Kerangka Kerja AWS Well-Architected

Pilar Keandalan

Pilar Keandalan: Kerangka Kerja AWS Well-Architected

Table of Contents

Pilar Keandalan - Kerangka Kerja AWS Well-Architected

Tanggal publikasi: 6 November 2024 (Revisi dokumen)

Laporan ini berfokus pada pilar keandalan Kerangka Kerja AWS Well-Architected. Laporan ini menyediakan panduan untuk membantu pelanggan menerapkan praktik terbaik dalam desain, pengiriman, dan pemeliharaan lingkungan Amazon Web Services (AWS).

Pengantar

Kerangka Kerja AWS Well-Architected akan membantu Anda mengetahui kelebihan dan kekurangan dari keputusan yang Anda ambil saat membangun beban kerja di AWS. Dengan menggunakan Kerangka Kerja ini, Anda akan mengetahui praktik-praktik terbaik berkaitan dengan arsitektur untuk mendesain dan mengoperasikan beban kerja yang andal, aman, efisien, hemat biaya, dan ramah lingkungan di cloud. Layanan ini menyediakan cara untuk menilai arsitektur Anda secara terus menerus berdasarkan praktik terbaik dan mengidentifikasi area yang perlu diperbaiki. Kami percaya bahwa memiliki beban kerja yang didesain dengan baik akan meningkatkan peluang keberhasilan bisnis.

Kerangka Kerja AWS Well-Architected berlandaskan pada enam pilar kerangka kerja:

- Keunggulan Operasional
- Keamanan
- Keandalan
- Efisiensi Kinerja
- Pengoptimalan Biaya
- Keberlanjutan

Artikel ini berfokus pada pilar keandalan dan cara menerapkannya ke solusi Anda. Mencapai keandalan dapat menjadi sebuah tantangan dalam lingkungan on-premise tradisional karena titik tunggal kegagalan, kurangnya otomatisasi, dan kurangnya elastisitas. Dengan mengadopsi praktik dalam artikel ini, Anda dapat membangun arsitektur yang memiliki fondasi kuat, arsitektur tangguh, manajemen perubahan yang konsisten, dan proses pemulihan kegagalan yang telah terbukti.

Artikel ini dimaksudkan untuk orang-orang yang memiliki peran di bidang teknologi, seperti kepala pejabat teknologi (CTO), arsitek, developer, dan anggota tim operasi. Setelah membaca artikel ini,

Anda akan memahami praktik terbaik AWS dan strategi yang digunakan ketika merancang arsitektur cloud untuk keandalan. Artikel ini menyertakan detail implementasi tingkat tinggi dan pola arsitektur, juga referensi ke sumber daya tambahan.

Keandalan

Pilar keandalan berkenaan dengan kemampuan beban kerja untuk menjalankan fungsinya dengan benar dan konsisten sesuai dengan yang diharapkan. Ini termasuk kemampuan untuk mengoperasikan dan menguji beban kerja di seluruh siklus hidupnya. Laporan resmi ini berisi panduan praktik terbaik yang mendalam untuk mengimplementasikan beban kerja yang andal di AWS.

Topik

Model Tanggung Jawab Bersama untuk Ketangguhan

Ketangguhan merupakan tanggung jawab bersama antara AWS dan Anda. Anda harus memahami cara pemulihan bencana (DR) dan ketersediaan beroperasi, sebagai bagian dari ketangguhan, menurut model tanggung jawab bersama ini.

tanggung jawab AWS - Ketahanan cloud

AWS bertanggung jawab atas ketangguhan infrastruktur yang menjalankan semua layanan yang ditawarkan di AWS Cloud. Infrastruktur ini terdiri dari perangkat keras, perangkat lunak, jaringan, dan fasilitas yang menjalankan layanan AWS Cloud. AWS menggunakan upaya yang wajar dan secara komersial membuat layanan AWS Cloud ini tersedia, memastikan ketersediaan layanan memenuhi atau melampaui Perjanjian Tingkat Layanan (SLA) AWS.

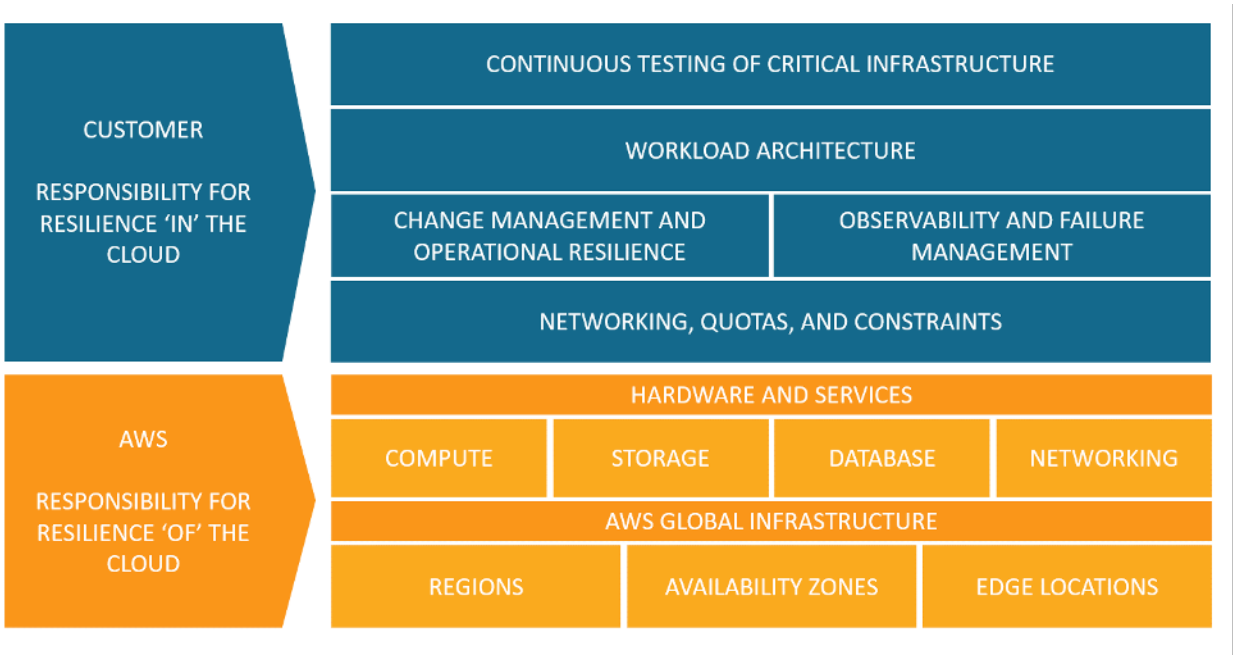
Infrastruktur Cloud Global AWS dirancang untuk memungkinkan pelanggan membangun arsitektur beban kerja yang sangat tangguh. Setiap Wilayah AWS sepenuhnya terisolasi dan terdiri dari beberapa Zona Ketersediaan, yang merupakan partisi infrastruktur yang sepenuhnya terisolasi. Zona Ketersediaan mengisolasi kesalahan yang dapat memengaruhi ketangguhan beban kerja, yang akan mencegahnya untuk memengaruhi zona-zona lain di Wilayah. Tetapi pada saat yang sama, semua zona di Wilayah AWS saling terhubung dengan bandwidth tinggi, jaringan berlatensi rendah, melalui serat metro khusus yang sepenuhnya redundan, yang menyediakan jaringan throughput yang tinggi dan latensi yang rendah antara zona. Semua lalu lintas antara zona dienkripsi. Performa jaringan

cukup untuk mendapatkan replikasi sinkron antara zona. Ketika sebuah aplikasi dipartisi secara lintas AZ, perusahaan akan menjadi lebih terisolasi dan terlindungi dari permasalahan-permasalahan seperti pemadaman listrik, sambaran petir, angin topan, angin puting beliung, dan lain-lain.

Tanggung jawab pelanggan - Ketahanan di cloud

Tanggung jawab Anda ditentukan oleh layanan-layanan AWS Cloud yang Anda pilih. Hal ini akan menentukan jumlah konfigurasi kerja yang harus Anda lakukan sebagai bagian dari tanggung jawab ketangguhan Anda. Contohnya, sebuah layanan seperti Amazon Elastic Compute Cloud (Amazon EC2) mengharuskan pelanggan melakukan semua tugas manajemen dan konfigurasi ketangguhan yang diperlukan. Pelanggan yang menerapkan instans Amazon EC2 bertanggung jawab untuk men-deploy instans Amazon EC2 di beberapa lokasi (seperti Zona Ketersediaan AWS), menerapkan penyembuhan mandiri dengan menggunakan layanan seperti Auto Scaling (penskalaan otomatis), dan menggunakan praktik terbaik arsitektur beban kerja tangguh untuk aplikasi yang diinstal pada instans tersebut. Untuk layanan-layanan terkelola, seperti Amazon S3 dan Amazon DynamoDB, AWS mengoperasikan lapisan infrastruktur, sistem operasi, dan platform, sedangkan pelanggan mengakses titik akhir untuk menyimpan dan mengambil data. Anda bertanggung jawab untuk mengelola ketangguhan data Anda, termasuk strategi pencadangan, penentuan versi, dan replikasi.

Melakukan deployment beban kerja Anda ke beberapa Zona Ketersediaan di Wilayah AWS merupakan bagian dari strategi ketersediaan tinggi yang didesain untuk melindungi beban kerja dengan mengisolasi masalah ke satu Zona Ketersediaan, yang menggunakan redundansi Zona Ketersediaan lain untuk terus melayani permintaan secara berkelanjutan. Arsitektur Multi-AZ juga merupakan bagian dari strategi DR yang didesain untuk membuat beban kerja menjadi lebih terisolasi dan terlindungi dari masalah-masalah seperti pemadaman listrik, sambaran petir, angin topan, gempa bumi, dan lain-lain. Strategi DR juga dapat menggunakan beberapa Wilayah AWS. Contohnya, dalam sebuah konfigurasi aktif/pasif, layanan untuk beban kerja mengalami failover dari Wilayah aktifnya ke Wilayah DR-nya jika Wilayah aktif tidak dapat lagi melayani permintaan.



Tanggung jawab untuk ketahanan di dalam dan dari cloud untuk pelanggan dan AWS.

Anda dapat menggunakan layanan-layanan AWS untuk mencapai sasaran ketangguhan Anda. Sebagai pelanggan, Anda bertanggung jawab atas manajemen aspek-aspek berikut dari sistem Anda untuk mencapai ketangguhan di cloud. Untuk detail lebih lanjut tentang masing-masing layanan secara khusus, lihat dokumentasi AWS.

Jaringan, kuota, dan kendala

- Praktik terbaik untuk area model tanggung jawab bersama ini dijelaskan secara rinci di bagian Landasan.
- Rencanakan arsitektur Anda dengan ruang yang memadai untuk menskalakan dan pahami kuota layanan (service quotas) dan kendala layanan yang Anda sertakan, berdasarkan peningkatan permintaan beban yang diharapkan jika berlaku.
- Rancang desain topologi jaringan Anda agar mempunyai ketersediaan yang tinggi, redundan, dan dapat diskalakan.

Manajemen perubahan dan ketahanan operasional

- Manajemen perubahan mencakup cara memperkenalkan dan mengelola perubahan yang dibuat di lingkungan Anda. Menerapkan perubahan memerlukan pembuatan dan pemeliharaan runbook agar tetap mutakhir dan strategi deployment untuk aplikasi dan infrastruktur Anda.

- Strategi tangguh untuk memantau sumber daya beban kerja harus mempertimbangkan semua komponen, termasuk metrik teknis dan bisnis, notifikasi, otomatisasi, dan analisis.
- Beban kerja di cloud harus beradaptasi dengan perubahan pada penskalaan permintaan sebagai reaksi terhadap gangguan atau fluktuasi penggunaan.

Manajemen observabilitas dan kegagalan

- Mengamati kegagalan melalui pemantauan diperlukan untuk melakukan otomatisasi penyembuhan sehingga beban kerja Anda dapat menahan kegagalan komponen.
- Manajemen kegagalan memerlukan pencadangan data, menerapkan praktik terbaik untuk memungkinkan beban kerja Anda bisa bertahan dari kegagalan komponen, dan perencanaan pemulihan bencana.

Arsitektur beban kerja

- Arsitektur beban kerja Anda mencakup bagaimana Anda merancang layanan-layanan di sekitar domain bisnis, menerapkan SOA dan desain sistem terdistribusi untuk mencegah kegagalan, dan membangun kemampuan seperti throttling, percobaan ulang, manajemen antrean, batas waktu, dan tuas darurat.
- Andalkan solusi AWS yang telah terbukti, Amazon Builders Library, dan pola nirserver untuk menyelaraskan dengan praktik terbaik dan implementasi jump start.
- Gunakan peningkatan berkelanjutan untuk menguraikan sistem Anda menjadi layanan-layanan terdistribusi guna menskalakan dan berinovasi lebih cepat. Gunakan panduan layanan mikro AWS dan opsi layanan terkelola untuk menyederhanakan dan mempercepat kemampuan Anda untuk memperkenalkan perubahan dan melahirkan inovasi.

Pengujian terus-menerus atas infrastruktur penting

- Menguji keandalan adalah pengujian yang dilakukan pada tingkat fungsional, kinerja, dan kekacauan, serta mengadopsi analisis insiden dan praktik game day untuk membangun keahlian dalam menyelesaikan masalah yang tidak dipahami dengan baik.
- Untuk aplikasi cloud all-in maupun aplikasi hibrida, mengetahui perilaku aplikasi ketika ada masalah yang timbul atau ketika ada komponen yang tidak berfungsi akan memampukan Anda untuk pulih dari penghentian dengan cepat dan andal.

- Buat dan dokumentasikan eksperimen yang dapat diulang untuk memahami perilaku sistem Anda ketika operasi tidak berjalan sesuai harapan. Pengujian ini akan membuktikan keefektifan ketangguhan secara keseluruhan dan memberikan Anda lingkaran umpan balik untuk prosedur operasional Anda sebelum menghadapi skenario kegagalan yang sebenarnya.

Prinsip desain

Di cloud, ada beberapa prinsip yang dapat membantu Anda meningkatkan keandalan. Pertimbangkan selalu prinsip-prinsip ini saat kita membahas praktik terbaik:

- Secara otomatis memulihkan data dari kegagalan: Dengan memantau beban kerja untuk indikator kinerja utama (KPI), Anda dapat menjalankan otomatisasi ketika ambang batas terlampaui. KPI ini harus menjadi ukuran dari nilai bisnis, bukan menjadi ukuran dari aspek-aspek teknis operasi layanan. Hal ini akan memungkinkan notifikasi otomatis dan pelacakan kegagalan, serta proses pemulihan otomatis yang menangani atau memperbaiki kegagalan. Dengan otomatisasi yang lebih canggih, antisipasi dan perbaikan kegagalan dapat dilakukan sebelum kegagalan itu terjadi.
- Prosedur pemulihan pengujian: Dalam lingkungan on-premise, pengujian biasanya dijalankan untuk membuktikan bahwa beban kerja bekerja dalam skenario tertentu. Pengujian biasanya tidak digunakan untuk memvalidasi strategi pemulihan. Di cloud, Anda dapat melakukan pengujian tentang bagaimana beban kerja mengalami kegagalan, dan Anda juga dapat memvalidasi prosedur-prosedur pemulihan. Anda dapat menggunakan otomatisasi untuk memicu berbagai kegagalan atau untuk membuat ulang skenario yang mengarah pada kegagalan sebelumnya. Pendekatan ini akan memperlihatkan jalur kegagalan yang dapat diuji dan diperbaiki sebelum sebuah skenario kegagalan benar-benar terjadi, sehingga dapat mengurangi risiko.
- Tingkatkan (skalakan) kapasitas secara horizontal untuk meningkatkan ketersediaan keseluruhan beban kerja: Ganti satu sumber daya besar dengan beberapa sumber daya kecil untuk mengurangi dampak kegagalan tunggal terhadap beban kerja keseluruhan. Distribusikan permintaan ke beberapa sumber daya yang lebih kecil untuk memastikan bahwa tidak terdapat titik kegagalan yang sama.
- Berhenti mengira-ngira besar kapasitas: Penyebab umum kegagalan dalam beban kerja on-premise adalah saturasi sumber daya, yaitu ketika permintaan yang ditempatkan di beban kerja melebihi kapasitas beban kerjanya (ini sering kali menjadi sasaran dari serangan denial of service). Di cloud, Anda dapat memantau pemanfaatan beban kerja dan permintaan, serta melakukan otomatisasi terhadap penambahan atau penghapusan sumber daya untuk mempertahankannya agar berada di tingkat optimal dalam memenuhi permintaan tanpa mengalami kekurangan atau

kelebihan penyediaan. Batasan tentunya masih ada, tetapi sebagian kuota dapat dikontrol dan sebagian lainnya dapat dikelola (lihat Kelola Kuota Layanan (Service Quotas) dan Pembatasan).

- Kelola perubahan dengan otomatisasi: Perubahan-perubahan yang dibuat untuk infrastruktur Anda harus dibuat dengan menggunakan otomatisasi. Perubahan-perubahan yang harus dikelola mencakup perubahan yang dibuat untuk otomatisasi, yang kemudian dapat dilacak dan ditinjau.

