생성 섹션을 참조하세요.

리소스 부족

작업 정의에서 CPU 또는 메모리 리소스가 컴퓨팅 리소스가 할당할 수 있는 크기를 초과하여 지정되어 있으면 작업이 배치되지 않습니다. 예를 들어 작업에 4GiB의 메모리가 지정되어 있고 컴퓨팅 리소스의 메모리가 사용 가능한 메모리보다 적다고 가정해 보겠습니다. 그러면 해당 컴퓨팅 리소스에 작업을 배치할 수 없는 경우가 발생합니다. 이러한 경우에는 작업 정의에서 지정한 메모리 크기를 줄이거나, 혹은 용량이 더욱 큰 컴퓨팅 리소스를 환경에 추가해야 합니다. 일부 메모리는 Amazon ECS 컨테이너 에이전트 및 기타 중요한 시스템 프로세스용으로 예약되어 있습니다. 자세한 내용은 컴퓨팅 리소스 메모리 관리 단원을 참조하십시오.

컴퓨팅 리소스에 대한 인터넷 액세스 없음

컴퓨팅 리소스는 Amazon ECS 서비스 엔드포인트와 통신하기 위한 액세스 권한이 필요합니다. 이는 인터페이스 VPC 엔드포인트를 통하거나 퍼블릭 IP 주소가 있는 컴퓨팅 리소스를 통해 이루어질 수 있습니다.

인터페이스 VPC 엔드포인트에 대한 자세한 정보는 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS 인터페이스 VPC 엔드포인트(AWS PrivateLink)를 참조하세요.

인터페이스 VPC 엔드포인트가 구성되어 있지 않고 컴퓨팅 리소스에 퍼블릭 IP 주소가 없는 경우 NAT(Network Address Translation)를 사용하여 이 액세스 권한을 제공해야 합니다. 자세한 내용은 Amazon VPC 사용 설명서의 NAT 게이트웨이 섹션을 참조하세요. 자세한 내용은 the section called "VPC 생성" 단원을 참조하십시오.

Amazon EC2 인스턴스 한도 도달

계정이에서 시작할 수 있는 Amazon EC2 인스턴스 수는 EC2 인스턴스 할당량에 따라 AWS 리전 결정됩니다. 일부 인스턴스 유형은 인스턴스 유형마다 할당량이 있습니다. 한도 증가를 요청하는 방법을 포함하여 계정의 Amazon EC2 인스턴스 할당량에 대한 자세한 내용은 Amazon EC2 사용 설명서의 Amazon EC2 서비스 한도를 참조하세요.

Amazon ECS 컨테이너 에이전트가 설치되지 않음

AWS Batch 가 작업을 실행하려면 Amazon Machine Image(AMI) 에 Amazon ECS 컨테이너 에이전트를 설치해야 합니다. Amazon ECS 컨테이너 에이전트는 Amazon ECS에 최적화된 AMI에 기본적으로 설치됩니다. Amazon ECS 컨테이너에 대한 자세한 내용은 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS 컨테이너 에이전트를 참조하세요.

시작 템플릿에서 장기 실행 사용자 데이터 스크립트

시작 템플릿에 완료하는 데 시간이 오래 걸리는 사용자 데이터 스크립트가 포함된 경우 인스턴스가 Amazon ECS에 등록되기 전에 시간이 초과될 수 있습니다. 이 경우 인스턴스는 작업을 픽업할 수 없게 되어 모든 작업이 `RUNNABLE` 상태로 유지됩니다. 인스턴스가 Amazon ECS에 등록하고 작업 실행을 시작하려면 모든 사용자 데이터 스크립트가 완료되어야 합니다.

이 문제를 해결하려면 시작 템플릿 사용자 데이터에서 장기 실행 또는 차단 작업을 검토하세요. 스크립트를 최적화하여 실행 시간을 줄이거나, 중요하지 않은 작업을 비동기적으로 실행하거나, 초기화 로직을 사용자 데이터 밖으로 완전히 이동하는 것이 좋습니다. 자세한 내용은 에서 Amazon EC2 시작 템플릿 사용 AWS Batch 단원을 참조하십시오.

자세한 내용은 내 AWS Batch 작업이 `RUNNABLE` 상태로 멈춰 있는 이유를 참조하세요. re:Post의 .

생성 시 태그가 지정되지 않은 스팟 인스턴스

AWS Batch 컴퓨팅 리소스에 대한 스팟 인스턴스 태그 지정이 2017년 10월 25일자로 지원됩니다. 이전에는 Amazon EC2 스팟 플릿 역할에 대한 권장된 IAM 관리형 정책(`AmazonEC2SpotFleetRole`)에 시작 시 스팟 인스턴스에 태그를 지정할 권한이 없었습니다. 새로운 권장 IAM 관리형 정책을 `AmazonEC2SpotFleetTaggingRole`이라고 합니다. 이 정책은 시작 시 스팟 인스턴스 태그 지정을 지원합니다.

스팟 인스턴스 생성 시 태그 지정을 수정하려면 다음 절차에 따라 현재 권장 IAM 관리형 정책을 Amazon EC2 스팟 플릿 역할에 적용합니다. 이렇게 하면 향후 해당 역할로 생성되는 모든 스팟 인스턴스는 생성 시 인스턴스 태그를 적용할 권한을 갖게 됩니다.