Definisi

Laporan resmi ini membahas tentang keandalan di cloud, yang menjelaskan praktik terbaik untuk keempat area ini:

- Fondasi
- Arsitektur Beban Kerja
- Manajemen Perubahan
- Manajemen Kegagalan

Untuk mencapai keandalan, Anda harus mengawalinya dengan fondasi—sebuah lingkungan tempat kuota layanan (service quotas) dan topologi jaringan mengakomodasi beban kerja. Arsitektur beban kerja sistem terdistribusi harus didesain untuk dapat mencegah dan memitigasi kegagalan. Beban kerja harus menangani perubahan-perubahan yang terjadi dalam permintaan dan persyaratan, serta harus didesain untuk dapat mendeteksi kegagalan dan melakukan pemulihan mandiri secara otomatis.

Topik

- Ketahanan, dan komponen keandalan
- Ketersediaan
- Tujuan Pemulihan Bencana (DR)

Ketahanan, dan komponen keandalan

Keandalan beban kerja di cloud bergantung pada sejumlah faktor, dan faktor utamanya adalah Ketahanan:

- Ketahanan adalah kemampuan beban kerja untuk pulih dari gangguan infrastruktur atau layanan, secara dinamis memperoleh sumber daya komputasi untuk memenuhi permintaan, dan memitigasi gangguan, seperti kesalahan konfigurasi atau masalah jaringan sementara.

Faktor-faktor lain yang memengaruhi keandalan beban kerja adalah sebagai berikut:

- Keunggulan Operasional, yang mencakup otomatisasi perubahan, penggunaan playbook untuk merespons kegagalan, dan Peninjauan Kesiapan Operasional (ORR) untuk mengonfirmasi bahwa aplikasi sudah siap untuk operasi produksi.
- Keamanan, yang mencakup pencegahan bahaya terhadap data atau infrastruktur yang mungkin disebabkan oleh pelaku kejahatan, yang dapat mengganggu ketersediaan. Misalnya, enskripsi cadangan untuk memastikan keamanan data.
- Efisiensi Kinerja, yang mencakup pembuatan desain untuk tingkat permintaan maksimum dan meminimalkan latensi untuk beban kerja Anda.
- Optimasi Biaya, yang mencakup kompromi-kompromi seperti pilihan untuk mengeluarkan lebih banyak biaya untuk instans EC2 guna mencapai stabilitas statis, atau untuk mengandalkan penskalaan otomatis saat kapasitas yang lebih besar diperlukan.

Ketahanan adalah fokus utama laporan resmi ini.

Empat aspek lainnya juga penting dan tercakup ke dalam masing-masing pilar Kerangka Kerja AWS Well-Architected. Banyak praktik terbaik yang diuraikan di sini juga menangani aspek-aspek keandalan tersebut, namun yang difokuskan di sini adalah ketahanan.

Ketersediaan

Ketersediaan (juga dikenal dengan ketersediaan layanan) adalah metrik yang umum digunakan untuk mengukur ketahanan secara kuantitatif, serta merupakan target tujuan ketahanan.

- Ketersediaan adalah persentase waktu ketika beban kerja tersedia untuk digunakan.

Tersedia untuk digunakan artinya beban kerja berhasil memberikan kinerja sesuai fungsinya yang disepakati ketika diperlukan.

Persentase ini dihitung dengan jangka waktu tertentu, seperti bulan, tahun, atau tiga tahun ke belakang. Dengan memberlakukan interpretasi seketat mungkin, maka ketersediaan akan berkurang

setiap kali aplikasi tidak beroperasi secara normal, termasuk gangguan terjadwal dan di luar jadwal. Kami mendefinisikan ketersediaan seperti berikut:

$$Availability = \frac{Available\ for\ Use\ Time}{Total\ Time}$$

- Ketersediaan adalah persentase waktu aktif (seperti 99,9%) selama jangka waktu tertentu (umumnya satu bulan atau tahun)
- Penulisan singkat yang umum digunakan adalah “jumlah angka sembilan”; misalnya, “lima angka sembilan” artinya tingkat ketersediaannya adalah 99,999%
- Beberapa pelanggan memilih untuk mengecualikan waktu henti layanan terjadwal (misalnya pemeliharaan terencana) dari Total Waktu dalam formula. Namun demikian, ini tidak disarankan, karena pengguna Anda kemungkinan ingin menggunakan layanan Anda selama waktu-waktu tersebut.

Berikut ini adalah tabel tujuan desain ketersediaan aplikasi umum dan durasi maksimum gangguan yang dapat terjadi dalam satu tahun namun tetap memenuhi tujuan-tujuan yang sudah ditetapkan. Tabel ini berisi contoh-contoh tipe aplikasi yang umum kita lihat pada tiap-tiap tingkatan ketersediaan. Di sepanjang dokumen ini, kita mengacu pada nilai-nilai ini.

Ketersediaan	Ketidaktersediaan Maksimum (per tahun)	Kategori Aplikasi
99%	3 hari 15 jam	Pemrosesan batch, ekstraksi data, transfer, dan tugas-tugas pemuatan
99,9%	8 jam 45 menit	Alat internal seperti manajemen pengetahuan, pelacakan proyek
99,95%	4 jam 22 menit	Niaga online, titik penjualan
99,99%	52 menit	Pengiriman video, beban kerja siaran

Ketersediaan	Ketidaktersediaan Maksimum (per tahun)	Kategori Aplikasi
99,999%	5 menit	Transaksi ATM, beban kerja telekomunikasi

Mengukur ketersediaan berdasarkan permintaan. Untuk layanan Anda, mungkin akan lebih mudah untuk menghitung permintaan yang berhasil dan gagal daripada “waktu yang tersedia untuk digunakan”. Dalam kasus ini, dapat digunakan hitungan berikut:

$$Availability = \frac{Successful\ Responses}{Valid\ Requests}$$

Ini sering kali dihitung untuk periode satu menit atau lima menit. Lalu, persentase waktu aktif bulanan (pengukuran ketersediaan berbasis waktu) dapat dihitung dari rata-rata semua periode ini. Jika tidak ada permintaan yang diterima dalam jangka waktu tertentu, maka aplikasi terhitung 100% tersedia untuk waktu tersebut.

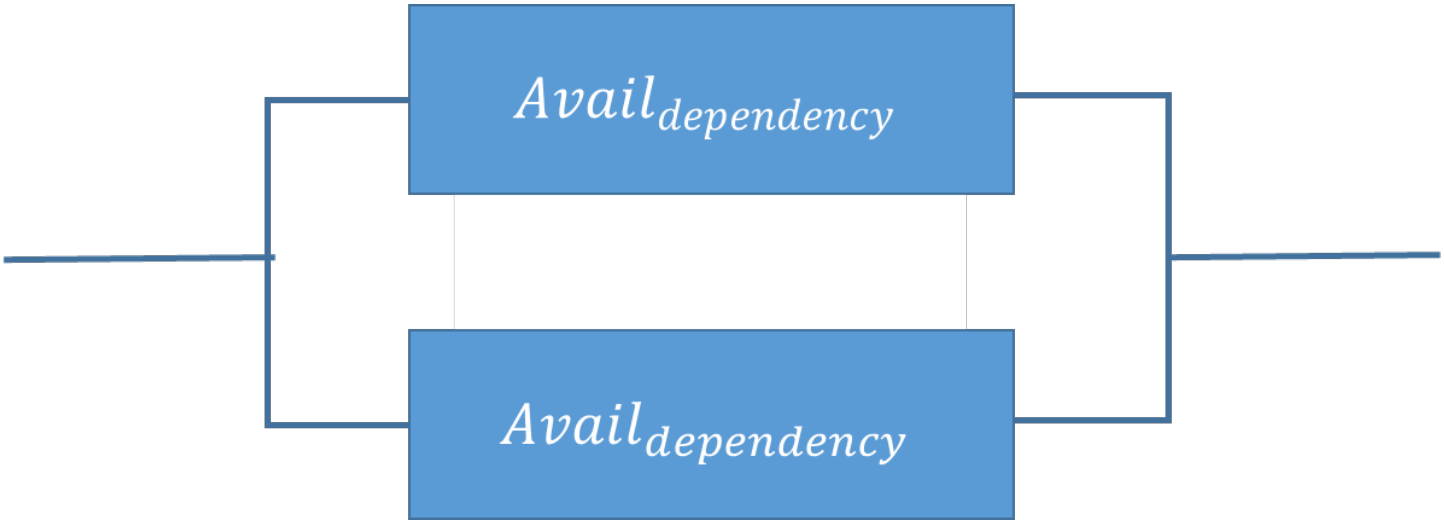
Menghitung ketersediaan dengan dependensi keras. Banyak sekali sistem yang memiliki dependensi keras (hard dependency) pada sistem lain, di mana gangguan pada sebuah sistem dependen secara langsung diartikan sebagai gangguan pada sistem yang menginvokasi. Ini adalah kebalikan dari dependensi lembut (soft dependency), di mana sebuah kegagalan yang terjadi pada sistem dependen diimbangi di dalam aplikasi. Ketika dependensi keras terjadi, ketersediaan sistem yang menginvokasi adalah produk dari ketersediaan sistem dependen tersebut. Misalnya, jika Anda memiliki sistem yang dirancang untuk ketersediaan 99,99% yang memiliki dependensi keras pada dua sistem independen lain yang masing-masing dirancang untuk ketersediaan 99,99%, maka secara teori beban kerja tersebut dapat meraih ketersediaan 99,97%:

$Avail_{invok} \times Avail_{dep1} \times Avail_{dep2} = Avail_{workload}$

99,99% × 99,99% × 99,99% = 99,97%

Oleh karena itu, penting untuk memahami dependensi Anda serta tujuan-tujuan desain ketersediaannya saat Anda menghitung ketersediaan Anda sendiri.

Menghitung ketersediaan dengan komponen redundan. Ketika suatu sistem melibatkan penggunaan komponen redundan dan independen (misalnya sumber daya redundan di beberapa Zona Ketersediaan yang berbeda), maka secara teori, ketersediaan sistem tersebut dihitung 100% dikurangi hasil tingkat kegagalan komponen. Misalnya, jika suatu sistem memanfaatkan dua komponen independen, masing-masing memiliki ketersediaan 99,9%, maka ketersediaan efektif dependensi ini adalah 99,9999%:



$Avail_{effective} = Avail_{MAX} - ((100\% - Avail_{dependency}) \times (100\% - Avail_{dependency}))$

$99{,}9999\% = 100\% - (0{,}1\% \times 0{,}1\%)$

Perhitungan pintasan: Jika ketersediaan semua komponen di dalam hitungan Anda hanya terdiri dari angka sembilan, maka Anda dapat menjumlah hitungan jumlah angka sembilan untuk mendapatkan jawaban. Pada contoh di atas, dua komponen independen dan redundan dengan ketersediaan tiga angka sembilan menghasilkan enam angka sembilan.

Menghitung ketersediaan dependensi. Beberapa dependensi menyediakan panduan tentang ketersediaannya, termasuk tujuan desain ketersediaan untuk banyak layanan AWS. Tetapi pada kasus-kasus di mana ini tidak tersedia (misalnya, komponen yang produsennya tidak mempublikasikan informasi ketersediaan), satu cara untuk menghitung adalah dengan menentukan Waktu Rata-rata Antara Kegagalan (MTBF) dan Waktu Rata-rata Pemulihan (MTTR). Hitungan ketersediaan dapat dilakukan dengan:

$$Avail_{EST} = \frac{MTBF}{MTBF + MTTR}$$

Misalnya, jika MTBF 150 hari dan MTTR 1 jam, maka hitungan ketersediaannya adalah 99,97%.

Untuk detail tambahan, silakan lihat Availability and Beyond: Memahami dan meningkatkan ketahanan sistem terdistribusi AWS, yang dapat membantu Anda menghitung ketersediaan Anda.

Biaya ketersediaan. Merancang desain aplikasi untuk tingkat ketersediaan yang lebih tinggi umumnya akan menyebabkan terjadinya peningkatan biaya, sehingga sebaiknya Anda mengidentifikasi kebutuhan ketersediaan yang sebenarnya sebelum mulai merancang desain aplikasi Anda. Tingkat ketersediaan yang tinggi menghadirkan persyaratan pengujian dan validasi yang lebih ketat berdasarkan skenario kegagalan yang lengkap. Tingkat ketersediaan ini memerlukan otomatisasi untuk pemulihan dari semua bentuk kegagalan, dan mengharuskan semua aspek yang ada dalam operasi sistem tersebut dibangun serupa dan diuji dengan standar yang sama. Misalnya, penambahan atau penghapusan kapasitas, deployment atau pembatalan (roll back) pembaruan perangkat lunak atau perubahan konfigurasi, atau migrasi data sistem harus dilakukan sesuai tujuan-tujuan ketersediaan yang dikehendaki. Selain biaya untuk pengembangan perangkat lunak, pada tingkat ketersediaan yang sangat tinggi, inovasi juga terkena dampak dikarenakan adanya kebutuhan untuk bergerak lebih lambat dalam melakukan deployment sistem. Oleh karena itu, panduan harus dibuat menyeluruh dalam menerapkan standar dan mempertimbangkan target-target ketersediaan yang tepat untuk seluruh siklus hidup pengoperasian sistem.

Faktor lain yang menyebabkan peningkatan biaya dalam sistem yang beroperasi dengan tujuan desain ketersediaan yang lebih tinggi adalah pemilihan dependensi. Pada tujuan-tujuan yang lebih tinggi ini, serangkaian perangkat lunak atau layanan yang dapat dipilih sebagai dependensi berkurang berdasarkan layanan mana yang telah memiliki investasi mendalam yang telah kami jelaskan sebelumnya untuk Anda. Seiring dengan meningkatnya tujuan-tujuan desain ketersediaan, merupakan hal umum untuk menemukan lebih sedikit layanan multi-tujuan (seperti basis data relasional) dan lebih banyak layanan yang dibuat khusus. Alasannya adalah bahwa jenis-jenis layanan yang dibuat khusus lebih mudah untuk dievaluasi, diuji, dan diotomatisasi, serta memiliki lebih sedikit potensi interaksi kejutan dengan fungsionalitas yang disertakan namun tidak digunakan.

Tujuan Pemulihan Bencana (DR)

Selain tujuan ketersediaan, strategi ketahanan Anda juga harus menyertakan tujuan-tujuan Pemulihan Bencana (DR) berdasarkan strategi yang digunakan untuk memulihkan beban kerja Anda apabila terjadi peristiwa bencana. Pemulihan Bencana berfokus pada tujuan-tujuan pemulihan satu kali untuk merespons terjadinya bencana alam, kegagalan teknis skala besar, atau ancaman manusia seperti serangan atau kesalahan. Ini berbeda dengan ketersediaan yang mengukur rata-rata

ketahanan selama kurun waktu tertentu ketika merespons terjadinya kegagalan komponen, lonjakan beban, atau bug perangkat lunak.

sasaran waktu pemulihan (RTO) Didefinisikan oleh organisasi. RTO adalah jeda waktu maksimum yang dapat diterima antara gangguan layanan dan pemulihan layanan. Ini menentukan apa yang dianggap sebagai jendela waktu yang dapat diterima ketika layanan tidak tersedia.

Sasaran Titik Pemulihan (RPO) Didefinisikan oleh organisasi. RPO adalah jumlah waktu maksimum yang dapat diterima sejak titik pemulihan data terakhir. Ini menentukan apa yang dianggap sebagai kehilangan data yang dapat diterima antara titik pemulihan terakhir dan gangguan layanan.

Hubungan antara RPO (Sasaran Titik Pemulihan), RTO (Sasaran Waktu Pemulihan), dan peristiwa bencana.

RTO serupa dengan MTTR (Rata-Rata Waktu hingga Pemulihan) yakni keduanya sama-sama mengukur waktu antara sejak dari dimulainya penghentian (outage) hingga pemulihan beban kerja. Namun, MTTR adalah nilai rata-rata yang diambil selama beberapa peristiwa yang berdampak pada ketersediaan dalam kurun waktu tertentu, sedangkan RTO adalah target, atau nilai maksimum yang diizinkan, untuk satu peristiwa yang berdampak terhadap ketersediaan.

Memahami kebutuhan ketersediaan

Merupakan hal normal ketika di awal, ketersediaan suatu aplikasi dianggap sebagai satu target tunggal untuk aplikasi secara keseluruhan. Namun demikian, setelah diselidiki lebih teliti, kita sering

menemukan bahwa aspek-aspek tertentu yang ada dalam suatu aplikasi atau layanan memiliki persyaratan ketersediaan yang berbeda. Sebagai contoh, beberapa sistem mungkin mengutamakan kemampuan untuk menerima dan menyimpan data baru sebelum mengambil data yang sudah ada. Sedangkan sistem lainnya mengutamakan operasi waktu nyata daripada operasi yang akan mengubah konfigurasi atau lingkungan sistem. Layanan-layanan mungkin memiliki persyaratan ketersediaan yang sangat tinggi selama jam tertentu, tetapi dapat memberikan toleransi terhadap gangguan yang jauh lebih lama di luar jam-jam tersebut. Ini adalah beberapa cara untuk mengurai satu aplikasi menjadi bagian-bagian komponen, dan mengevaluasi persyaratan ketersediaan untuk masing-masing komponen. Manfaat dari melakukan hal ini adalah untuk memfokuskan upaya (serta pengeluaran) Anda pada ketersediaan berdasarkan kebutuhan khusus, bukan merekayasa keseluruhan sistem sesuai persyaratan paling ketat.

Rekomendasi
Secara kritis, lakukan evaluasi terhadap aspek-aspek unik untuk aplikasi Anda dan, jika perlu, bedakan tujuan desain ketersediaan dan pemulihan bencana untuk mencerminkan kebutuhan bisnis Anda.

Di dalam AWS, kita umumnya membagi layanan ke dalam “bidang data” dan “bidang kontrol.” Bidang data bertanggung jawab untuk menghadirkan layanan waktu nyata sedangkan bidang kontrol digunakan untuk mengonfigurasi lingkungan. Misalnya, instans Amazon EC2, basis data Amazon RDS, dan operasi baca/tulis tabel Amazon DynamoDB semuanya adalah operasi bidang data. Sebaliknya, peluncuran instans EC2 atau basis data RDS baru, atau penambahan atau pengubahan metadata tabel di DynamoDB dianggap sebagai operasi bidang kontrol. Meskipun tingkat ketersediaan yang tinggi penting untuk semua kemampuan ini, bidang data umumnya memiliki tujuan desain ketersediaan yang lebih tinggi daripada bidang kontrol. Oleh karena itu, beban kerja dengan persyaratan ketersediaan yang tinggi harus menghindari dependensi run-time pada operasi bidang kontrol.

Banyak pelanggan AWS menerapkan pendekatan serupa dalam mengevaluasi aplikasi mereka secara kritis dan mengidentifikasi sub-komponen dengan kebutuhan ketersediaan yang berbeda-beda. Tujuan desain ketersediaan kemudian disesuaikan dengan berbagai aspek, dan upaya kerja yang tepat dijalankan untuk merekayasa sistem. AWS memiliki banyak pengalaman dalam merekayasa aplikasi dengan berbagai tujuan desain ketersediaan, termasuk layanan dengan ketersediaan 99,999% atau lebih. AWS Solution Architects (SA) dapat membantu Anda merancang tujuan ketersediaan Anda dengan baik. Melibatkan AWS sejak awal dalam proses desain Anda

dapat meningkatkan kemampuan kami untuk membantu Anda memenuhi tujuan-tujuan ketersediaan Anda. Perencanaan ketersediaan tidak hanya dilakukan sebelum peluncuran beban kerja Anda. Perencanaan ini juga dilakukan secara berkelanjutan untuk menyempurnakan desain Anda seiring Anda mendapatkan pengalaman operasional, belajar dari peristiwa dunia nyata, dan bertahan di tengah berbagai jenis kegagalan yang terjadi. Anda kemudian dapat menerapkan upaya kerja yang tepat untuk melakukan perbaikan setelah implementasi Anda.

Kebutuhan ketersediaan yang diperlukan untuk suatu beban kerja harus diselaraskan dengan kebutuhan dan kekritisan bisnis. Dengan menetapkan terlebih dulu kerangka kerja kekritisan bisnis dengan RTO, RPO, dan ketersediaan yang ditetapkan, Anda dapat menilai masing-masing beban kerja. Pendekatan seperti ini akan mengharuskan orang-orang yang terlibat di dalam implementasi beban kerja untuk memahami kerangka kerja tersebut, dan dampak yang diberikan oleh beban kerja mereka terhadap kebutuhan bisnis.

Fondasi

Persyaratan mendasar memiliki cakupan yang lebih luas dari satu beban kerja atau proyek. Sebelum merancang sistem apa pun, persyaratan mendasar yang memengaruhi keandalan harus diterapkan. Misalnya, Anda harus memiliki bandwidth jaringan yang memadai untuk pusat data Anda.

Di lingkungan on-premise, persyaratan ini dapat menyebabkan waktu jeda yang lama yang disebabkan oleh dependensi sehingga harus disertakan selama perencanaan awal. Namun, dengan AWS, sebagian besar persyaratan mendasar ini sudah disertakan atau dapat ditangani sesuai kebutuhan. Cloud dirancang agar bersifat hampir tidak terbatas, sehingga AWS bertanggung jawab untuk memenuhi persyaratan jaringan dan kapasitas komputasi yang memadai, agar Anda dapat dengan leluasa mengubah alokasi dan ukuran sumber daya sesuai permintaan.

Bagian-bagian berikutnya akan menjelaskan praktik terbaik yang berfokus pada pertimbangan keandalan ini.

Topik

- Mengelola kuota layanan (service quotas) dan kendala
- Merencanakan topologi jaringan Anda

Mengelola kuota layanan (service quotas) dan kendala

Untuk arsitektur beban kerja berbasis cloud, ada service quotas (kuota layanan) (yang juga disebut sebagai batas layanan). Kuota ini ada untuk mencegah penyediaan sumber daya lebih yang tidak disengaja melebihi kebutuhan Anda dan untuk membatasi tingkat permintaan di operasi API sehingga melindungi layanan dari penyalahgunaan. Ada juga kendala-kendala sumber daya, misalnya, laju Anda dapat mendorong bit di kabel serat optik, atau jumlah penyimpanan di diska secara fisik.

Jika Anda menggunakan aplikasi AWS Marketplace, Anda harus memahami batasan-batasan penggunaannya. Jika Anda menggunakan layanan web atau perangkat lunak sebagai layanan pihak ketiga, maka Anda juga harus mengetahui batasan-batasannya.

Praktik terbaik

- REL01-BP01 Kesadaran tentang kuota (service quotas) dan kendala layanan
- REL01-BP02 Mengelola kuota layanan (service quotas) di seluruh akun dan wilayah

- REL01-BP03 Mengakomodasi kuota layanan (service quotas) tetap dan kendala melalui arsitektur
- REL01-BP04 Memantau dan mengelola kuota
- REL01-BP05 Mengotomatiskan manajemen kuota
- REL01-BP06 Memastikan adanya selisih yang memadai antara kuota saat ini dan penggunaan maksimum untuk mengakomodasi failover

REL01-BP01 Kesadaran tentang kuota (service quotas) dan kendala layanan

Perhatikan kuota default Anda dan kelola permintaan penambahan kuota untuk arsitektur beban kerja Anda. Ketahui kendala sumber daya cloud mana, seperti disk atau jaringan, yang berpotensi memberi dampak.

Hasil yang diinginkan: Pelanggan dapat mencegah terjadinya degradasi atau gangguan layanan dalam Akun AWS mereka dengan menerapkan pedoman yang tepat untuk melakukan pemantauan metrik utama, peninjauan infrastruktur, dan langkah-langkah remediasi otomatisasi untuk memverifikasi bahwa kuota layanan (service quotas) dan batas layanan tidak tercapai, dan hal ini dapat menyebabkan terjadinya degradasi atau gangguan terhadap layanan.

Anti-pola umum:

- Melakukan deployment beban kerja tanpa memahami kuota keras dan lunak serta batasan-batasannya untuk layanan yang digunakan.
- Melakukan deployment beban kerja pengganti tanpa menganalisis dan mengonfigurasi ulang kuota yang diperlukan atau menghubungi tim Dukungan sebelumnya.
- Berasumsi bahwa layanan cloud tidak memiliki batasan dan layanan dapat digunakan tanpa mempertimbangkan angka permintaan, batas, hitungan, dan kuantitas.
- Berasumsi bahwa kuota akan ditingkatkan secara otomatis.
- Tidak mengetahui proses dan lini waktu permintaan kuota.
- Berasumsi bahwa kuota layanan (service quotas) cloud default selalu sama untuk setiap layanan di semua wilayah.
- Berasumsi bahwa kendala layanan dapat ditembus dan sistem akan diskalakan secara otomatis dan meningkatkan batas di luar kendala sumber daya
- Tidak menguji aplikasi saat lalu lintas memuncak untuk membebani pemanfaatan sumber dayanya.

- Melakukan penyediaan sumber daya tanpa melakukan analisis terhadap ukuran sumber daya yang diperlukan.
- Melakukan penyediaan berlebihan dalam hal kapasitas dengan memilih jenis sumber daya yang jauh melampaui kebutuhan riil atau perkiraan lalu lintas puncak.
- Tidak menilai persyaratan kapasitas untuk tingkat lalu lintas yang baru sebelum adanya peristiwa pelanggan baru atau deployment teknologi baru.

Manfaat menerapkan praktik terbaik ini: Pemantauan dan manajemen otomatis kuota layanan (service quotas) dan kendala sumber daya dapat mengurangi kegagalan secara proaktif. Perubahan pada pola lalu lintas untuk layanan pelanggan dapat menyebabkan terjadinya gangguan atau penurunan kualitas jika praktik terbaik tidak diikuti. Dengan memantau dan mengelola nilai-nilai ini di semua wilayah dan semua akun, aplikasi dapat memiliki ketahanan yang lebih baik saat terjadi peristiwa yang tidak direncanakan atau tidak diinginkan.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Tinggi

Panduan implementasi

Service Quotas adalah sebuah layanan AWS yang akan membantu Anda mengelola kuota Anda untuk lebih dari 250 layanan AWS dari satu lokasi. Di samping mencari nilai kuota, Anda juga dapat meminta dan melacak peningkatan kuota dari konsol Service Quotas atau menggunakan SDK AWS. AWS Trusted Advisor menawarkan pemeriksaan kuota layanan (service quotas) yang menampilkan penggunaan dan kuota Anda untuk beberapa aspek dari beberapa layanan. Service Quotas default per layanan juga ada di dalam dokumentasi AWS berdasarkan layanan masing-masing (misalnya, lihat Kuota Amazon VPC).

Beberapa batas layanan, seperti batas tingkat pada API yang diberikan throttling diatur di Amazon API Gateway itu sendiri dengan mengonfigurasi rencana penggunaan. Beberapa batasan yang ditetapkan sebagai konfigurasi pada layanan masing-masing termasuk IOPS yang Disediakan, penyimpanan Amazon RDS yang dialokasikan, dan alokasi volume Amazon EBS. Amazon Elastic Compute Cloud memiliki dasbor batas layanannya sendiri yang akan dapat membantu Anda mengelola instans, Amazon Elastic Block Store, dan batas-batas alamat IP Elastic. Jika Anda memiliki kasus penggunaan di mana kuota layanan (service quotas) memengaruhi kinerja aplikasi Anda dan tidak dapat disesuaikan dengan kebutuhan Anda, hubungi Dukungan untuk mengetahui apakah terdapat langkah mitigasi.

Kuota layanan (service quotas) bisa menurut Wilayah atau bersifat global. Menggunakan sebuah layanan AWS yang mencapai kuotanya, tidak akan memberikan manfaat yang diharapkan

sebagaimana dalam penggunaan normal dan dapat menyebabkan terjadinya gangguan atau penurunan kualitas layanan. Misalnya, sebuah kuota layanan (service quotas) membatasi jumlah instans DL Amazon EC2 yang digunakan di Wilayah. Batas tersebut dapat dicapai selama peristiwa penskalaan lalu lintas dengan menggunakan grup Auto Scaling (ASG).

Kuota layanan (service quotas) untuk setiap akun harus dinilai secara rutin dalam hal penggunaan untuk menentukan batas layanan yang tepat untuk akun tersebut. Kuota layanan (service quotas) ini dibuat sebagai pagar pembatas operasional, agar Anda tidak melakukan pengadaan sumber daya lebih dari yang dibutuhkan tanpa disadari. Kuota layanan (service quotas) juga berfungsi untuk membatasi angka permintaan pada operasi API guna melindungi layanan dari penyalahgunaan.

Kendala layanan berbeda dari kuota layanan (service quotas). Kendala-kendala layanan mewakili batasan sumber daya tertentu yang ditentukan oleh jenis sumber daya tersebut. Kendala tersebut dapat berupa kapasitas penyimpanan (misalnya, gp2 memiliki batas ukuran 1 GB - 16 TB) atau throughput diska. Sangat penting untuk merancang dan terus menilai kendala jenis sumber daya untuk penggunaan yang mungkin mencapai batasnya. Jika suatu kendala tercapai di luar perkiraan, maka aplikasi dan layanan akun dapat mengalami gangguan atau mengalami penurunan kualitas.

Jika terdapat kasus penggunaan di mana kuota layanan (service quotas) memengaruhi kinerja aplikasi dan tidak dapat disesuaikan dengan kebutuhan, hubungi Dukungan untuk mengetahui apakah terdapat langkah mitigasi. Untuk detail lebih lanjut tentang cara menyesuaikan kuota tetap, lihat REL01-BP03 Mengakomodasi kuota layanan (service quotas) tetap dan kendala melalui arsitektur.

Terdapat sejumlah layanan dan alat AWS yang dapat membantu Anda dalam melakukan pemantauan dan pengelolaan Service Quotas. Layanan dan alat harus dimanfaatkan untuk menyediakan pemeriksaan level kuota otomatis atau manual.

- AWS Trusted Advisor menawarkan pemeriksaan kuota layanan (service quotas) yang menampilkan penggunaan dan kuota Anda untuk beberapa aspek dari beberapa layanan. Alat ini dapat membantu Anda mengidentifikasi layanan yang sudah mendekati kuota.
- Konsol Manajemen AWS menyediakan metode untuk menampilkan nilai kuota layanan (service quotas), mengelola, meminta kuota baru, memantau status permintaan kuota, dan menampilkan riwayat kuota.
- AWS CLI dan CDK menawarkan metode terprogram untuk mengelola dan memantau level serta penggunaan kuota layanan (service quotas) secara otomatis.

Langkah-langkah implementasi

Untuk Service Quotas:

- Tinjau AWS Service Quotas.
- Untuk mengetahui kuota layanan (service quotas) Anda saat ini, tentukan layanan (seperti IAM Access Analyzer) yang digunakan. Terdapat sekitar 250 layanan AWS yang dikontrol oleh kuota layanan (service quotas). Lalu, tentukan nama kuota layanan (service quotas) tertentu yang dapat digunakan di dalam setiap akun dan Wilayah. Terdapat sekitar 3000 nama kuota layanan (service quotas) per Wilayah.
- Tambahkan analisis kuota ini dengan AWS Config untuk menemukan semua sumber daya AWS yang digunakan dalam Akun AWS Anda.
- Gunakan data AWS CloudFormation untuk menentukan sumber daya AWS yang Anda gunakan. Lihatlah sumber daya yang dibuat baik di Konsol Manajemen AWS atau dengan perintah `list-stack-resources` AWS CLI. Anda juga dapat melihat sumber daya yang dikonfigurasi untuk di-deploy di templat itu sendiri.
- Tentukan semua layanan yang diperlukan oleh beban kerja Anda dengan melihat kode deployment.
- Tentukan kuota layanan (service quotas) yang berlaku. Gunakan informasi yang dapat diakses secara terprogram dari Trusted Advisor dan Service Quotas.
- Tetapkan metode pemantauan otomatis (lihat REL01-BP02 Mengelola kuota layanan (service quotas) di seluruh akun dan wilayah dan REL01-BP04 Memantau dan mengelola kuota) untuk memperingatkan dan menginformasikan apakah kuota layanan (service quotas) sudah dekat atau telah mencapai batasnya.
- Tetapkan metode otomatis dan terprogram untuk memeriksa apakah kuota layanan (service quotas) telah diubah di satu wilayah tetapi tidak di wilayah lain dalam akun yang sama (lihat REL01-BP02 Mengelola kuota layanan (service quotas) di seluruh akun dan wilayah dan REL01-BP04 Memantau dan mengelola kuota).
- Lakukan otomatisasi terhadap pemindaian log dan metrik aplikasi untuk menentukan apakah terdapat kesalahan kuota atau kendala layanan yang terjadi. Jika terdapat kesalahan, kirimkan peringatan ke sistem pemantauan.
- Tetapkan prosedur teknik untuk menghitung perubahan kuota yang diperlukan (lihat REL01-BP05 Mengotomatiskan manajemen kuota) setelah diidentifikasi bahwa kuota yang lebih besar diperlukan untuk layanan tertentu.
- Buat alur kerja pengadaan dan persetujuan untuk meminta perubahan dalam kuota layanan (service quotas). Sertakan di dalamnya alur kerja pengecualian untuk mengantisipasi jika permintaan ditolak atau disetujui sebagian.

- Buat metode rekayasa untuk meninjau kuota layanan (service quotas) sebelum pengadaan dan menggunakan layanan AWS baru sebelum digulirkan ke produksi atau lingkungan yang berisi data (misalnya akun pengujian beban).

Untuk kendala layanan:

- Bangun metode pemantauan dan metrik untuk memberi peringatan jika terdapat pembacaan sumber daya yang mendekati kendala sumber dayanya. Manfaatkan CloudWatch apabila diperlukan untuk melakukan pemantauan metrik atau log.
- Tetapkan ambang batas peringatan untuk masing-masing sumber daya yang memiliki kendala yang dapat berpengaruh terhadap aplikasi atau sistem.
- Ciptakan prosedur manajemen alur kerja dan infrastruktur untuk mengubah jenis sumber daya jika pemanfaatan mendekati kendala. Alur kerja ini harus mencakup pengujian beban sebagai sebuah praktik terbaik untuk memverifikasi bahwa jenis sumber daya baru tersebut sudah tepat dengan kendala-kendala baru.
- Migrasikan sumber daya yang diidentifikasi ke jenis sumber daya baru yang disarankan, dengan menggunakan prosedur dan proses yang ada.