현재 IAM 관리형 정책을 Amazon EC2 스팟 플릿 역할에 적용하려면

1. https://console.aws.amazon.com/iam/에서 IAM 콘솔을 엽니다.
2. 역할을 선택한 후 Amazon EC2 스팟 플릿 역할을 선택합니다.
3. 정책 연결을 선택합니다.
4. AmazonEC2SpotFleetTaggingRole을 선택하고 정책 연결을 선택합니다.
5. Amazon EC2 스팟 플릿 역할을 다시 선택하여 이전 정책을 제거합니다.
6. Amazon EC2SpotFleetRole 정책 오른쪽에 있는 x를 선택하고 분리를 선택합니다.

스팟 인스턴스가 스케일 다운되지 않음

AWS Batch 는 2021년 3월 10일에 AWSServiceRoleForBatch 서비스 연결 역할을 도입했습니다. 컴퓨팅 환경의 `serviceRole` 파라미터에 역할이 지정되지 않은 경우 이 서비스 연결 역할이 서비스 역할로 사용됩니다. 하지만 서비스 연결 역할이 EC2 스팟 컴퓨팅 환경에서 사용되지만 사용되는 스팟 역할에는 AmazonEC2SpotFleetTaggingRole 관리형 정책이 포함되지 않는다고 가정해 보겠습니다. 그러면 스팟 인스턴스는 스케일 다운되지 않습니다. 그 결과 “이 작업을 수행할 권한이 없습니다.” 라는 오류 메시지가 표시됩니다. 다음 단계를 사용하여 `spotIamFleetRole` 파라미터에 사용하는 스팟 플릿 역할을 업데이트합니다. 자세한 내용은 IAM 사용 설명서의 서비스 연결 역할 사용 및 AWS 서비스에 권한을 위임하는 역할 생성을 참조하세요.

주제

- AmazonEC2SpotFleetTaggingRole 관리형 정책을의 스팟 플릿 역할에 연결 AWS Management Console
- 를 사용하여 AmazonEC2SpotFleetTaggingRole 관리형 정책을 스팟 플릿 역할에 연결 AWS CLI

AmazonEC2SpotFleetTaggingRole 관리형 정책을의 스팟 플릿 역할에 연결 AWS Management Console

현재 IAM 관리형 정책을 Amazon EC2 스팟 플릿 역할에 적용하려면

1. IAM 콘솔(https://console.aws.amazon.com/iam/)을 엽니다.
2. 역할을 선택한 후 Amazon EC2 스팟 플릿 역할을 선택합니다.

3. 정책 연결을 선택합니다.
4. AmazonEC2SpotFleetTaggingRole을 선택하고 정책 연결을 선택합니다.
5. Amazon EC2 스팟 플릿 역할을 다시 선택하여 이전 정책을 제거합니다.
6. Amazon EC2SpotFleetRole 정책 오른쪽에 있는 x를 선택하고 분리를 선택합니다.

를 사용하여 AmazonEC2SpotFleetTaggingRole 관리형 정책을 스팟 플릿 역할에 연결 AWS CLI

예제 명령은 Amazon EC2 스팟 플릿 역할의 이름이 *AmazonEC2SpotFleetRole*이라고 가정합니다. 역할이 다른 이름을 사용하는 경우 그에 맞게 명령을 조정합니다.

AmazonEC2SpotFleetTaggingRole 관리형 정책을 스팟 플릿 역할에 연결하려면

1. AmazonEC2SpotFleetTaggingRole 관리형 IAM 정책을 *AmazonEC2SpotFleetRole* 역할에 연결하려면 AWS CLI를 사용하여 다음 명령을 실행합니다.

```
$ aws iam attach-role-policy \
    --policy-arn arn:aws:iam::aws:policy/service-role/AmazonEC2SpotFleetTaggingRole
 \
    --role-name AmazonEC2SpotFleetRole
```

2. AmazonEC2SpotFleetRole 관리형 IAM 정책을 *AmazonEC2SpotFleetRole* 역할에서 분리하려면 AWS CLI를 사용하여 다음 명령을 실행합니다.

```
$ aws iam detach-role-policy \
    --policy-arn arn:aws:iam::aws:policy/service-role/AmazonEC2SpotFleetRole \
    --role-name AmazonEC2SpotFleetRole
```

Secrets Manager 암호를 검색할 수 없음

버전 1.16.0-1보다 낮은 Amazon ECS 에이전트와 함께 AMI를 사용하는 경우 이 기능을 사용하려면 Amazon ECS 에이전트 구성 변수 `ECS_ENABLE_AWSLOGS_EXECUTIONROLE_OVERRIDE=true`를 사용해야 합니다. 새 컨테이너 인스턴스를 생성할 때 새 컨테이너 인스턴스로 `./etc/ecs/ecs.config` 파일에 추가할 수 있습니다. 또는 기존 인스턴스에 추가할 수 있습니다. 기존 인스턴스에 추가하는 경우 추가 후 ECS 에이전트를 다시 시작해야 합니다. 자세한 내용을 알아보려면 Amazon Elastic Container Service 개발자 안내서의 Amazon ECS 컨테이너 에이전트 구성을 참조하세요.

작업 정의 리소스 요구 사항을 재정의할 수 없음

SubmitJob으로 전달된 containerOverrides 구조의 memory 및 vcpus 멤버에 지정된 메모리 및 vCPU 재정의는 작업 정의의 resourceRequirements 구조에 지정된 메모리 및 vCPU 요구 사항을 재정의할 수 없습니다.

이러한 리소스 요구 사항을 재정의하려고 하면 다음 오류 메시지가 표시될 수 있습니다.

"이 값은 더 이상 사용되지 않는 키로 제출되었으며 작업 정의의 리소스 요구 사항에서 제공하는 값과 충돌할 수 있습니다."

이 문제를 해결하려면 containerOverrides의 resourceRequirements 멤버에 메모리 및 vCPU 요구 사항을 지정합니다. 메모리 및 vCPU 재정의가 다음 줄에 지정된 경우를 예로 들어보겠습니다.

```
"containerOverrides": {
   "memory": 8192,
   "vcpus": 4
}
```

다음과 같이 변경합니다.

```
"containerOverrides": {
   "resourceRequirements": [
      {
         "type": "MEMORY",
         "value": "8192"
      },
      {
         "type": "VCPU",
         "value": "4"
      }
   ],
}
```

작업 정의의 containerProperties 객체에 지정된 메모리 및 vCPU 요구 사항도 동일하게 변경합니다. 메모리 및 vCPU 요구 사항이 다음 줄에 지정된 경우를 예로 들어보겠습니다.

```
{
   "containerProperties": {
      "memory": 4096,
```

```
      "vcpus": 2,
}
```

다음과 같이 변경합니다.

```
"containerProperties": {
   "resourceRequirements": [
      {
         "type": "MEMORY",
         "value": "4096"
      },
      {
         "type": "VCPU",
         "value": "2"
      }
   ],
}
```

`desiredvCpus` 설정을 업데이트할 때 나타나는 오류 메시지

AWS Batch API를 사용하여 원하는 vCPUs(`desiredvCpus`) 설정을 업데이트할 때 다음 오류 메시지가 표시됩니다.

`Manually scaling down compute environment is not supported. Disconnecting job queues from compute environment will cause it to scale-down to minvCpus.`

이 문제는 업데이트된 `desiredvCpus` 값이 현재 `desiredvCpus` 값보다 작을 경우 발생합니다. `desiredvCpus` 값을 업데이트할 때 다음 두 조건이 충족되어야 합니다.

- `desiredvCpus` 값은 `minvCpus` 및 `maxvCpus` 값 사이에 있어야 합니다.
- 업데이트된 `desiredvCpus` 값은 현재 `desiredvCpus` 값보다 크거나 같아야 합니다.

AWS Batch Amazon EKS의

주제

- INVALID 컴퓨팅 환경
- AWS Batch Amazon EKS 작업의 상태가 멈춤 RUNNABLE

- AWS Batch Amazon EKS 작업의 상태가 멈춤 STARTING
- aws-auth ConfigMap 필드가 제대로 구성되었는지 확인
- RBAC 권한 또는 바인딩이 제대로 구성되지 않음

다음 주제를 검토하여 Amazon Elastic Kubernetes Service AWS Batch 에서를 사용할 때 발생할 수 있는 일반적인 문제에 대한 검토 프로세스와 잠재적 솔루션을 찾습니다.

INVALID 컴퓨팅 환경

관리형 컴퓨팅 환경을 잘못 구성했을 수 있습니다. 잘못 구성한 경우 컴퓨팅 환경이 INVALID 상태가 되어 배치 작업을 수락할 수 없습니다. 다음 섹션에서는 발생 가능한 원인과 원인에 따른 문제 해결 방법을 설명합니다.

지원되지 않는 Kubernetes 버전

CreateComputeEnvironment API 작업 또는 UpdateComputeEnvironment API 작업을 사용하여 컴퓨팅 환경을 생성하거나 업데이트할 때 다음과 유사한 오류 메시지가 표시될 수 있습니다. EC2Configuration에서 지원되지 않는 Kubernetes 버전을 지정하는 경우 이 문제가 발생합니다.

```
At least one imageKubernetesVersion in EC2Configuration is not supported.
```

이 문제를 해결하려면 컴퓨팅 환경을 삭제하고 지원되는 Kubernetes 버전으로 다시 생성하세요.

Amazon EKS 클러스터에서 마이너 버전 업그레이드를 수행할 수 있습니다. 예를 들어 마이너 버전이 지원되지 않는 경우에도 클러스터를 1.xx에서 1.yy로 업그레이드할 수 있습니다.

하지만 메이저 버전 업데이트 후에는 컴퓨팅 환경 상태가 INVALID로 변경될 수 있습니다. 메이저 버전을 1.xx에서 2.yy로 업그레이드하는 경우를 예로 들 수 있습니다. 메이저 버전이에서 지원되지 않는 경우 다음과 유사한 오류 메시지가 AWS Batch표시됩니다.

```
reason=CLIENT_ERROR - ... EKS Cluster version [2.yy] is unsupported
```

이 문제를 해결하려면 API 작업을 사용하여 컴퓨팅 환경을 생성하거나 업데이트할 때, 지원되는 Kubernetes 버전을 지정합니다.