Sumber daya

Praktik-praktik terbaik terkait:

- REL01-BP02 Mengelola kuota layanan (service quotas) di seluruh akun dan wilayah
- REL01-BP03 Mengakomodasi kuota layanan (service quotas) tetap dan kendala melalui arsitektur
- REL01-BP04 Memantau dan mengelola kuota
- REL01-BP05 Mengotomatiskan manajemen kuota
- REL01-BP06 Memastikan adanya selisih yang memadai antara kuota saat ini dan penggunaan maksimum untuk mengakomodasi failover
- REL03-BP01 Pilih cara mengelompokkan beban kerja Anda
- REL10-BP01 Melakukan deployment beban kerja ke beberapa lokasi
- REL11-BP01 Memantau semua komponen beban kerja untuk mendeteksi kegagalan
- REL11-BP03 Melakukan otomatisasi pemulihan di semua lapisan
- REL12-BP04 Menguji ketahanan menggunakan chaos engineering

Dokumen terkait:

- AWS Pilar Keandalan Kerangka Kerja Well-Architected: Ketersediaan
- AWS Service Quotas (sebelumnya disebut batas layanan)
- AWS Trusted Advisor Pemeriksaan Praktik Terbaik (lihat bagian Batas Layanan)
- Pemantau batas AWS di jawaban AWS
- Batas Layanan Amazon EC
- Apa itu Service Quotas?
- Cara Meminta Peningkatan Kuota
- Titik akhir dan kuota layanan (service quotas
- Panduan Pengguna Service Quotas
- Pemantau Kuota untuk AWS
- AWS Batas Isolasi Kesalahan
- Ketersediaan dengan redundansi
- AWS untuk Data
- Apa itu Integrasi Berkelanjutan?
- Apa itu Pengiriman Berkelanjutan?
- Partner APN: partner yang dapat membantu Anda mengatasi manajemen konfigurasi
- Mengelola siklus hidup akun di lingkungan SaaS akun per penyewa di AWS
- Mengelola dan memantau throttling API di beban kerja Anda
- Lihat rekomendasi AWS Trusted Advisor dalam skala besar dengan AWS Organizations
- Melakukan Otomatisasi Peningkatan Batas Layanan dan Dukungan Perusahaan dengan AWS Control Tower

Video terkait:

- AWS Live re:Inforce 2019 - Service Quotas
- Melihat dan Mengelola Kuota untuk Layanan AWS Menggunakan Service Quotas
- Demo Kuota AWS IAM

Alat terkait:

- Amazon CodeGuru Reviewer

- AWS CodeDeploy
- AWS CloudTrail
- Amazon CloudWatch
- Amazon EventBridge
- Amazon DevOps Guru
- AWS Config
- AWS Trusted Advisor
- AWS CDK
- AWS Systems Manager
- AWS Marketplace

REL01-BP02 Mengelola kuota layanan (service quotas) di seluruh akun dan wilayah

Jika Anda menggunakan beberapa akun atau Wilayah, minta kuota yang sesuai di semua lingkungan tempat beban kerja produksi Anda dijalankan.

Hasil yang diinginkan: Layanan dan aplikasi tidak boleh terpengaruh oleh kehabisan kuota layanan (service quotas) untuk konfigurasi yang menjangkau akun-akun atau Wilayah atau yang memiliki desain ketahanan dengan menggunakan zona, Wilayah, atau failover akun.

Anti-pola umum:

- Membiarkan penggunaan sumber daya di satu Wilayah terisolasi untuk terus bertambah, tanpa mekanisme untuk mempertahankan kapasitas di wilayah-wilayah lainnya.
- Mengatur semua kuota di Wilayah isolasi secara manual dan independen.
- Tidak mempertimbangkan efek arsitektur ketahanan (seperti aktif atau pasif) dalam kebutuhan kuota masa depan saat terjadi penurunan kualitas di dalam Wilayah non-primer.
- Tidak melakukan evaluasi kuota secara rutin dan membuat perubahan yang diperlukan di setiap Wilayah dan akun tempat beban kerja dijalankan.
- Tidak memanfaatkan templat permintaan kuota untuk meminta peningkatan di beberapa Wilayah dan akun.
- Tidak memperbarui kuota layanan (service quotas) disebabkan pemikiran yang tidak tepat bahwa peningkatan kuota memiliki dampak biaya seperti permintaan pemesanan komputasi.

Manfaat menerapkan praktik terbaik ini: Memverifikasi bahwa Anda dapat menangani beban Anda saat ini di wilayah sekunder atau akun jika layanan regional tidak tersedia. Hal ini dapat membantu Anda mengurangi jumlah kesalahan atau tingkat penurunan kualitas yang terjadi selama kerugian wilayah (region loss).

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Tinggi

Panduan implementasi

Kuota layanan (service quotas) dilacak per akun. Setiap kuota disesuaikan dengan Wilayah AWS, kecuali ditetapkan sebaliknya. Selain lingkungan produksi, kelola juga kuota yang ada di semua lingkungan non-produksi yang berlaku agar pengujian dan pengembangan tidak terhambat. Untuk mempertahankan tingkat ketahanan yang tinggi diperlukan penilaian kuota layanan (service quotas) secara kontinu (baik otomatis maupun manual).

Dengan lebih banyak beban kerja yang mencakup Wilayah karena penerapan desain dengan menggunakan pendekatan Active/Active, Active/Passive – Hot, Active/Passive – Cold, dan Active/Passive – Pilot Light, penting untuk memahami semua tingkat kuota Wilayah dan akun. Pola lalu lintas terdahulu tidak selalu menjadi indikator yang tepat bahwa kuota layanan (service quotas) diatur dengan benar.

Sama pentingnya, batas nama kuota layanan (service quotas) tidak selalu sama untuk setiap Wilayah. Di satu Wilayah, nilainya bisa jadi lima, sedangkan di wilayah lain nilainya bisa jadi sepuluh. Pengelolaan atas kuota-kuota ini harus meliputi semua layanan, akun, dan Wilayah yang sama untuk menyediakan ketahanan yang konsisten saat beban meningkat.

Sesuaikan semua perbedaan kuota layanan (service quotas) di semua Wilayah yang berbeda (Wilayah Aktif atau Wilayah Pasif) dan buat proses untuk menyesuaikan semua perbedaan tersebut secara kontinu. Rencana pengujian failover Wilayah pasif sangat jarang diskalakan ke kapasitas aktif puncak, sehingga kegiatan game day atau table top bisa gagal menemukan perbedaan kuota layanan (service quotas) antar-Wilayah dan juga mempertahankan batas-batas yang tepat.

Penyimpangan kuota layanan (service quotas), kondisi di mana batas kuota layanan (service quotas) untuk kuota bernama tertentu diubah dalam satu Wilayah dan tidak semua Wilayah, sangat penting untuk dilacak dan dilakukan evaluasi. Anda harus mempertimbangkan untuk mengubah kuota di Wilayah yang memiliki lalu lintas atau yang berpotensi mendatangkan lalu lintas.

- Pilih Wilayah dan akun yang relevan berdasarkan persyaratan layanan, latensi, peraturan, dan pemulihan bencana (DR) Anda.

- Identifikasikan kuota layanan (service quotas) di semua akun, Wilayah, dan Zona Ketersediaan yang relevan. Batasannya mencakup akun dan Wilayah. Nilai-nilai ini harus dibandingkan untuk mengetahui perbedaannya.

Langkah-langkah implementasi

- Lakukan peninjauan atas nilai Service Quotas yang mungkin telah melampaui level risiko penggunaan a. AWS Trusted Advisor menyediakan peringatan untuk pelanggaran 80% dan 90% ambang batas.
- Tinjau nilai untuk kuota layanan (service quotas) di Wilayah Pasif mana pun (dalam desain Aktif/Pasif). Verifikasi bahwa muatan akan berhasil dijalankan di dalam Wilayah sekunder apabila terjadi kegagalan di dalam Wilayah primer.
- Otomatiskan penilaian jika penyelewengan kuota layanan (service quotas) apa pun telah terjadi antar-Wilayah di dalam akun yang sama dan lakukan tindakan yang semestinya untuk mengubah batas.
- Jika Unit Organisasi (OU) pelanggan disusun dengan cara yang didukung, templat kuota layanan (service quotas) harus diperbarui agar mencerminkan perubahan pada kuota apa pun yang harus diterapkan ke beberapa Wilayah dan akun.
  - Buat templat dan kaitkan Wilayah dengan perubahan kuota.
  - Tinjau semua templat kuota layanan (service quotas) yang ada untuk mengetahui jika ada perubahan yang diperlukan (Wilayah, batas, dan akun).

Sumber daya

Praktik-praktik terbaik terkait:

- REL01-BP01 Kesadaran tentang kuota (service quotas) dan kendala layanan
- REL01-BP03 Mengakomodasi kuota layanan (service quotas) tetap dan kendala melalui arsitektur
- REL01-BP04 Memantau dan mengelola kuota
- REL01-BP05 Mengotomatiskan manajemen kuota
- REL01-BP06 Memastikan adanya selisih yang memadai antara kuota saat ini dan penggunaan maksimum untuk mengakomodasi failover
- REL03-BP01 Pilih cara mengelompokkan beban kerja Anda
- REL10-BP01 Melakukan deployment beban kerja ke beberapa lokasi

- REL11-BP01 Memantau semua komponen beban kerja untuk mendeteksi kegagalan
- REL11-BP03 Melakukan otomatisasi pemulihan di semua lapisan
- REL12-BP04 Menguji ketahanan menggunakan chaos engineering

Dokumen terkait:

- Pilar Keandalan Kerangka Kerja AWS Well-Architected: Ketersediaan
- AWS Service Quotas (sebelumnya disebut batas layanan)
- Pemeriksaan Praktik Terbaik AWS Trusted Advisor (lihat bagian Batas Layanan)
- Pemantau batas AWS di jawaban AWS
- Batas Layanan Amazon EC
- Apa itu Service Quotas?
- Cara Meminta Peningkatan Kuota
- Titik akhir dan kuota layanan (service quotas
- Panduan Pengguna Service Quotas
- Pemantau Kuota untuk AWS
- Batas Isolasi Kesalahan AWS
- Ketersediaan dengan redundansi
- AWS untuk Data
- Apa itu Integrasi Berkelanjutan?
- Apa itu Pengiriman Berkelanjutan?
- Partner APN: partner yang dapat membantu Anda mengatasi manajemen konfigurasi
- Mengelola siklus hidup akun di lingkungan SaaS akun per penyewa di AWS
- Mengelola dan memantau throttling API di beban kerja Anda
- Lihat rekomendasi AWS Trusted Advisor dalam skala besar dengan AWS Organizations
- Melakukan Otomatisasi Peningkatan Batas Layanan dan Dukungan Perusahaan dengan AWS Control Tower

Video terkait:

- AWS Live re:Inforce 2019 - Service Quotas
- Melihat dan Mengelola Kuota untuk Layanan AWS Menggunakan Service Quotas

- Demo Kuota AWS IAM

Layanan terkait:

- Amazon CodeGuru Reviewer
- AWS CodeDeploy
- AWS CloudTrail
- Amazon CloudWatch
- Amazon EventBridge
- Amazon DevOps Guru
- AWS Config
- AWS Trusted Advisor
- AWS CDK
- AWS Systems Manager
- AWS Marketplace

REL01-BP03 Mengakomodasi kuota layanan (service quotas) tetap dan kendala melalui arsitektur

Perhatikan kuota layanan (service quotas), batasan layanan, dan batas sumber daya fisik yang tidak dapat diubah. Rancang arsitektur untuk aplikasi dan layanan agar batas ini tidak memengaruhi keandalan.

Contohnya antara lain adalah bandwidth jaringan, ukuran payload invokasi fungsi nirserver, tingkat lonjakan throttling untuk gateway API, dan sambungan pengguna serentak ke basis data.

Hasil yang diinginkan: Aplikasi atau layanan berfungsi sebagaimana yang diharapkan, baik saat kondisi lalu lintas normal atau pun sedang tinggi. Aplikasi dan layanan telah dirancang untuk berfungsi dengan batasan untuk kuota layanan (service quotas) atau kendala tetap sumber daya tersebut.

Anti-pola umum:

- Memilih sebuah desain yang menggunakan sebuah sumber daya dari sebuah layanan, tidak menyadari bahwa terdapat kendala desain yang akan menyebabkan desain ini gagal begitu Anda menyesuaikan skala.

- Melakukan tolok ukur yang tidak realistis dan akan mencapai kuota tetap layanan selama pengujian. Contohnya, menjalankan pengujian pada batas lonjakan tetapi dalam jangka waktu yang lama.
- Memilih desain yang tidak dapat diskalakan atau dimodifikasi jika kuota layanan (service quotas) tetap akan terlampaui. Contohnya, ukuran payload SQS sebesar 256 KB.
- Observabilitas belum dirancang dan diimplementasikan untuk memantau dan memberikan peringatan tentang ambang batas kuota layanan (service quotas) yang mungkin berisiko selama lalu lintas sedang tinggi

Manfaat menerapkan praktik terbaik ini: Memverifikasi bahwa aplikasi akan berjalan berdasarkan semua tingkat beban layanan yang diproyeksikan tanpa terjadi gangguan atau degradasi.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Sedang

Panduan implementasi

Tidak seperti sumber daya atau kuota layanan relatif yang diganti dengan unit kapasitas lebih tinggi, kuota tetap layanan AWS tidak dapat diubah. Ini berarti semua jenis layanan-layanan AWS ini harus dievaluasi untuk mengetahui potensi batas kapasitas keras ketika digunakan dalam sebuah desain aplikasi.

Batas absolut ditunjukkan di konsol Service Quotas. Jika kolom menampilkan `ADJUSTABLE = No`, layanan memiliki batas keras. Batas keras juga ditunjukkan di beberapa halaman konfigurasi sumber daya. Contohnya, Lambda memiliki batas keras spesifik yang tidak dapat disesuaikan.

Sebagai contoh, ketika merancang desain dari sebuah aplikasi python untuk beroperasi dalam fungsi Lambda, aplikasi tersebut harus dievaluasi untuk menentukan apakah ada peluang bagi Lambda untuk beroperasi lebih lama dari 15 menit. Jika kode mungkin akan dijalankan lebih dari batas kuota layanan (service quotas) ini, desain atau teknologi lain harus dipertimbangkan. Jika batas ini tercapai setelah deployment produksi, maka aplikasi akan mengalami penurunan kualitas dan gangguan sampai aplikasi itu dapat diperbaiki. Tidak seperti kuota relatif, tidak ada metode untuk mengubah batas-batas ini meskipun dalam kondisi darurat Keparahan 1.

Setelah aplikasi telah di-deploy ke lingkungan pengujian, Anda harus berstrategi untuk menemukan apakah batas keras dapat tercapai. Pengujian stres, pengujian beban, dan pengujian kekacauan harus menjadi bagian dari rencana pengujian pendahuluan.

Langkah-langkah implementasi

- Tinjau daftar lengkap layanan AWS yang dapat digunakan dalam fase desain aplikasi.
- Tinjau batas kuota lunak dan batas kuota keras untuk semua layanan ini. Tidak semua batas ditunjukkan di konsol Service Quotas. Beberapa layanan menggambarkan batasan ini di lokasi yang lain.
- Saat Anda mendesain aplikasi Anda, lakukan peninjauan terhadap pendorong teknologi dan bisnis beban kerja Anda, seperti hasil bisnis, kasus penggunaan, sistem yang dependen, target ketersediaan, dan objek pemulihan bencana. Biarkan pendorong teknologi dan bisnis Anda memandu proses untuk mengidentifikasi sistem terdistribusi yang tepat untuk beban kerja Anda.
- Analisis beban layanan di berbagai Wilayah dan akun. Banyak batas keras yang berbasis wilayah untuk layanan. Tetapi, beberapa batas berbasis akun.
- Analisis arsitektur ketangguhan untuk penggunaan sumber daya selama terjadi kegagalan zona dan kegagalan Wilayah. Jika kemajuan desain multi-Wilayah menggunakan pendekatan aktif/aktif, aktif/pasif – panas, aktif/pasif - dingin, dan aktif/pasif - pilot light, kasus-kasus kegagalan ini akan menyebabkan penggunaan yang lebih tinggi. Hal ini akan menimbulkan potensi kasus penggunaan untuk mencapai batas keras.

Sumber daya

Praktik-praktik terbaik terkait:

- REL01-BP01 Kesadaran tentang kuota (service quotas) dan kendala layanan
- REL01-BP02 Mengelola kuota layanan (service quotas) di seluruh akun dan wilayah
- REL01-BP04 Memantau dan mengelola kuota
- REL01-BP05 Mengotomatiskan manajemen kuota
- REL01-BP06 Memastikan adanya selisih yang memadai antara kuota saat ini dan penggunaan maksimum untuk mengakomodasi failover
- REL03-BP01 Pilih cara mengelompokkan beban kerja Anda
- REL10-BP01 Melakukan deployment beban kerja ke beberapa lokasi
- REL11-BP01 Memantau semua komponen beban kerja untuk mendeteksi kegagalan
- REL11-BP03 Melakukan otomatisasi pemulihan di semua lapisan
- REL12-BP04 Menguji ketahanan menggunakan chaos engineering

Dokumen terkait:

- AWS Pilar Keandalan Kerangka Kerja Well-Architected: Ketersediaan
- AWS Service Quotas (sebelumnya disebut batas layanan)
- AWS Trusted Advisor Pemeriksaan Praktik Terbaik (lihat bagian Batas Layanan)
- Pemantau batas AWS di jawaban AWS
- Batas Layanan Amazon EC
- Apa itu Service Quotas?
- Cara Meminta Peningkatan Kuota
- Titik akhir dan kuota layanan (service quotas
- Panduan Pengguna Service Quotas
- Pemantau Kuota untuk AWS
- Batas Isolasi Kesalahan AWS
- Ketersediaan dengan redundansi
- AWS untuk Data
- Apa itu Integrasi Berkelanjutan?
- Apa itu Pengiriman Berkelanjutan?
- Partner APN: partner yang dapat membantu Anda mengatasi manajemen konfigurasi
- Mengelola siklus hidup akun di lingkungan SaaS akun per penyewa di AWS
- Mengelola dan memantau throttling API di beban kerja Anda
- Lihat rekomendasi AWS Trusted Advisor dalam skala besar dengan AWS Organizations
- Melakukan Otomatisasi Peningkatan Batas Layanan dan Dukungan Perusahaan dengan AWS Control Tower
- Tindakan, sumber daya, dan kunci syarat untuk layanan Service Quotas

Video terkait:

- AWS Live re:Inforce 2019 - Service Quotas
- Melihat dan Mengelola Kuota untuk Layanan AWS Menggunakan Service Quotas
- AWS Demo Kuota IAM
- AWS re:Invent 2018: Loop Tertutup dan Pikiran Terbuka: Cara Mengendalikan Sistem, Besar dan Kecil

Alat terkait:

- AWS CodeDeploy
- AWS CloudTrail
- Amazon CloudWatch
- Amazon EventBridge
- Amazon DevOps Guru
- AWS Config
- AWS Trusted Advisor
- AWS CDK
- AWS Systems Manager
- AWS Marketplace

REL01-BP04 Memantau dan mengelola kuota

Evaluasi potensi penggunaan Anda dan tingkatkan kuota dengan semestinya, sehingga terjadi pertumbuhan penggunaan sesuai rencana.