AWS Batch Amazon EKS의는 현재 다음 Kubernetes 버전을 지원합니다.

- 1.34

- 1.33
- 1.32
- 1.31
- 1.30
- 1.29

인스턴스 프로파일이 존재하지 않음

지정된 인스턴스 프로파일이 없는 경우 Amazon EKS 컴퓨팅 환경 AWS Batch 의 상태가 로 변경됩니다INVALID. statusReason 파라미터에 다음과 유사한 오류 세트가 표시됩니다.

```
CLIENT_ERROR - Instance profile arn:aws:iam::...:instance-profile/<name> does not exist
```

이 문제를 해결하려면 작업 인스턴스 프로파일을 지정하거나 생성합니다. 자세한 내용을 알아보려면 Amazon EKS 사용 설명서의 Amazon EKS 노드 IAM 역할을 참조하세요.

유효하지 않은 Kubernetes 네임스페이스

Amazon EKS AWS Batch 에서 컴퓨팅 환경의 네임스페이스를 검증할 수 없는 경우 컴퓨팅 환경 상태가 로 변경됩니다INVALID. 예를 들어 네임스페이스가 존재하지 않는 경우 이 문제가 발생할 수 있습니다.

statusReason 파라미터에 다음과 유사한 오류 메시지 세트가 표시됩니다.

```
CLIENT_ERROR - Unable to validate Kubernetes Namespace
```

다음 중 하나에 해당하면 이 문제가 발생할 수 있습니다.

- CreateComputeEnvironment 호출의 Kubernetes 네임스페이스 문자열이 존재하지 않습니다. 자세한 내용은 CreateComputeEnvironment를 참조하세요.
- 네임스페이스를 관리하는 데 필요한 역할 기반 액세스 제어(RBAC) 권한이 제대로 구성되지 않습니다.
- AWS Batch 는 Amazon EKS Kubernetes API 서버 엔드포인트에 액세스할 수 없습니다.

이 문제를 해결하려면 aws-auth ConfigMap 필드가 제대로 구성되었는지 확인 섹션을 참조하세요. 자세한 내용은 Amazon EKS AWS Batch 에서 시작하기 단원을 참조하십시오.

삭제된 컴퓨팅 환경

Amazon EKS 컴퓨팅 환경에서 연결된를 삭제하기 전에 AWS Batch Amazon EKS 클러스터를 삭제한다고 가정해 보겠습니다. 그러면 컴퓨팅 환경 상태가 `INVALID`로 변경됩니다. 이 시나리오에서 동일한 이름으로 Amazon EKS 클러스터를 다시 생성하면 컴퓨팅 환경이 제대로 작동하지 않습니다.

이 문제를 해결하려면 Amazon EKS 컴퓨팅 환경에서 AWS Batch 를 삭제한 다음 다시 생성합니다.

노드가 Amazon EKS 클러스터에 조인하지 않음

AWS Batch Amazon EKS의는 모든 노드가 Amazon EKS 클러스터에 조인되지 않은 것으로 확인되면 컴퓨팅 환경을 축소합니다. Amazon EKS AWS Batch 에서 컴퓨팅 환경을 스케일 다운하면 컴퓨팅 환경 상태가 로 변경됩니다`INVALID`.

> **Note**
>
> AWS Batch 는 문제를 디버깅할 수 있도록 컴퓨팅 환경 상태를 즉시 변경하지 않습니다.

`statusReason` 파라미터에 설정된 오류 메시지는 다음 중 하나와 유사합니다.

```
Your compute environment has been INVALIDATED and scaled down because
none of the instances joined the underlying ECS Cluster. Common issues
preventing instances joining are the following: VPC/Subnet configuration
preventing communication to ECS, incorrect Instance Profile policy
preventing authorization to ECS, or customized AMI or LaunchTemplate
configurations affecting ECS agent.
```

```
Your compute environment has been INVALIDATED and scaled down because
none of the nodes joined the underlying Amazon EKS Cluster. Common issues
preventing nodes joining are the following: networking configuration
preventing communication to Amazon EKS Cluster, incorrect Amazon EKS
Instance Profile or Kubernetes RBAC policy preventing authorization
to Amazon EKS Cluster, customized AMI or LaunchTemplate configurations
affecting Amazon EKS/Kubernetes node bootstrap.
```

기본 Amazon EKS AMI를 사용하는 경우 이 문제의 가장 일반적인 원인은 다음과 같습니다.

- 인스턴스 역할이 올바르게 구성되지 않았습니다. 자세한 내용을 알아보려면 Amazon EKS 사용 설명서의 Amazon EKS 노드 IAM 역할을 참조하세요.

- 서브넷이 제대로 구성되지 않았습니다. 자세한 내용은 Amazon EKS 사용 설명서의 Amazon EKS VPC 및 서브넷 요구 사항 및 고려 사항을 참조하세요.
- 보안 그룹이 올바르게 구성되지 않았습니다. 자세한 내용은 Amazon EKS 사용 설명서의 Amazon EKS 보안 그룹 요구 사항 및 고려 사항을 참조하세요.