Hasil yang diinginkan: Sistem aktif dan otomatis yang mengelola dan memantau telah di-deploy. Solusi-solusi operasi ini memastikan ambang batas penggunaan kuota hampir dicapai. Hal ini akan secara proaktif diperbaiki berdasarkan perubahan kuota yang diminta.

Anti-pola umum:

- Tidak mengonfigurasi pemantauan untuk memeriksa ambang batas kuota layanan (service quotas)
- Tidak mengonfigurasi pemantauan untuk batas keras, meskipun nilai-nilai tersebut tidak dapat diubah.
- Berasumsi bahwa jumlah waktu yang diperlukan untuk meminta dan mendapatkan perubahan kuota lunak adalah segera atau singkat.
- Mengonfigurasi alarm untuk ketika kuota layanan (service quotas) sudah dekat, namun tidak memiliki proses untuk merespons pemberitahuan.
- Hanya mengonfigurasi alarm untuk layanan yang didukung oleh Service AWS Quotas dan tidak memantau layanan lain. AWS
- Tidak mempertimbangkan manajemen kuota untuk beberapa desain ketangguhan Wilayah, seperti pendekatan aktif/aktif, aktif/pasif - panas, aktif/pasif - dingin, dan aktif/pasif - pilot light.
- Tidak menilai perbedaan kuota antara Wilayah.

- Tidak menilai kebutuhan di setiap Wilayah untuk permintaan peningkatan kuota spesifik.
- Tidak memanfaatkan templat untuk pengelolaan kuota multi Wilayah.

Manfaat menetapkan praktik terbaik ini: Pelacakan otomatis terhadap AWS Service Quotas dan pemantauan penggunaan Anda terhadap kuota tersebut akan memungkinkan Anda untuk melihat kapan Anda mendekati batas kuota. Anda juga dapat menggunakan data pemantauan ini untuk membantu Anda membatasi penurunan kualitas karena kehabisan kuota.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Sedang

Panduan implementasi

Untuk layanan-layanan yang didukung, Anda dapat memantau kuota Anda dengan mengonfigurasikan berbagai layanan yang berbeda yang dapat menilai dan kemudian mengirimkan pemberitahuan atau alarm. Hal ini dapat membantu Anda memantau penggunaan dan dapat memberitahukan kuota yang akan segera habis kepada Anda. Alarm ini dapat dipanggil dari, fungsi AWS Config Lambda, Amazon CloudWatch, atau dari. AWS Trusted Advisor Anda juga dapat menggunakan filter metrik pada CloudWatch Log untuk mencari dan mengekstrak pola di log untuk menentukan apakah penggunaan mendekati ambang kuota.

Langkah-langkah implementasi

Untuk pemantauan:

- Catat pemakaian sumber daya saat ini (misalnya, bucket atau instans). Gunakan API operasi layanan, seperti Amazon EC2 `DescribeInstances`API, untuk mengumpulkan konsumsi sumber daya saat ini.
- Catat kuota saat ini yang penting dan berlaku untuk layanan dengan menggunakan:
  - AWS Service Quotas
  - AWS Trusted Advisor
  - AWS dokumentasi
  - AWS halaman khusus layanan
  - AWS Command Line Interface (AWS CLI)
  - AWS Cloud Development Kit (AWS CDK)
- Gunakan AWS Service Quotas, AWS layanan yang membantu Anda mengelola kuota untuk lebih dari 250 AWS layanan dari satu lokasi.

- Gunakan batas Trusted Advisor layanan untuk memantau batas layanan Anda saat ini di berbagai ambang batas.
- Gunakan riwayat kuota layanan (konsol atau AWS CLI) untuk memeriksa kenaikan regional.
- Bandingkan perubahan kuota layanan (service quotas) di setiap Wilayah dan setiap akun untuk membuat kesetaraan, jika diperlukan.

Untuk manajemen:

- Otomatis: Siapkan aturan AWS Config khusus untuk memindai kuota layanan di seluruh Wilayah dan membandingkan perbedaannya.
- Otomatis: Siapkan fungsi Lambda terjadwal untuk memindai kuota layanan (service quotas) di berbagai Wilayah dan bandingkan untuk mengetahui perbedaannya.
- Petunjuk: Pindai kuota layanan melalui AWS CLI,API, atau AWS Konsol untuk memindai kuota layanan di seluruh Wilayah dan membandingkan perbedaannya. Laporkan perbedaannya.
- Jika perbedaan kuota diidentifikasi antara Wilayah, mintalah perubahan kuota, jika perlu.
- Tinjau hasil dari semua permintaan.

Sumber daya

Praktik-praktik terbaik terkait:

- REL01-BP01 Kesadaran tentang kuota (service quotas) dan kendala layanan
- REL01-BP02 Mengelola kuota layanan (service quotas) di seluruh akun dan wilayah
- REL01-BP03 Mengakomodasi kuota layanan (service quotas) tetap dan kendala melalui arsitektur
- REL01-BP05 Mengotomatiskan manajemen kuota
- REL01-BP06 Memastikan adanya selisih yang memadai antara kuota saat ini dan penggunaan maksimum untuk mengakomodasi failover
- REL03-BP01 Pilih cara mengelompokkan beban kerja Anda
- REL10-BP01 Melakukan deployment beban kerja ke beberapa lokasi
- REL11-BP01 Memantau semua komponen beban kerja untuk mendeteksi kegagalan
- REL11-BP03 Melakukan otomatisasi pemulihan di semua lapisan
- REL12-BP04 Menguji ketahanan menggunakan chaos engineering

Dokumen terkait:

- AWS Pilar Keandalan Well-Architected Framework: Ketersediaan
- AWS Service Quotas (sebelumnya disebut sebagai batas layanan)
- AWS Trusted Advisor Pemeriksaan Praktik Terbaik (lihat bagian Batas Layanan)
- AWS batasi monitor pada AWS jawaban
- Batas EC2 Layanan Amazon
- Apa itu Service Quotas?
- Cara Meminta Peningkatan Kuota
- Titik akhir dan kuota layanan (service quotas)
- Panduan Pengguna Service Quotas
- Monitor Kuota untuk AWS
- AWS Batas Isolasi Kesalahan
- Ketersediaan dengan redundansi
- AWS untuk Data
- Apa itu Integrasi Berkelanjutan?
- Apa itu Pengiriman Berkelanjutan?
- APNMitra: mitra yang dapat membantu manajemen konfigurasi
- Mengelola siklus hidup akun di lingkungan SaaS account-per-tenant AWS
- Mengelola dan memantau API pembatasan dalam beban kerja Anda
- Lihat AWS Trusted Advisor rekomendasi dalam skala besar dengan AWS Organizations
- Mengotomatisasi Peningkatan Batas Layanan dan Dukungan Perusahaan dengan AWS Control Tower
- Tindakan, sumber daya, dan kunci syarat untuk layanan Service Quotas

Video terkait:

- AWS Live re: Inforce 2019 - Service Quotas
- Melihat dan Mengelola Kuota untuk AWS Layanan Menggunakan Service Quotas
- AWS IAMKuota Demo

- AWS RE: Invent 2018: Tutup Loop dan Membuka Pikiran: Cara Mengendalikan Sistem, Besar dan Kecil

Alat terkait:

- AWS CodeDeploy
- AWS CloudTrail
- Amazon CloudWatch
- Amazon EventBridge
- DevOpsGuru Amazon
- AWS Config
- AWS Trusted Advisor
- AWS CDK
- AWS Systems Manager
- AWS Marketplace

REL01-BP05 Mengotomatiskan manajemen kuota

Kuota layanan, juga disebut sebagai batasan dalam layanan AWS, adalah jumlah maksimum untuk sumber daya dalam Akun AWS Anda. Setiap layanan AWS menetapkan serangkaian kuota dan nilai default-nya. Untuk memberikan beban kerja Anda akses ke semua sumber daya yang dibutuhkannya, Anda mungkin perlu meningkatkan nilai kuota layanan Anda.

Pertumbuhan konsumsi sumber daya AWS oleh beban kerja dapat mengancam stabilitas beban kerja dan memengaruhi pengalaman pengguna jika kuota terlampaui. Terapkan alat untuk memperingatkan Anda ketika beban kerja Anda mendekati batas dan pertimbangkan untuk membuat permintaan peningkatan kuota secara otomatis.

Hasil yang diinginkan: Kuota dikonfigurasi dengan tepat untuk beban kerja yang berjalan di setiap Akun AWS dan Wilayah.

Anti-pola umum:

- Anda gagal mempertimbangkan dan menyesuaikan kuota dengan tepat untuk memenuhi persyaratan beban kerja.

- Anda melacak kuota dan penggunaan menggunakan metode yang bisa menjadi usang, seperti dengan spreadsheet.
- Anda hanya memperbarui batas layanan berdasarkan jadwal berkala.
- Organisasi Anda tidak memiliki proses operasional untuk meninjau kuota yang ada dan meminta peningkatan kuota layanan ketika diperlukan.

Manfaat menjalankan praktik terbaik ini:

- Peningkatan ketahanan beban kerja: Anda mencegah kesalahan yang disebabkan oleh terlampauinya kuota sumber daya AWS.
- Pemulihan bencana yang disederhanakan: Anda dapat menggunakan kembali mekanisme manajemen kuota otomatis yang dibangun di Wilayah primer selama pengaturan DR di Wilayah AWS lain.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Sedang

Panduan implementasi

Lihat kuota saat ini dan lacak konsumsi kuota yang berjalan melalui berbagai mekanisme seperti konsol AWS Service Quotas, AWS Command Line Interface (AWS CLI), dan SDK AWS. Anda juga dapat mengintegrasikan basis data manajemen konfigurasi (CMDB) dan sistem manajemen layanan IT (ITSM) Anda dengan API AWS Service Quotas.

Buat peringatan otomatis jika penggunaan kuota mencapai ambang batas yang Anda tentukan, dan tetapkan proses untuk mengirimkan permintaan peningkatan kuota ketika Anda menerima peringatan. Jika beban kerja yang mendasarinya sangat penting bagi bisnis Anda, Anda dapat mengotomatiskan permintaan peningkatan kuota, tetapi uji otomatisasi dengan cermat untuk menghindari risiko tindakan yang tidak terkendali seperti "growth feedback loop".

Peningkatan kuota yang lebih kecil sering kali disetujui secara otomatis. Permintaan kuota yang lebih besar mungkin perlu diproses secara manual oleh tim dukungan AWS dan dapat memakan waktu lebih lama untuk ditinjau dan diproses. Berikan waktu tambahan untuk memproses beberapa permintaan atau permintaan peningkatan yang besar.

Langkah-langkah implementasi

- Terapkan pemantauan otomatis terhadap kuota layanan, dan keluarkan peringatan jika pertumbuhan pemanfaatan sumber daya beban kerja Anda mendekati batas kuota Anda. Misalnya,

Pemantau Kuota untuk AWS dapat memberikan pemantauan otomatis terhadap kuota layanan. Alat ini terintegrasi dengan AWS Organizations dan di-deploy menggunakan Cloudformation StackSets sehingga akun baru secara otomatis dipantau saat dibuat.

- Gunakan fitur seperti templat permintaan Service Quotas atau AWS Control Tower untuk menyederhanakan penyiapan Service Quotas untuk akun baru.
- Buat dasbor penggunaan kuota layanan Anda saat ini di semua Akun AWS dan wilayah dan lihat dasbor tersebut jika diperlukan untuk mencegah terlampauinya kuota Anda. Trusted Advisor Organizational (TAO) Dashboard, bagian dari Cloud Intelligence Dashboard, dapat membantu Anda mulai membangun dasbor semacam itu dengan cepat.
- Lacak permintaan peningkatan batas layanan. Wawasan Terkonsolidasi dari Beberapa Akun (CIMA) dapat menyediakan tampilan tingkat Organisasi dari semua permintaan Anda.
- Uji pembuatan peringatan dan otomatisasi permintaan peningkatan kuota apa pun dengan menetapkan ambang batas kuota yang lebih rendah di akun non-produksi. Jangan melakukan pengujian ini di akun produksi.

Sumber daya

Praktik-praktik terbaik terkait:

- OPS10-BP07 Melakukan otomatisasi respons terhadap peristiwa

Dokumen terkait:

- Partner APN: partner yang dapat membantu Anda mengatasi manajemen konfigurasi
- AWS Marketplace: Produk CMDB yang membantu melacak batasan
- AWS Service Quotas (sebelumnya disebut batas layanan)
- Pemeriksaan Praktik Terbaik AWS Trusted Advisor (lihat bagian Batas Layanan)
- Solusi Pemantau Kuota di AWS - Solusi AWS
- Apa itu Service Quotas?
- Apa itu templat permintaan Service Quotas?

Video terkait:

- AWS Live re:Inforce 2019 - Service Quotas

- Melakukan Otomatisasi Peningkatan Batas Layanan dan Dukungan Korporasi dengan AWS Control Tower

Alat terkait:

- Pemantau Kuota untuk AWS

REL01-BP06 Memastikan adanya selisih yang memadai antara kuota saat ini dan penggunaan maksimum untuk mengakomodasi failover

Artikel ini akan menjelaskan kepada Anda tentang cara menjaga ruang antara kuota sumber daya dan penggunaan Anda, dan bagaimana hal itu dapat memberikan manfaat bagi organisasi Anda. Setelah Anda selesai menggunakan sebuah sumber daya, kuota penggunaan tersebut dapat terus memperhitungkan sumber daya itu. Hal ini dapat mengakibatkan terjadinya sumber daya yang gagal atau tidak dapat diakses. Hindari kegagalan sumber daya dengan memastikan apakah kuota meliputi sumber daya yang tumpang tindih dan tidak dapat diakses serta penggantinya. Pertimbangkan kasus penggunaan seperti kegagalan jaringan, kegagalan Zona Ketersediaan, atau kegagalan Wilayah ketika menghitung selisih ini.

Hasil yang diinginkan: Kegagalan kecil atau besar dalam sumber daya atau aksesibilitas sumber daya dapat tercakup dalam ambang batas layanan yang berlaku saat ini. Kegagalan zona, kegagalan jaringan, atau bahkan kegagalan Wilayah telah dipertimbangkan dalam pembuatan rencana sumber daya.

Anti-pola umum:

- Mengatur kuota layanan (service quotas) berdasarkan kebutuhan saat ini tanpa memperhitungkan skenario failover.
- Tidak mempertimbangkan pengguna utama stabilitas statis ketika menghitung kuota puncak untuk sebuah layanan.
- Tidak mempertimbangkan potensi kemungkinan sumber daya yang tidak dapat diakses dalam menghitung kuota total yang diperlukan untuk masing-masing Wilayah.
- Tidak mempertimbangkan batas isolasi kesalahan layanan AWS untuk beberapa layanan dan potensi kemungkinan pola penggunaannya yang abnormal.

Manfaat menerapkan praktik terbaik ini: Ketika terjadi peristiwa gangguan layanan yang memengaruhi ketersediaan aplikasi, gunakan cloud untuk menerapkan strategi-strategi untuk pulih dari peristiwa ini. Contoh strategi adalah membuat sumber daya tambahan untuk menggantikan sumber daya yang tidak dapat diakses untuk mengakomodasi kondisi failover tanpa menghabiskan batas layanan Anda.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Sedang

Panduan implementasi

Ketika melakukan evaluasi batas kuota, Anda juga harus mempertimbangkan kasus-kasus failover yang dapat terjadi karena adanya degradasi. Pertimbangkan kasus-kasus failover berikut ini.

- VPC yang mengalami gangguan atau tidak dapat diakses.
- Subnet yang tidak dapat diakses.
- Zona Ketersediaan yang mengalami degradasi dan memengaruhi aksesibilitas sumber daya.
- Berbagai rute titik keluar dan masuk atau rute jaringan yang diblokir atau berubah.
- Wilayah terdegradasi yang berdampak pada aksesibilitas sumber daya.
- Subset sumber daya yang dipengaruhi oleh kegagalan di Wilayah atau Zona Ketersediaan.

Keputusan untuk failover adalah keputusan yang bersifat unik untuk setiap situasi dan pelanggan, karena dampak bisnisnya bisa sangat berbeda. Atasi perencanaan kapasitas sumber daya di lokasi failover dan kuota sumber daya sebelum Anda memutuskan untuk melakukan failover aplikasi atau layanan.

Anda juga harus mempertimbangkan puncak aktivitas yang lebih tinggi dari normal saat meninjau kuota untuk setiap layanan. Puncak ini mungkin terkait dengan sumber daya yang tidak dapat diakses karena jaringan atau izin, tetapi masih aktif. Sumber daya aktif yang tidak dihentikan dihitung untuk memenuhi batas kuota layanan (service quotas).

Langkah-langkah implementasi

- Pertahankan ruang antara kuota layanan (service quotas) saat ini dan penggunaan maksimum untuk mengakomodasi failover atau hilangnya kemampuan akses.
- Tentukan service quotas Anda. Memperhitungkan pola deployment yang khas, persyaratan ketersediaan, dan pertumbuhan konsumsi.
- Minta peningkatan kuota, jika perlu. Antisipasi waktu tunggu untuk permintaan kenaikan kuota.

- Tentukan persyaratan keandalan (juga disebut sebagai jumlah angka sembilan Anda).
- Memahami skenario kesalahan potensial seperti hilangnya komponen, Zona Ketersediaan, atau Wilayah.
- Tetapkan metodologi deployment Anda (misalnya canary, blue/green, red/black, atau rolling).
- Sertakan buffer yang sesuai untuk batas saat ini. Contoh buffer bisa 15%.
- Sertakan perhitungan untuk stabilitas statis (Zona dan Wilayah), apabila sesuai.
- Rencanakan peningkatan pemakaian dan pantau tren pemakaian Anda.
- Pertimbangkan dampak stabilitas statis untuk beban kerja Anda yang paling penting. Lakukan penilaian terhadap sumber daya yang sesuai dengan sebuah sistem stabil secara statis di semua Wilayah dan Zona Ketersediaan.
- Pertimbangkan untuk menggunakan Reservasi Kapasitas Sesuai Permintaan untuk menjadwalkan kapasitas sebelum failover. Hal ini bisa menjadi strategi yang bermanfaat untuk mengimplementasikan penjadwalan bisnis yang paling penting guna mengurangi potensi risiko mendapatkan kuantitas dan jenis sumber daya yang benar selama failover.

Sumber daya

Praktik-praktik terbaik terkait:

- REL01-BP01 Kesadaran tentang kuota (service quotas) dan kendala layanan
- REL01-BP02 Mengelola kuota layanan (service quotas) di seluruh akun dan wilayah
- REL01-BP03 Mengakomodasi kuota layanan (service quotas) tetap dan kendala melalui arsitektur
- REL01-BP04 Memantau dan mengelola kuota
- REL01-BP05 Mengotomatiskan manajemen kuota
- REL03-BP01 Pilih cara mengelompokkan beban kerja Anda
- REL10-BP01 Melakukan deployment beban kerja ke beberapa lokasi
- REL11-BP01 Memantau semua komponen beban kerja untuk mendeteksi kegagalan
- REL11-BP03 Melakukan otomatisasi pemulihan di semua lapisan
- REL12-BP04 Menguji ketahanan menggunakan chaos engineering

Dokumen terkait:

- Pilar Keandalan Kerangka Kerja AWS Well-Architected: Ketersediaan

- AWS Service Quotas (sebelumnya disebut batas layanan)
- Pemeriksaan Praktik Terbaik AWS Trusted Advisor (lihat bagian Batas Layanan)
- Pemantau batas AWS di jawaban AWS
- Batas Layanan Amazon EC2
- Apa itu Service Quotas?
- Cara Meminta Peningkatan Kuota
- Titik akhir dan kuota layanan (service quotas)
- Panduan Pengguna Service Quotas
- Pemantau Kuota untuk AWS
- AWS Batas Isolasi Kesalahan
- Ketersediaan dengan redundansi
- AWS untuk Data
- Apa itu Integrasi Berkelanjutan?
- Apa itu Pengiriman Berkelanjutan?
- Partner APN: partner yang dapat membantu Anda mengatasi manajemen konfigurasi
- Mengelola siklus hidup akun di lingkungan SaaS akun per penyewa di AWS
- Mengelola dan memantau throttling API di beban kerja Anda
- Lihat rekomendasi AWS Trusted Advisor dalam skala besar dengan AWS Organizations
- Melakukan Otomatisasi Peningkatan Batas Layanan dan Dukungan Perusahaan dengan AWS Control Tower
- Tindakan, sumber daya, dan kunci syarat untuk layanan Service Quotas

Video terkait:

- AWS Live re:Inforce 2019 - Service Quotas
- Melihat dan Mengelola Kuota untuk Layanan AWS Menggunakan Service Quotas
- Demo Kuota AWS IAM
- AWS re:Invent 2018: Loop Tertutup dan Pikiran Terbuka: Cara Mengendalikan Sistem, Besar dan Kecil

Alat terkait:

- AWS CodeDeploy
- AWS CloudTrail
- Amazon CloudWatch
- Amazon EventBridge
- Amazon DevOps Guru
- AWS Config
- AWS Trusted Advisor
- CDK AWS
- AWS Systems Manager
- AWS Marketplace

Merencanakan topologi jaringan Anda

Sering kali beban kerja ada di beberapa lingkungan. Ini termasuk beberapa lingkungan cloud (baik yang dapat diakses publik maupun privat) dan kemungkinan infrastruktur pusat data Anda yang ada sekarang. Rencana harus mencakup pertimbangan jaringan, seperti konektivitas dalam sistem dan antar sistem, pengelolaan alamat IP publik, pengelolaan alamat IP privat, dan resolusi nama domain.

Saat merancang arsitektur sistem menggunakan jaringan berbasis alamat IP, Anda harus merencanakan topologi dan penanganan jaringan untuk mengantisipasi adanya kemungkinan kegagalan, dan untuk mengakomodasi pertumbuhan dan integrasi mendatang dengan sistem-sistem lain serta jaringannya.

Amazon Virtual Private Cloud (Amazon VPC) akan memungkinakn Anda menyediakan bagian AWS Cloud privat yang terisolasi di mana Anda dapat meluncurkan sumber daya AWS dalam sebuah jaringan virtual.

Praktik terbaik

- REL02-BP01 Gunakan konektivitas jaringan dengan ketersediaan tinggi untuk titik akhir publik beban kerja Anda
- REL02-BP02 Menyediakan konektivitas redundan antara jaringan privat di cloud dan lingkungan on-premise
- REL02-BP03 Pastikan alokasi subnet IP menjelaskan ekspansi dan ketersediaan
- REL02-BP04 Mengutamakan topologi hub-and-spoke daripada mesh many-to-many

- REL02-BP05 Terapkan rentang alamat IP privat yang tidak tumpang tindih di semua ruang alamat privat tempat semuanya saling terhubung

REL02-BP01 Gunakan konektivitas jaringan dengan ketersediaan tinggi untuk titik akhir publik beban kerja Anda

Membuat konektivitas jaringan dengan ketersediaan tinggi untuk titik akhir publik beban kerja Anda dapat membantu Anda mengurangi waktu henti karena hilangnya konektivitas dan meningkatkan ketersediaan serta SLA beban kerja Anda. Untuk mencapai ini, gunakan DNS dengan ketersediaan tinggi, jaringan pengiriman konten (CDN), gateway API, penyeimbangan beban, atau proksi mundur.

Hasil yang diinginkan: Sangat penting bagi Anda untuk merencanakan, membangun, dan mengoperasionalkan konektivitas jaringan dengan ketersediaan tinggi untuk titik akhir publik Anda. Jika beban kerja Anda menjadi tidak terjangkau karena hilangnya konektivitas, meskipun beban kerja Anda beroperasi dan tersedia, para pelanggan Anda akan menganggap sistem Anda tidak berfungsi (down). Dengan menggabungkan konektivitas jaringan yang tangguh dan memiliki ketersediaan yang tinggi untuk titik akhir publik beban kerja Anda, bersama dengan arsitektur yang tangguh untuk beban kerja itu sendiri, Anda dapat memberikan tingkat layanan dan ketersediaan sebaik mungkin untuk para pelanggan Anda.

AWS Global Accelerator, Amazon CloudFront, Amazon API Gateway, URL Fungsi AWS Lambda, API AWS AppSync, dan Elastic Load Balancing (ELB) semuanya menyediakan titik akhir publik dengan ketersediaan yang tinggi. Amazon Route 53 menyediakan layanan DNS dengan ketersediaan yang tinggi untuk resolusi nama domain guna memverifikasi bahwa alamat titik akhir publik Anda dapat diselesaikan.

Anda juga dapat mengevaluasi peralatan perangkat lunak AWS Marketplace untuk melakukan penyeimbangan beban dan proksi.

Anti-pola umum:

- Mendesain sebuah beban kerja dengan ketersediaan yang tinggi tanpa merencanakan konektivitas jaringan dan DNS untuk ketersediaan tinggi.
- Menggunakan alamat internet publik di masing-masing instans atau kontainer dan mengelola konektivitasnya dengan DNS.
- Menggunakan alamat IP, bukan nama domain untuk mencari layanan.
- Tidak menguji skenario di mana konektivitas ke titik akhir publik Anda hilang.

- Tidak menganalisis pola distribusi dan kebutuhan throughput jaringan.
- Tidak menguji dan merencanakan skenario di mana konektivitas jaringan internet ke titik akhir publik beban kerja Anda mungkin mengalami gangguan.
- Memberikan konten (seperti halaman web, aset statis, atau file media) ke area-area geografis besar dan tidak menggunakan CDN.
- Tidak membuat rencana untuk serangan penolakan layanan terdistribusi (DDoS). Serangan DDoS berisiko menghalangi lalu lintas sah dan menurunkan ketersediaan untuk para pengguna Anda.

Manfaat menjalankan praktik terbaik ini: Membuat desain untuk konektivitas jaringan dengan ketersediaan tinggi dan tangguh akan memastikan bahwa beban kerja Anda dapat diakses dan tersedia bagi para pengguna Anda.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Tinggi

Panduan implementasi

Inti dari pembangunan konektivitas jaringan dengan ketersediaan tinggi untuk titik akhir publik adalah pengarahan rute lalu lintas. Untuk memverifikasi bahwa lalu lintas Anda dapat menjangkau titik akhir, DNS harus dapat mengatur nama domain ke alamat IP-nya yang bersangkutan. Gunakan Sistem Nama Domain (DNS) yang dapat diskalakan dan dengan ketersediaan tinggi seperti Amazon Route 53 untuk mengelola catatan DNS domain Anda. Anda juga dapat menggunakan pemeriksaan kondisi yang disediakan oleh Amazon Route 53. Pemeriksaan kondisi memverifikasi bahwa aplikasi Anda dapat dijangkau, tersedia, dan berfungsi, dan pemeriksaan ini dapat diatur sedemikian sehingga dapat menyerupai perilaku pengguna Anda, seperti meminta halaman web atau URL tertentu. Jika terjadi kegagalan, Amazon Route 53 merespons permintaan resolusi DNS dan mengarahkan lalu lintas ke titik akhir dengan kondisi bagus saja. Anda juga dapat mempertimbangkan penggunaan kemampuan Perutean Berbasis Latensi dan DNS Geo yang ditawarkan oleh Amazon Route 53.

Untuk memverifikasi bahwa beban kerja Anda sendiri mempunyai ketersediaan yang tinggi, gunakan Elastic Load Balancing (ELB). Amazon Route 53 dapat digunakan untuk menargetkan lalu lintas ke ELB, yang mendistribusikan lalu lintas ke instans komputasi target. Anda juga dapat menggunakan Amazon API Gateway bersama dengan AWS Lambda untuk solusi nirserver. Pelanggan juga dapat menjalankan beban kerja di beberapa Wilayah AWS. Dengan pola aktif/aktif multi-situs, beban kerja dapat melayani lalu lintas yang datang dari beberapa Wilayah. Dengan pola aktif/pasif multi-situs, beban kerja dapat melayani lalu lintas dari wilayah aktif sementara data akan direplikasi ke wilayah sekunder dan menjadi aktif jika terjadi kegagalan di wilayah primer. Pemeriksaan kondisi Route 53 kemudian dapat digunakan untuk mengontrol failover DNS dari titik akhir mana pun di Wilayah primer

ke titik akhir di Wilayah sekunder, sehingga memverifikasi bahwa beban kerja Anda dapat dijangkau dan tersedia untuk pengguna Anda.

Amazon CloudFront memberikan API yang sederhana untuk mendistribusikan konten dengan laju transfer data yang tinggi dan latensi yang rendah dengan melayani permintaan menggunakan jaringan lokasi edge di seluruh dunia. Jaringan pengiriman konten (CDN) melayani para pelanggan dengan menghadirkan konten yang berada di atau di-cache di lokasi yang dekat dengan pengguna. Ini juga akan meningkatkan ketersediaan aplikasi Anda karena beban untuk konten dialihkan dari server Anda ke lokasi edge CloudFront. Lokasi edge dan cache edge regional menyimpan cache salinan konten Anda berada dekat dengan penonton Anda sehingga konten dapat diambil dengan cepat, konten lebih mudah dijangkau, dan beban kerja memiliki ketersediaan yang lebih tinggi.

Untuk beban kerja yang memiliki pengguna yang tersebar secara geografis, AWS Global Accelerator akan membantu Anda dalam meningkatkan ketersediaan dan performa aplikasi. AWS Global Accelerator memberikan alamat IP statis Anycast yang berfungsi sebagai titik masuk tetap ke aplikasi Anda yang di-host di satu atau beberapa Wilayah AWS. Hal ini akan memungkinkan lalu lintas masuk ke jaringan global AWS sedekat mungkin dengan para pengguna Anda, sehingga akan meningkatkan keterjangkauan dan ketersediaan beban kerja Anda. AWS Global Accelerator juga memantau kondisi titik akhir aplikasi Anda dengan menggunakan TCP, HTTP, dan pemeriksaan kondisi HTTPS. Setiap perubahan yang terjadi pada kondisi atau konfigurasi titik akhir Anda akan mengizinkan pengarahan ulang lalu lintas pengguna ke titik akhir yang memiliki kondisi kesehatan yang bagus yang memberikan ketersediaan dan performa terbaik bagi pengguna Anda. Selain itu, AWS Global Accelerator juga memiliki desain yang dapat mengisolasi kesalahan yang menggunakan dua alamat IPv4 status yang dilayani oleh zona jaringan mandiri sehingga dapat meningkatkan ketersediaan aplikasi Anda.

Untuk membantu Anda melindungi pelanggan dari serangan DDoS, AWS menyediakan AWS Shield Standard. Shield Standard secara otomatis dinyalakan dan akan melindungi dari serangan infrastruktur umum (lapisan 3 dan 4) seperti banjir SYN/UDP dan serangan refleksi untuk mendukung ketersediaan aplikasi Anda yang tinggi di AWS. Untuk perlindungan tambahan dari serangan yang lebih besar dan lebih canggih (seperti banjir UDP), serangan penghentian layanan (seperti banjir TCP SYN), dan untuk membantu melindungi aplikasi Anda yang dijalankan di Amazon Elastic Compute Cloud (Amazon EC2), Elastic Load Balancing (ELB), Amazon CloudFront, AWS Global Accelerator, dan Route 53, Anda dapat mempertimbangkan untuk menggunakan AWS Shield Advanced. Untuk perlindungan dari serangan lapisan Aplikasi seperti HTTP POST atau GET flood, gunakan AWS WAF. AWS WAF dapat menggunakan alamat IP, header HTTP, bodi HTTP, string URI, injeksi SQL, dan kondisi skrip lintas situs untuk menentukan apakah permintaan harus diblokir atau diizinkan.

Langkah-langkah implementasi

1. Siapkan DNS dengan ketersediaan tinggi: Amazon Route 53 adalah layanan web Sistem Nama Domain (DNS) yang dapat diskalakan dan memiliki ketersediaan tinggi. Route 53 menghubungkan permintaan pengguna ke aplikasi internet yang dijalankan di AWS atau on-premise. Untuk informasi selengkapnya, silakan lihat Mengonfigurasi Amazon Route 53 sebagai layanan DNS Anda.
2. Siapkan pemeriksaan kondisi: Ketika menggunakan Route 53, verifikasi bahwa hanya target dengan kondisi bagus yang dapat diselesaikan. Mulailah dengan membuat pemeriksaan kondisi Route 53 dan mengonfigurasi failover DNS. Aspek-aspek berikut penting untuk Anda pertimbangkan ketika mempersiapkan pemeriksaan kondisi:
    a. Bagaimana Amazon Route 53 menentukan apakah pemeriksaan kondisi hasilnya sehat atau tidak
    b. Membuat, memperbarui, dan menghapus pemeriksaan kondisi
    c. Memantau status pemeriksaan kondisi dan mendapatkan notifikasi
    d. Praktik terbaik untuk DNS Amazon Route 53
3. Hubungkan layanan DNS Anda ke titik akhir Anda.
    a. Saat menggunakan Penyeimbangan Beban Elastis sebagai target untuk lalu lintas Anda, buat catatan alias menggunakan Amazon Route 53 yang mengarah ke titik akhir regional penyeimbang beban Anda. Selama pembuatan catatan alias, atur opsi Evaluasi kondisi target ke Ya.
    b. Untuk beban kerja nirserver atau API pribadi saat API Gateway digunakan, gunakan Route 53 untuk mengarahkan lalu lintas ke API Gateway.
4. Buat keputusan mengenai jaringan pengiriman konten.
    a. Untuk mengirimkan konten dengan menggunakan lokasi edge yang lebih dekat dengan pengguna, mulailah dengan memahami cara CloudFront memberikan konten.
    b. Memulai dengan distribusi CloudFront yang sederhana. Kemudian CloudFront akan mengetahui dari mana Anda ingin konten dikirimkan, dan detail tentang cara melacak dan mengelola pengiriman konten. Aspek-aspek berikut penting untuk dipahami dan dipertimbangkan ketika Anda mempersiapkan distribusi CloudFront:
        i. Cara kerja caching dengan lokasi edge CloudFront
        ii. Meningkatkan proporsi permintaan yang disajikan secara langsung dari cache CloudFront (rasio temuan cache)
        iii. Menggunakan Amazon CloudFront Origin Shield

    iv. Mengoptimalkan ketersediaan yang tinggi dengan failover asal CloudFront

5. Siapkan perlindungan lapisan aplikasi: AWS WAF akan membantu Anda melindungi dari bot dan eksploitasi web umum yang dapat memengaruhi ketersediaan, mengancam keamanan, atau memakai sumber daya secara berlebihan. Untuk mendapatkan pemahaman yang lebih dalam, silakan tinjau cara kerja AWS WAF dan kapan Anda siap menerapkan perlindungan dari lapisan aplikasi HTTP POST DAN GET floods, tinjau Memulai dengan AWS WAF. Anda juga dapat menggunakan AWS WAF dengan CloudFront, silakan lihat dokumentasi tentang cara kerja AWS WAF dengan fitur Amazon CloudFront.
6. Siapkan perlindungan DDoS tambahan: Secara default, semua pelanggan AWS akan menerima perlindungan dari serangan DDoS lapisan transpor dan jaringan paling sering terjadi yang menargetkan situs web atau aplikasi Anda, dengan menggunakan AWS Shield Standard tanpa biaya tambahan. Untuk perlindungan tambahan terhadap aplikasi yang terhubung ke internet yang berjalan di Amazon EC2, Penyeimbangan Beban Elastis, Amazon CloudFront, AWS Global Accelerator, dan Amazon Route 53, Anda dapat mempertimbangkan AWS Shield Advanced dan meninjau contoh arsitektur tangguh DDoS. Untuk melindungi beban kerja dan titik akhir publik Anda dari serangan DDoS, silakan tinjau Memulai AWS Shield Advanced.