> **Note**
>
> PHD(Personal Health Dashboard)에 오류 알림이 표시될 수도 있습니다.

AWS Batch Amazon EKS 작업의 상태가 멈춤 **RUNNABLE**

`aws-auth ConfigMap`은 관리형 노드 그룹을 생성하거나 `eksctl`을 사용하여 노드 그룹을 생성할 때 자동으로 생성되어 클러스터에 적용됩니다. `aws-auth ConfigMap`은 처음에 노드가 클러스터에 조인할 수 있도록 하기 위해 생성됩니다. 하지만 `aws-authConfigMap`을 사용하여 사용자 및 역할에 역할 기반 액세스 제어(RBAC) 액세스를 추가할 수도 있습니다.

`aws-auth ConfigMap`이 제대로 구성되었는지 확인하려면

1. `aws-auth ConfigMap`에서 매핑된 역할을 검색합니다.

   ```
   $ kubectl get configmap -n kube-system aws-auth -o yaml
   ```

2. `roleARN`이 다음과 같이 구성되어 있는지 확인합니다.

   ```
   rolearn: arn:aws:iam::aws_account_number:role/AWSServiceRoleForBatch
   ```

   > **Note**
   >
   > Amazon EKS 컨트롤 플레인 로그를 검토할 수도 있습니다. 자세한 내용을 알아보려면 Amazon EKS 사용 설명서의 Amazon EKS 클러스터 컨트롤 플레인 로깅을 참조하세요.

작업이 `RUNNABLE` 상태에서 멈추는 문제를 해결하려면 `kubectl`을 사용하여 매니페스트를 다시 적용하는 것이 좋습니다. 자세한 내용은 2단계: Amazon EKS 클러스터 준비 AWS Batch 단원을 참조하십시오. 또는 `kubectl`을 사용하여 `aws-auth ConfigMap`을 수동으로 편집할 수 있습니다. 자세한 내용은 Amazon EKS 사용 설명서의 클러스터에 대한 IAM 사용자 및 역할 액세스 활성화를 참조하세요.

AWS Batch Amazon EKS 작업의 상태가 멈춤 **STARTING**

kubelet의 장기 실행 요청(pull, log, exec 및 attach)에 대해 포드가 ContainerCreating에서 PENDING에 멈춰 있으면, 포드 시작 문제가 해결되거나 작업이 종료될 때까지 작업이 STARTING 상태로 유지될 수 있습니다. 아래 적격 시나리오에서 AWS Batch 는 사용자를 대신하여 작업을 종료합니다. 그렇지 않으면 TerminateJob API를 사용하여 작업을 수동으로 종료해야 합니다.

STARTING에서 작업이 멈출 수 있는 사유를 확인하려면 자습서: 실행 중인 작업을 포드 및 노드에 매핑하기을 사용하여 podName을 찾고 포드를 설명합니다.

```
% kubectl describe pod aws-batch.000c8190-87df-31e7-8819-176fe017a24a -n my-aws-batch-
namespace
Name:             aws-batch.000c8190-87df-31e7-8819-176fe017a24a
Namespace:        my-aws-batch-namespace
...
Containers:
  default:
...
    State:          Waiting
      Reason:       ContainerCreating
    Ready:          False
...
Conditions:
  Type                        Status
  PodReadyToStartContainers   False
  Initialized                 True
  Ready                       False
  ContainersReady             False
  PodScheduled                True
...
Events:
  Type     Reason       Age    From     Message
  ----     ------       ----   ----     -------
  Warning  FailedMount  2m32s  kubelet  Unable to attach or mount volumes: ...
```

완전한 가시성을 위해 컨트롤 플레인 로그를 CloudWatch Logs로 전송하도록 EKS 클러스터를 구성하는 것을 고려하세요.

시나리오: 영구 볼륨 클레임 연결 또는 탑재 실패

볼륨이 연결되지 않거나 마운트되지 않는 영구 볼륨 클레임을 사용하는 작업은 종료 대상입니다. 이는 잘못 구성된 작업 정의의 결과일 수 있습니다. 자세한 내용은 Amazon EKS 리소스에 단일 노드 작업 정의 생성 섹션을 참조하세요.

aws-auth ConfigMap 필드가 제대로 구성되었는지 확인

aws-auth ConfigMap이 제대로 구성되었는지 확인하려면

1. aws-auth ConfigMap에서 매핑된 역할을 검색합니다.

```
$ kubectl get configmap -n kube-system aws-auth -o yaml
```

2. roleARN이 다음과 같이 구성되어 있는지 확인합니다.

```
rolearn: arn:aws:iam::aws_account_number:role/AWSServiceRoleForBatch
```

> **Note**
>
> 서비스 연결 역할의 ARN에서 aws-service-role/batch.amazonaws.com/ 경로가 제거되었습니다. 이는 구성 맵에 aws-auth 문제가 있기 때문입니다. 자세한 내용은 경로가 aws-authconfigmap의 ARN에 포함될 때 경로가 있는 역할이 작동하지 않는 경우를 참조하세요.