Sumber daya

Praktik-praktik terbaik terkait:

- REL10-BP01 Melakukan deployment beban kerja ke beberapa lokasi
- REL11-BP04 Mengandalkan bidang data dan bukan bidang kontrol selama pemulihan
- REL11-BP06 Mengirimkan notifikasi ketika peristiwa memengaruhi ketersediaan

Dokumen terkait:

- Partner APN: partner yang dapat membantu Anda merencanakan jaringan Anda
- AWS Marketplace untuk Infrastruktur Jaringan
- Apa itu AWS Global Accelerator?
- Apa yang dimaksud dengan Amazon CloudFront?
- Apa itu Amazon Route 53?
- Apa itu Penyeimbangan Beban Elastis?
- Kemampuan Konektivitas Jaringan - Membangun Fondasi Cloud Anda

- Apa itu Amazon API Gateway?
- Apa itu AWS WAF, AWS Shield, dan AWS Firewall Manager?
- Apa itu Pengontrol Pemulihan Aplikasi Amazon?
- Konfigurasikan pemeriksaan kondisi kustom untuk failover DNS

Video terkait:

- AWS re:Invent 2022 - Meningkatkan performa dan ketersediaan dengan AWS Global Accelerator
- AWS re:Invent 2020: Manajemen lalu lintas global dengan Amazon Route 53
- AWS re:Invent 2022 - Mengoperasikan aplikasi Multi-AZ dengan ketersediaan tinggi
- AWS re:Invent 2022 - Memahami lebih dalam infrastruktur jaringan AWS
- AWS re:Invent 2022 - Membangun jaringan yang tangguh

Contoh terkait:

- Pemulihan Bencana dengan Pengendali Pemulihan Aplikasi (ARC) Amazon
- Lokakarya AWS Global Accelerator

REL02-BP02 Menyediakan konektivitas redundan antara jaringan privat di cloud dan lingkungan on-premise

Implementasikan redundansi dalam koneksi Anda antara jaringan pribadi di cloud dan lingkungan on-premise untuk mencapai ketahanan konektivitas. Ini dapat dicapai dengan melakukan deployment dua atau lebih tautan dan jalur lalu lintas, sehingga menjaga konektivitas jika terjadi kegagalan jaringan.

Anti-pola umum:

- Anda bergantung hanya pada satu koneksi jaringan, yang akan menciptakan satu titik kegagalan.
- Anda hanya menggunakan satu terowongan VPN atau beberapa terowongan yang berakhir di Zona Ketersediaan yang sama.
- Anda mengandalkan satu ISP untuk konektivitas VPN, yang dapat menyebabkan terjdainya kegagalan total selama terjadi penghentian (outage) ISP.
- Tidak mengimplementasikan protokol perutean dinamis seperti BGP, yang sangat penting untuk melakukan perutean ulang lalu lintas saat terjadi gangguan jaringan.

- Anda mengabaikan batasan bandwidth terowongan VPN dan melebih-lebihkan kemampuan cadangannya.

Manfaat menerapkan praktik terbaik ini: Dengan mengimplementasikan konektivitas redundan antara lingkungan cloud dan lingkungan perusahaan atau on-premise Anda, maka layanan-layanan dependen yang ada di antara dua lingkungan tersebut dapat berkomunikasi dengan andal.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Tinggi

Panduan implementasi

Saat Anda menggunakan AWS Direct Connect untuk menghubungkan jaringan on-premise ke AWS, Anda dapat mencapai ketahanan jaringan maksimum (SLA 99,99%) dengan menggunakan koneksi terpisah yang berakhir pada perangkat berbeda yang ada di lebih dari satu lokasi on-premise dan lebih dari satu lokasi AWS Direct Connect. Topologi ini menawarkan Anda ketahanan terhadap kegagalan perangkat, masalah konektivitas, dan pemadaman (outage) lokasi total. Atau, Anda dapat mencapai ketahanan tinggi (SLA 99,9%) dengan menggunakan dua koneksi individual ke beberapa lokasi (setiap lokasi on-premise terhubung ke satu lokasi Direct Connect). Pendekatan ini akan melindungi Anda dari gangguan konektivitas yang disebabkan oleh pemotongan serat atau kegagalan perangkat dan membantu Anda mengurangi kegagalan lokasi total. Direct Connect Resiliency Toolkit dapat membantu merancang desain topologi AWS Direct Connect Anda.

Anda juga dapat mempertimbangkan untuk mengunakan AWS Site-to-Site VPN yang berakhir pada AWS Transit Gateway sebagai cadangan yang hemat biaya ke koneksi AWS Direct Connect utama Anda. Pengaturan ini akan memungkinkan Anda melakukan perutean multi-jalur biaya setara (ECMP) di beberapa terowongan VPN, yang dapat memungkinkan throughput hingga 50Gbps, meskipun setiap terowongan VPN dibatasi 1,25 Gbps. Namun demikian, penting untuk diingat bahwa AWS Direct Connect tetap merupakan pilihan paling efektif untuk meminimalkan gangguan jaringan dan menyediakan konektivitas yang stabil.

Saat menggunakan VPN melalui internet untuk menghubungkan lingkungan cloud Anda ke pusat data on-premise Anda, konfigurasikan dua terowongan VPN sebagai bagian dari sebuah koneksi VPN situs-ke-situs tunggal. Setiap terowongan harus berakhir di Zona Ketersediaan yang berbeda untuk mendapatkan ketersediaan yang tinggi dan harus menggunakan perangkat keras redundan untuk mencegah kegagalan perangkat on-premise. Selain itu, pertimbangkan juga untuk menggunakan beberapa koneksi internet dari berbagai penyedia layanan internet (ISP) di lokasi on-premise Anda untuk menghindari gangguan konektivitas VPN total karena terjadi satu pemadaman

(outage) ISP. Memilih ISP dengan perutean dan infrastruktur yang beragam, terutama yang memiliki jalur fisik terpisah ke titik akhir AWS, akan memberikan Anda ketersediaan konektivitas yang tinggi.

Selain redundansi fisik yang memiliki beberapa koneksi AWS Direct Connect dan beberapa terowongan VPN (atau kombinasi keduanya), implementasi perutean dinamis Protokol Gateway Batas (BGP) juga merupakan hal penting. BGP yang dinamis akan menyediakan pengalihan rute lalu lintas secara otomatis dari satu jalur ke jalur yang lain berdasarkan kondisi jaringan waktu nyata dan kebijakan yang sudah dikonfigurasikan. Perilaku dinamis ini sangat bermanfaat dalam menjaga ketersediaan jaringan dan kontinuitas layanan jika ada kegagalan tautan atau jaringan yang terjadi. Jalur alternatif dipilih dengan cepat, sehingga dapat meningkatkan ketahanan dan keandalan jaringan.

Langkah-langkah implementasi

- Akuisisi konektivitas dengan ketersediaan tinggi antara AWS dan lingkungan on-premise Anda.
  - Gunakan beberapa koneksi AWS Direct Connect atau terowongan VPN antara jaringan privat yang di-deploy secara terpisah.
  - Gunakan lebih dari satu lokasi Direct Connect untuk ketersediaan tinggi.
  - Ketika menggunakan beberapa Wilayah AWS, ciptakan redundansi setidaknya di dalam dua di antaranya.
- Gunakan AWS Transit Gateway, jika memungkinkan, untuk mengakhiri koneksi VPN Anda.
- Evaluasi peralatan AWS Marketplace untuk mengakhiri VPN atau untuk memperluas SD-WAN Anda ke AWS. Ketika Anda menggunakan peralatan AWS Marketplace, lakukan deployment instans redundan untuk ketersediaan yang tinggi di Zona Ketersediaan yang berbeda.
- Sediakan koneksi redundan ke lingkungan on-premise Anda.
  - Anda mungkin memerlukan koneksi redundan ke beberapa Wilayah AWS agar ketersediaan yang Anda butuhkan bisa dicapai.
  - Menggunakan Direct Connect Resiliency Toolkit untuk memulai.

Sumber daya

Dokumen terkait:

- Rekomendasi Ketahanan AWS Direct Connect
- Menggunakan Koneksi VPN Site-to-Site Redundan untuk Memberikan Failover
- Kebijakan Perutean dan Komunitas BGP

- Konfigurasi Aktif/Aktif dan Aktif/Pasif di AWS Direct Connect
- Partner APN: partner yang dapat membantu Anda merencanakan jaringan Anda
- AWS Marketplace untuk Infrastruktur Jaringan
- Laporan Resmi Pilihan Konektivitas Amazon Virtual Private Cloud
- Membangun Infrastruktur Jaringan AWS Multi-VPC yang Dapat Diskalakan dan Aman
- Menggunakan koneksi VPN Site-to-Site redundan untuk memberikan failover
- Menggunakan Direct Connect Resiliency Toolkit untuk memulai
- Titik Akhir VPC dan Layanan Titik Akhir VPC (AWS PrivateLink)
- Apa yang Dimaksud Amazon VPC?
- Apa itu gateway transit?
- Apa itu AWS Site-to-Site VPN?
- Bekerja dengan gateway Direct Connect

Video terkait:

- AWS re:Invent 2018: Desain VPC Tingkat Lanjut dan Kemampuan Baru untuk Amazon VPC
- AWS re:Invent 2019: Arsitektur referensi AWS Transit Gateway untuk berbagai VPC

REL02-BP03 Pastikan alokasi subnet IP menjelaskan ekspansi dan ketersediaan

Rentang alamat IP Amazon VPC harus cukup besar untuk mengakomodasi persyaratan beban kerja, termasuk pertimbangan ekspansi mendatang dan alokasi alamat IP ke subnet di seluruh Zona Ketersediaan. Ini mencakup penyeimbang beban, instans EC2, dan aplikasi berbasis kontainer.

Ketika Anda merencanakan topologi jaringan Anda, langkah pertama yang harus dilakukan adalah menetapkan ruang alamat IP itu sendiri. Rentang alamat IP privat (mengikuti pedoman RFC 1918) harus dialokasikan untuk masing-masing VPC. Akomodasikan persyaratan berikut sebagai bagian dari proses ini:

- Berikan ruang alamat IP untuk lebih dari satu VPC untuk setiap Wilayah.
- Di dalam sebuah VPC, berikan ruang untuk beberapa subnet sehingga Anda dapat mencakup beberapa Zona Ketersediaan.

- Pertimbangkan untuk membiarkan ruang blok CIDR yang tidak digunakan di dalam sebuah VPC untuk melakukan ekspansi di masa mendatang.
- Pastikan ada ruang alamat IP untuk memenuhi kebutuhan armada sementara instans Amazon EC2 yang mungkin Anda gunakan, seperti Armada Spot untuk machine learning, klaster Amazon EMR, atau klaster Amazon Redshift. Pertimbangan serupa sebaiknya diberikan ke klaster Kubernetes, seperti Amazon Elastic Kubernetes Service (Amazon EKS), karena setiap pod Kubernetes diberi alamat yang dapat dirutekan dari blok CIDR VPC secara default.
- Perlu diperhatikan bahwa empat alamat IP pertama dan alamat IP terakhir di setiap blok CIDR subnet sudah terpesan dan tidak tersedia untuk Anda gunakan.
- Perlu diperhatikan bahwa blok CIDR VPC awal yang dialokasikan ke VPC Anda tidak dapat diubah atau dihapus, tetapi Anda dapat menambahkan tambahan blok CIDR yang tidak tumpang tindih ke VPC. CIDR IPv4 subnet tidak dapat diubah, tetapi CIDR IPv6 dapat diubah.
- Blok CIDR VPC terbesar yang memungkinkan adalah /16, dan yang terkecil adalah /28.
- Pertimbangkan untuk menggunakan jaringan terkoneksi lain (VPC, on-premise, atau penyedia cloud lainnya) dan pastikan ruang alamat IP tidak tumpang tindih. Untuk informasi selengkapnya, lihat REL02-BP05 Terapkan rentang alamat IP privat yang tidak tumpang tindih di semua ruang alamat privat tempat semuanya saling terhubung.

Hasil yang diinginkan: Sebuah subnet IP yang dapat diskalakan dapat membantu Anda mengakomodasi pertumbuhan yang mungkin terjadi di masa depan dan menghindari pemborosan yang tidak perlu.

Anti-pola umum:

- Jika Anda gagal mempertimbangkan pertumbuhan yang mungkin terjadi di masa depan, maka hal itu akan mengakibatkan blok CIDR yang terlalu kecil dan membutuhkan konfigurasi ulang, serta berpotensi menyebabkan waktu henti.
- Salah memperkirakan jumlah alamat IP yang dapat digunakan oleh satu penyeimbang beban elastis.
- Melakukan deployment banyak penyeimbang beban lalu lintas tinggi ke subnet yang sama
- Menggunakan mekanisme penskalaan otomatis tetapi gagal memantau pemakaian alamat IP.
- Menentukan rentang CIDR yang terlalu besar melampaui ekspektasi pertumbuhan yang mungkin terjadi di masa depan, sehingga menyebabkan adanya kesulitan untuk melakukan peering dengan jaringan lain dengan rentang alamat yang tumpang tindih.

Manfaat menerapkan praktik terbaik ini: Ini akan memastikan bahwa Anda dapat mengakomodasi pertumbuhan beban kerja Anda dan akan terus memberikan ketersediaan saat Anda menaikkan skala.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Sedang

Panduan implementasi

Rencanakan jaringan Anda untuk mengakomodasi pertumbuhan, kepatuhan terhadap peraturan, dan integrasi dengan jaringan yang lain. Pertumbuhan mungkin saja lebih besar dari yang diperkirakan, kepatuhan terhadap peraturan dapat berubah, dan koneksi jaringan privat atau akuisisi bisa jadi akan sulit diimplementasikan tanpa melakukan perencanaan yang baik.

- Pilih Akun AWS dan Wilayah yang relevan berdasarkan persyaratan layanan, latensi, peraturan, dan pemulihan bencana (DR) Anda.
- Identifikasi kebutuhan Anda untuk deployment VPC regional.
- Identifikasi ukuran VPC.
  - Tentukan apakah Anda akan melakukan deployment konektivitas multi-VPC.
    - Apa Itu Gateway Transit?
    - Konektivitas Multi-VPC Satu Wilayah
  - Tentukan apakah Anda membutuhkan jaringan terpisah (segregated) untuk persyaratan berdasarkan regulasi.
  - Buat VPC dengan blok CIDR yang mempunyai ukuran yang sesuai untuk mengakomodasi kebutuhan Anda saat ini dan kebutuhan di masa depan.
    - Jika Anda memiliki proyeksi pertumbuhan yang tidak diketahui, maka Anda sebaiknya melakukan kesalahan di sisi blok CIDR yang lebih besar untuk mengurangi potensi konfigurasi ulang di masa depan
  - Pertimbangkan untuk menggunakan pengalamatan IPv6 untuk subnet sebagai bagian dari VPC tumpukan ganda. IPv6 sangat cocok untuk digunakan di subnet privat yang memuat armada instans sementara atau kontainer yang seharusnya membutuhkan alamat IPv4 dalam jumlah besar.