> **Note**
>
> Amazon EKS 컨트롤 플레인 로그를 검토할 수도 있습니다. 자세한 내용을 알아보려면 Amazon EKS 사용 설명서의 Amazon EKS 클러스터 컨트롤 플레인 로깅을 참조하세요.

작업이 RUNNABLE 상태에서 멈추는 문제를 해결하려면 kubectl을 사용하여 매니페스트를 다시 적용하는 것이 좋습니다. 자세한 내용은 2단계: Amazon EKS 클러스터 준비 AWS Batch 단원을 참조하십시오. 또는 kubectl을 사용하여 aws-auth ConfigMap을 수동으로 편집할 수 있습니다. 자세한 내용은 Amazon EKS 사용 설명서의 클러스터에 대한 IAM 사용자 및 역할 액세스 활성화를 참조하세요.

RBAC 권한 또는 바인딩이 제대로 구성되지 않음

RBAC 권한 또는 바인딩 문제가 발생하는 경우 `aws-batch` Kubernetes 역할이 Kubernetes 네임스페이스에 액세스할 수 있는지 확인합니다.

```
$ kubectl get namespace namespace --as=aws-batch
```

```
$ kubectl auth can-i get ns --as=aws-batch
```

`kubectl describe` 명령을 사용하여 클러스터 역할 또는 Kubernetes 네임스페이스에 대한 권한을 볼 수도 있습니다.

```
$ kubectl describe clusterrole aws-batch-cluster-role
```

출력의 예시는 다음과 같습니다.

```
Name:         aws-batch-cluster-role
Labels:       <none>
Annotations:  <none>
PolicyRule:
  Resources                                       Non-Resource URLs  Resource Names
 Verbs
  ---------                                       -----------------  --------------
 -----
  configmaps                                      []                 []
 [get list watch]
  nodes                                           []                 []
 [get list watch]
  pods                                            []                 []
 [get list watch]
  daemonsets.apps                                 []                 []
 [get list watch]
  deployments.apps                                []                 []
 [get list watch]
  replicasets.apps                                []                 []
 [get list watch]
  statefulsets.apps                               []                 []
 [get list watch]
  clusterrolebindings.rbac.authorization.k8s.io   []                 []
 [get list]
```

```
  clusterroles.rbac.authorization.k8s.io         []                 []
 [get list]
  namespaces                                     []                 []
 [get]
  events                                         []                 []
 [list]
```

```
$ kubectl describe role aws-batch-compute-environment-role -n my-aws-batch-namespace
```

출력의 예시는 다음과 같습니다.

```
Name:         aws-batch-compute-environment-role
Labels:       <none>
Annotations:  <none>
PolicyRule:
  Resources                                  Non-Resource URLs  Resource Names  Verbs
  ---------                                  -----------------  --------------  -----
  pods                                       []                 []              [create
 get list watch delete patch]
  serviceaccounts                            []                 []              [get list]
  rolebindings.rbac.authorization.k8s.io     []                 []              [get list]
  roles.rbac.authorization.k8s.io            []                 []              [get list]
```

이 문제를 해결하려면 RBAC 권한 및 `rolebinding` 명령을 다시 적용합니다. 자세한 내용은 2단계: Amazon EKS 클러스터 준비 AWS Batch 섹션을 참조하세요.

Resource: AWS Batch service 할당량

다음 표에는 변경할 AWS Batch 수 없는에 대한 서비스 할당량이 나와 있습니다. 각 할당량은 리전별로 적용됩니다.

Resource	할당량
최대 작업 대기열 수 자세한 내용은 작업 대기열 단원을 참조하십시오.	50
Amazon ECS 및 Amazon EKS 전반의 최대 컴퓨팅 환경 수. 자세한 내용은 컴퓨팅 환경 AWS Batch 단원을 참조하십시오.	50
Amazon EKS 클러스터당 최대 컴퓨팅 환경 수.	5
작업 대기열당 최대 컴퓨팅 환경 수	3
작업당 최대 작업 종속성 수	20
최대 작업 정의 크기(`RegisterJobDefinition` API 작업)	24KiB
최대 작업 페이로드 크기(`SubmitJob` API 작업의 경우)	30KiB
어레이 작업의 어레이 최대 크기	10000
`SUBMITTED` 상태 작업의 최대수	1000000
`SubmitJob` 작업에 대한 각 계정의 최대 초당 트랜잭션 수(TPS)	50
소모성 리소스의 최대 수	50k
서비스 환경의 최대 수. 자세한 내용은 서비스 환경 단원을 참조하십시오.	50
각 작업 대기열의 최대 서비스 환경 수	1
할당량 관리가 활성화된 각 서비스 환경에 연결된 최대 작업 대기열 수입니다. 자세한 내용은 the section called “할당량 관리”을 참조하세요.	1
각 할당량 관리 작업 대기열의 최대 할당량 공유 수입니다. 자세한 내용은 the section called “할당량 관리”을 참조하세요.	20

Resource	할당량
최대 SubmitServiceJob 요청 크기	30KiB
최대 작업 서비스 요청 페이로드 크기(SubmitServiceJob API 작업의 경우)	10KiB
SubmitServiceJob 작업에 대한 각 계정의 최대 초당 트랜잭션 수(TPS)	5
서비스 작업에 대한 재시도 전략의 최대 시도 횟수	10

사용 방법에 따라 AWS Batch추가 할당량이 적용될 수 있습니다. Amazon EC2 할당량에 대해 자세히 알아보려면 AWS 일반 참조의 Amazon EC2 Service Quotas 섹션을 참조하세요. Amazon ECS 할당량에 대한 자세한 내용은 AWS 일반 참조의 Amazon ECS Service Quotas 섹션을 참조하세요. Amazon EKS 할당량에 대한 자세한 내용은 AWS 일반 참조의 Amazon EKS Service Quotas 섹션을 참조하세요.

문서 기록

다음 표에서는의 최초 릴리스 이후 설명서에 대한 중요한 변경 사항을 설명합니다 AWS Batch. 사용자로부터 받은 의견을 수렴하기 위해 설명서가 자주 업데이트됩니다.

변경 사항	설명	날짜
업데이트된 AWSBatchServiceRolePolicyForSageMaker	`sagemaker:DeleteTrainingJob` 권한을 추가하도록 `AWSBatchServiceRolePolicyForSageMaker` 관리형 정책을 업데이트했습니다.	2026년 4월 16일
default_x86_64 및 default_arm64 추가	허용된 인스턴스 유형에 대해 새로운 `default_x86_64` 및 `default_arm64` 가 추가되었습니다.	2025년 8월 15일
서비스 환경 및 서비스 작업 추가	SageMaker AI와 AWS Batch 함께 사용하기 위한 서비스 환경 및 서비스 작업이 추가되었습니다.	2025년 7월 30일
AWSServiceRoleForAWSBatchWithSagemaker 및 AWSBatchServiceRolePolicyForSageMaker 추가	가 사용자를 대신하여 SageMaker AI AWS Batch 를 관리할 수 `AWSBatchServiceRolePolicyForSageMaker` 있도록 허용하는 새로운 AWS 서비스 연결 역할 `AWSServiceRoleForAWSBatchWithSagemaker` 및 관리형 정책이 추가되었습니다.	2025년 7월 30일
EKS AL 2023 AMI에 대한 지원 추가	EKS AL2에서 EKS AL2023으로 업그레이드하는 방법	2025년 6월 24일

FireLens 및 ECS Exec 명령에 대한 지원 추가	FireLens 및 ECS Exec 명령에 대한 지원이 추가되었습니다.	2025년 4월 15일
에 대한 리소스 인식 예약에 대한 지원을 추가합니다. AWS Batch	Amazon Elastic Container Service, Amazon Elastic Kubernetes Service 및에 대한 리소스 인식 예약에 AWS Batch 대한 지원을 추가합니다 AWS Fargate.	2025년 2월 27일
AWS Batch 지원되는 Amazon EKS 버전 업데이트	에서 버전 1.22를 제거하도록 AWS Batch 지원하는 Amazon EKS 버전을 업데이트했습니다.	2024년 3월 11일
AWS Batch 지원되는 Amazon EKS 버전 업데이트	버전 1.29를 포함하도록가 AWS Batch 지원하는 Amazon EKS 버전을 업데이트했습니다.	2024년 2월 29일
작업 자동 재시도	코드 샘플을 수정했습니다.	2024년 2월 29일
에 대한 멀티컨테이너 작업에 대한 지원을 추가합니다. AWS Batch	Amazon Elastic Container Service, Amazon Elastic Kubernetes Service 및에 대한 다중 컨테이너 작업에 AWS Batch 대한 지원을 추가합니다 AWS Fargate.	2024년 2월 28일
AWS Batch 지원되는 Amazon EKS 버전 업데이트	버전 1.28을 포함하도록가 AWS Batch 지원하는 Amazon EKS 버전을 업데이트했습니다.	2024년 1월 27일

BatchServiceRolePolicy 및 AWSBatchServiceRole 업데이트	BatchServiceRolePolicy 스팟 플릿 요청 기록 및 Amazon EC2 Auto Scaling 활동 설명을 위한 지원을 추가하도록 업데이트되었습니다. AWSBatchServiceRole 문 IDs 추가`ec2:DescribeSpotFleetRequestHistory` 하고 및에 AWS Batch 권한을 부여하도록 업데이트되었습니다 `autoscaling:DescribeScalingActivities` .	2023년 12월 5일
AWS Batch Amazon EKS의	AWS Batch 는 Amazon EKS 클러스터에서 작업 실행에 대한 지원을 추가합니다.	2022년 10월 25일
에 대한 교차 서비스 혼동된 대리자 방지 AWS Batch	AWS Batch 는 이제 엔터티(서비스 또는 계정)가 다른 엔터티에 의해 작업을 수행하도록 강요될 때 발생하는 혼동된 대리자 보안 문제에 대한 해결 방법을 제공합니다.	2022년 6월 6일
인터페이스 VPC 엔드포인트 (AWS PrivateLink)	기반 인터페이스 VPC 엔드포인트 구성에 대한 지원이 추가되었습니다 AWS PrivateLink. 즉, NAT 인스턴스, VPN 연결 또는를 통해 액세스할 필요 AWS Batch 없이 VPC와 간에 프라이빗 연결을 생성할 수 있습니다 Direct Connect.	2022년 4월 15일

향상된 컴퓨팅 환경 업데이트	AWS Batch 컴퓨팅 환경에 대한 지원 업데이트가 개선되었습니다.	2022년 4월 14일
AWS 관리형 정책 업데이트 - 기존 정책 업데이트	AWS Batch 는 기존 관리형 정책을 업데이트했습니다.	2021년 12월 6일
공정 공유 예약	AWS Batch 는 작업 대기열에 예약 정책을 추가하기 위한 지원을 추가합니다.	2021년 11월 9일
Amazon EFS	AWS Batch 는 Amazon EFS 파일 시스템을 작업 정의에 추가하는 지원을 추가합니다.	2021년 4월 1일
서비스 연결 역할 추가	AWS Batch 는 AWSServiceRoleForBatch 서비스 연결 역할을 추가합니다.	2021년 3월 10일
AWS Fargate 지원	AWS Batch 는 Fargate 리소스에서 작업 실행에 대한 지원을 추가합니다.	2020년 12월 3일
리소스에 태깅	AWS Batch 는 컴퓨팅 환경, 작업 정의, 작업 대기열 및 작업에 메타데이터 태그를 추가할 수 있는 지원을 추가합니다.	2020년 10월 7일
암호	AWS Batch 는 작업에 보안 암호를 전달하기 위한 지원을 추가합니다.	2020년 10월 1일
로깅	AWS Batch 는 작업에 대한 추가 로그 드라이버 지정에 대한 지원을 추가합니다.	2020년 10월 1일
할당 전략	AWS Batch 는 인스턴스 유형을 선택하는 여러 전략에 대한 지원을 추가합니다.	2019년 10월 16일

EFA 지원	AWS Batch 는 Elastic Fabric Adapter(EFA) 디바이스에 대한 지원을 추가합니다.	2019년 8월 2일
GPU 일정 예약	AWS Batch 는 GPU 스케줄링을 추가합니다. 이 기능을 사용하면 각 작업에 필요한 GPU 수를 지정하고 그에 따라 AWS Batch 가 인스턴스를 스케일 업할 수 있습니다.	2019년 4월 4일
다중 노드 병렬 작업	AWS Batch 는 다중 노드 병렬 작업에 대한 지원을 추가합니다. 이 기능을 사용하면 여러 Amazon EC2 인스턴스에 걸쳐 있는 단일 작업을 실행할 수 있습니다.	2018년 11월 19일
리소스 수준 권한	AWS Batch 는 여러 API 작업에 대한 리소스 수준 권한을 지원합니다.	2018년 11월 12일
Amazon EC2 시작 템플릿 지원	AWS Batch 는 컴퓨팅 환경에서 시작 템플릿 사용에 대한 지원을 추가합니다.	2018년 11월 12일
AWS Batch 작업 제한 시간	AWS Batch 는 작업 제한 시간에 대한 지원을 추가합니다. 이 지원을 사용하면 작업이 예상보다 오래 실행되면가 작업을 AWS Batch 종료하도록 작업에 대한 특정 제한 시간을 구성할 수 있습니다.	2018년 5월 4일

AWS Batch EventBridge 대상인 작업	AWS Batch 작업은 EventBridge 대상으로 사용할 수 있습니다. 간단한 규칙을 생성하여 이벤트를 일치시키고 이에 대한 응답으로 AWS Batch 작업을 제출할 수 있습니다.	2018년 3월 1일
에 대한 CloudTrail 감사 AWS Batch	CloudTrail은 AWS Batch API 작업에 대한 호출을 감사할 수 있습니다.	2018년 1월 10일
배열 작업	AWS Batch 는 배열 작업에 대한 지원을 추가합니다. 사용자는 파라미터 스윕 및 Monte Carlo 워크로드에 작업을 배열할 수 있습니다.	2017년 11월 28일
확장된 AWS Batch 태그 지정	AWS Batch 는 태그 지정 함수에 대한 지원을 확장합니다. 이 함수를 사용하여 관리형 컴퓨팅 환경에서 시작되는 Amazon EC2 스팟 인스턴스의 태그를 지정할 수 있습니다.	2017년 10월 26일
AWS Batch EventBridge의 이벤트 스트림	AWS Batch 는 EventBridge에 대한 이벤트 스트림을 추가합니다. AWS Batch 이벤트 스트림을 사용하여 작업 대기열에 제출된 작업 상태에 대한 알림을 거의 실시간으로 받을 수 있습니다.	2017년 10월 24일

작업 자동 재시도	AWS Batch 는 작업 재시도에 대한 지원을 추가합니다. 이 업데이트를 통해 작업이 실패할 경우 자동으로 작업을 다시 시도하도록 하는 재시도 전략을 작업과 작업 정의에 적용할 수 있습니다.	2017년 3월 28일
AWS Batch 일반 가용성	AWS Batch 는에서 배치 컴퓨팅 워크로드를 실행할 수 있는 수단으로 설계되었습니다 AWS 클라우드.	2017년 1월 5일

기계 번역으로 제공되는 번역입니다. 제공된 번역과 원본 영어의 내용이 상충하는 경우에는 영어 버전이 우선합니다.