Sumber daya

Praktik terbaik Well-Architected terkait:

- REL02-BP05 Terapkan rentang alamat IP privat yang tidak tumpang tindih di semua ruang alamat privat tempat semuanya saling terhubung

Dokumen terkait:

- Partner APN: partner yang dapat membantu Anda merencanakan jaringan Anda
- AWS Marketplace untuk Infrastruktur Jaringan
- Laporan Resmi Pilihan Konektivitas Amazon Virtual Private Cloud
- Konektivitas jaringan ketersediaan tinggi (HA) beberapa pusat data
- Konektivitas Multi-VPC Satu Wilayah
- Apa yang Dimaksud Amazon VPC?
- IPv6 di AWS
- IPv6 di arsitektur referensi
- Amazon Elastic Kubernetes Service meluncurkan dukungan IPv6
- Rekomendasi untuk VPC Anda - Penyeimbang Beban Klasik
- Subnet Zona Ketersediaan - Penyeimbang Beban Aplikasi
- Zona Ketersediaan - Penyeimbang Beban Jaringan

Video terkait:

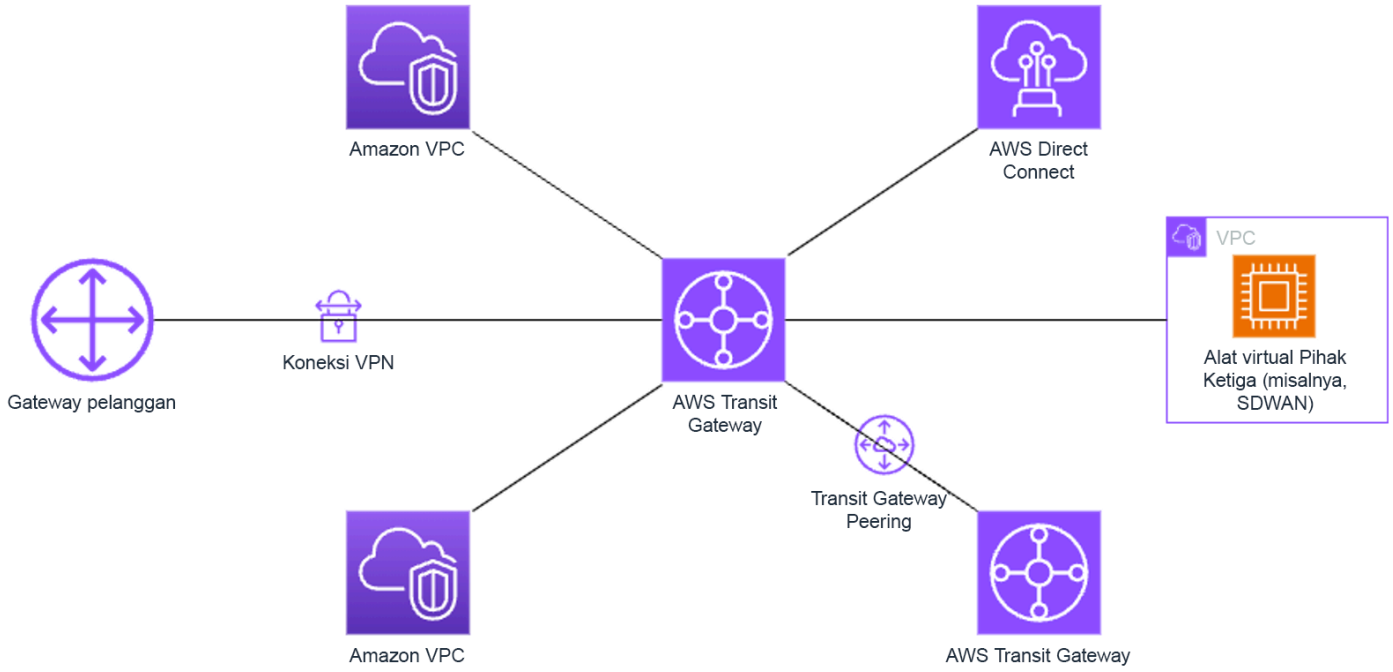
- AWS re:Invent 2018: Desain VPC Tingkat Lanjut dan Kemampuan Baru untuk Amazon VPC (NET303)
- AWS re:Invent 2019: Arsitektur referensi AWS Transit Gateway untuk berbagai VPC (NET406-R1)
- AWS re:Invent 2023: AWS Siap dengan yang berikutnya? Merancang jaringan untuk pertumbuhan dan fleksibilitas (NET310)

REL02-BP04 Mengutamakan topologi hub-and-spoke daripada mesh many-to-many

Saat menghubungkan beberapa jaringan privat, seperti Cloud Privat Virtual (VPC) dan jaringan on-premise, pilihlah topologi "hub and spoke", bukan topologi "mesh". Tidak seperti topologi mesh, di mana setiap jaringan terhubung langsung ke jaringan lain dan meningkatkan kompleksitas serta overhead manajemen, arsitektur hub and spoke memusatkan koneksi melalui satu hub tunggal.

Sentralisasi ini menyederhanakan struktur jaringan dan meningkatkan operabilitas, skalabilitas, dan kontrolnya.

AWS Transit Gateway adalah sebuah layanan terkelola yang dapat diskalakan dan memiliki ketersediaan tinggi yang dirancang untuk pembangunan konstruksi jaringan hub and spoke di AWS. Layanan ini berfungsi sebagai hub pusat jaringan Anda yang menyediakan segmentasi jaringan, perutean tersentralisasi, dan koneksi yang disederhanakan ke lingkungan cloud dan on-premise. Gambar berikut menunjukkan cara menggunakan AWS Transit Gateway untuk membangun topologi hub and spoke Anda.



Hasil yang diinginkan: Anda telah menghubungkan Cloud Privat Virtual (VPC) dan jaringan on-premise melalui hub pusat. Anda mengonfigurasi koneksi peering Anda melalui hub, yang bertindak sebagai router cloud yang sangat mudah diskalakan. Perutean disederhanakan karena Anda tidak perlu bekerja dengan hubungan peering yang kompleks. Lalu lintas antar-jaringan dienkripsi, dan Anda memiliki kemampuan untuk mengisolasi jaringan.

Anti-pola umum:

- Anda membuat aturan peering jaringan yang kompleks.
- Anda menyediakan rute antar-jaringan yang seharusnya tidak saling berkomunikasi (misalnya, beban kerja terpisah yang tidak memiliki interdependensi).
- Ada tata kelola yang tidak efektif untuk instans hub.

Manfaat menjalankan praktik terbaik ini: Ketika jumlah jaringan yang terhubung meningkat, manajemen dan perluasan konektivitas "mesh" menjadi semakin sulit. Arsitektur "mesh" menimbulkan kesulitan tambahan, seperti komponen infrastruktur tambahan, persyaratan konfigurasi, dan pertimbangan deployment. Mesh juga menimbulkan overhead tambahan untuk mengelola dan memantau komponen bidang data dan bidang kontrol. Anda harus memikirkan cara memberikan ketersediaan tinggi arsitektur mesh, cara memantau kesehatan dan kinerja mesh, serta cara menangani peningkatan komponen mesh.

Model hub-and-spoke, di sisi lain, menetapkan perutean lalu lintas terpusat di beberapa jaringan. Model ini memberikan pendekatan yang lebih sederhana untuk manajemen dan pemantauan komponen bidang data dan bidang kontrol.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Sedang

Panduan implementasi

Buat akun Layanan Jaringan jika belum ada. Tempatkan hub di akun Layanan Jaringan organisasi. Pendekatan ini memungkinkan hub dikelola secara terpusat oleh rekayasawan jaringan.

Hub model hub-and-spoke bertindak sebagai router virtual untuk lalu lintas yang mengalir antara Cloud Privat Virtual (VPC) dan jaringan on-premise. Pendekatan ini mengurangi kompleksitas jaringan dan memudahkan pemecahan masalah jaringan.

Pertimbangkan desain jaringan Anda, termasuk VPC, AWS Direct Connect, dan koneksi VPN Site-to-Site yang ingin Anda hubungkan.

Pertimbangkan untuk menggunakan subnet terpisah untuk setiap koneksi VPC gateway transit. Untuk setiap subnet, gunakan CIDR kecil (misalnya /28) agar Anda memiliki lebih banyak ruang alamat untuk sumber daya komputasi. Selain itu, buat satu ACL jaringan, dan kaitkan dengan semua subnet yang terkait dengan hub tersebut. Jaga agar ACL jaringan tetap terbuka di arah masuk dan keluar.

Rancang dan implementasikan tabel rute Anda sedemikian rupa agar rute hanya disediakan di antara jaringan yang seharusnya berkomunikasi. Hilangkan rute antar-jaringan yang seharusnya tidak saling berkomunikasi (misalnya, di antara beban kerja terpisah yang tidak memiliki interdependensi).

Langkah-langkah implementasi

1. Rencanakan jaringan Anda. Tentukan jaringan mana yang ingin Anda sambungkan, dan verifikasi bahwa jaringan tersebut tidak berbagi rentang CIDR yang tumpang-tindih.

2. Buat AWS Transit Gateway dan hubungkan VPC Anda.
3. Jika diperlukan, buatlah koneksi VPN atau gateway Direct Connect dan kaitkan dengan Transit Gateway.
4. Tentukan bagaimana lalu lintas dirutekan di antara VPC yang terhubung dan koneksi lain melalui konfigurasi tabel rute Transit Gateway Anda.
5. Gunakan Amazon CloudWatch untuk melakukan pemantauan dan penyesuaian konfigurasi yang diperlukan untuk optimalisasi performa dan biaya.

Sumber daya

Praktik-praktik terbaik terkait:

- REL02-BP03 Pastikan alokasi subnet IP memperhitungkan ekspansi dan ketersediaan
- REL02-BP05 Terapkan rentang alamat IP privat yang tidak tumpang-tindih di semua ruang alamat privat tempat semuanya saling terhubung

Dokumen terkait:

- Apa Itu Gateway Transit?
- Praktik terbaik desain gateway transit
- Membangun Infrastruktur Jaringan AWS Multi-VPC yang Dapat Diskalakan dan Aman
- Membangun jaringan global menggunakan peering AWS Transit Gateway Inter-Region
- Pilihan Konektivitas Amazon Virtual Private Cloud
- Partner APN: partner yang dapat membantu Anda merencanakan jaringan Anda
- AWS Marketplace untuk Infrastruktur Jaringan

Video terkait:

- AWS re:Invent 2023 - Fondasi jaringan AWS
- AWS re:Invent 2023 - Desain VPC tingkat lanjut dan kemampuan baru

Lokakarya terkait:

- Lokakarya AWS Transit Gateway

REL02-BP05 Terapkan rentang alamat IP privat yang tidak tumpang tindih di semua ruang alamat privat tempat semuanya saling terhubung

Rentang alamat IP dari setiap VPC Anda tidak boleh tumpang tindih ketika dilakukan peering, terhubung melalui Transit Gateway, atau terhubung melalui VPN. Hindari konflik alamat IP antara lingkungan on-premise dan VPC atau dengan penyedia cloud lain yang Anda gunakan. Selain itu, Anda harus memiliki cara untuk mengalokasikan rentang alamat IP privat ketika dibutuhkan. Sistem manajemen alamat IP (IPAM) dapat membantu mengotomatiskannya.

Hasil yang diinginkan:

- Tidak ada konflik rentang alamat IP antara VPC, lingkungan on-premise, atau penyedia cloud lain.
- Manajemen alamat IP yang tepat akan memungkinkan penskalaan infrastruktur jaringan yang lebih mudah untuk mengakomodasi pertumbuhan dan perubahan persyaratan jaringan.

Anti-pola umum:

- Menggunakan rentang IP yang sama di VPC Anda seperti yang Anda miliki on-premise, di jaringan perusahaan Anda, atau penyedia cloud lainnya
- Tidak melacak rentang IP VPC yang digunakan untuk deployment beban kerja Anda.
- Mengandalkan proses manajemen alamat IP manual, seperti spreadsheet.
- Blok CIDR yang terlalu besar atau kecil, yang mengakibatkan pemborosan alamat IP atau ruang alamat, tidak akan mencukupi untuk beban kerja Anda.

Manfaat menerapkan praktik terbaik ini: Perencanaan aktif jaringan Anda akan memastikan bahwa Anda tidak memiliki beberapa kejadian alamat IP yang sama di jaringan-jaringan yang saling terhubung. Ini akan mencegah timbulnya masalah perutean di bagian beban kerja yang menggunakan aplikasi yang berbeda.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Sedang

Panduan implementasi

Manfaatkan IPAM, seperti Amazon VPC IP Address Manager, untuk memantau dan mengelola penggunaan CIDR Anda. Beberapa IPAM juga tersedia dari AWS Marketplace. Evaluasi potensi penggunaan Anda di AWS, tambahkan rentang CIDR ke VPC yang ada, dan buatlah VPC untuk memungkinkan pertumbuhan yang direncanakan dalam penggunaan.

Langkah-langkah implementasi

- Catat konsumsi CIDR saat ini (misalnya, VPC dan subnet).
  - Gunakan operasi API layanan untuk mengumpulkan konsumsi CIDR saat ini.
  - Gunakan Manajer Alamat IP VPC Amazon untuk menemukan sumber daya.
- Catat penggunaan subnet Anda saat ini.
  - Gunakan operasi API layanan untuk mengumpulkan subnet per VPC di setiap Wilayah.
  - Gunakan Manajer Alamat IP VPC Amazon untuk menemukan sumber daya.
- Catat penggunaan saat ini.
- Tentukan apakah Anda telah membuat rentang IP yang tumpang tindih.
- Hitung kapasitas cadangan.
- Identifikasi rentang IP yang tumpang tindih. Anda dapat bermigrasi ke rentang alamat baru atau mempertimbangkan untuk menggunakan teknik seperti Gateway NAT privat atau AWS PrivateLink jika Anda perlu menghubungkan rentang yang tumpang tindih.

Sumber daya

Praktik-praktik terbaik terkait:

- Melindungi jaringan

Dokumen terkait:

- Partner APN: partner yang dapat membantu Anda merencanakan jaringan Anda
- AWS Marketplace untuk Infrastruktur Jaringan
- Laporan Resmi Pilihan Konektivitas Amazon Virtual Private Cloud
- Konektivitas jaringan ketersediaan tinggi (HA) beberapa pusat data
- Menghubungkan Jaringan dengan Rentang IP yang Tumpang Tindih
- Apa yang Dimaksud Amazon VPC?
- Apa itu IPAM?

Video terkait:

- AWS re:Invent 2023 - Desain VPC tingkat lanjut dan kemampuan baru

- AWS re:Invent 2019: Arsitektur referensi AWS Transit Gateway untuk berbagai VPC
- AWS re:Invent 2023 - Siap dengan yang berikutnya? Merancang jaringan untuk pertumbuhan dan fleksibilitas
- AWS re:Invent 2021 - {New Launch} Kelola alamat IP Anda dalam skala besar di AWS

Arsitektur beban kerja

Beban kerja yang andal dimulai dengan desain perangkat lunak dan infrastruktur yang diputuskan sejak awal. Pilihan arsitektur Anda akan memengaruhi perilaku beban kerja Anda di keenam pilar Well-Architected. Untuk keandalan, terdapat beberapa pola tertentu yang harus diikuti.

Bagian-bagian berikutnya menjelaskan praktik terbaik untuk digunakan dengan pola-pola keandalan ini.

Topik

Mendesain arsitektur layanan beban kerja Anda

Bangun beban kerja yang andal dan dapat diskalakan dengan mudah menggunakan arsitektur berorientasi layanan (SOA) atau arsitektur layanan mikro. Arsitektur berorientasi layanan (SOA) adalah praktik untuk membuat komponen perangkat lunak dapat digunakan ulang lewat antarmuka layanan. Arsitektur layanan mikro melangkah lebih jauh untuk membuat komponen menjadi lebih kecil dan lebih sederhana.

Antarmuka arsitektur berorientasi layanan (SOA) menggunakan standar komunikasi umum sehingga dapat dimasukkan dengan cepat ke dalam beban kerja baru. SOA menggantikan praktik pembangunan arsitektur monolit, yang terdiri dari unit-unit tak dapat dibagi dan saling bergantung satu sama lain.

Di AWS, kami selalu menggunakan SOA, tetapi kini kami sudah mulai membangun sistem-sistem kami menggunakan layanan mikro. Meskipun layanan mikro memiliki sejumlah kualitas yang menarik, manfaat paling penting untuk ketersediaan adalah ukurannya yang lebih kecil dan sifatnya yang lebih sederhana. Layanan mikro memungkinkan Anda untuk membedakan ketersediaan yang diperlukan dari layanan yang berbeda, dan oleh karena itu fokuskan investasi terutama ke layanan mikro yang memiliki kebutuhan ketersediaan paling besar. Sebagai contoh, untuk menghadirkan halaman informasi produk di Amazon.com (“halaman detail”), ratusan layanan mikro diinvokasi untuk membangun bagian-bagian halaman yang terpisah. Meskipun terdapat beberapa layanan yang harus tersedia untuk menyediakan harga dan detail produk, sebagian besar konten di halaman tersebut

dapat dikecualikan jika layanan tidak tersedia. Bahkan hal-hal seperti foto dan ulasan tidak diperlukan untuk menyediakan pengalaman bagi pelanggan dalam membeli sebuah produk.

Praktik terbaik

- REL03-BP01 Pilih cara mengelompokkan beban kerja Anda
- REL03-BP02 Bangun layanan yang berfokus pada domain dan fungsionalitas bisnis khusus
- REL03-BP03 Memberikan kontrak layanan per API

REL03-BP01 Pilih cara mengelompokkan beban kerja Anda

Segmentasi beban kerja penting saat menentukan persyaratan ketahanan aplikasi Anda. Arsitektur monolitik harus dihindari jika memungkinkan. Sebagai gantinya, pertimbangkan dengan cermat komponen-komponen aplikasi mana yang dapat dipecah menjadi layanan-layanan mikro. Tergantung pada persyaratan aplikasi Anda, ini mungkin berakhir menjadi kombinasi arsitektur berorientasi layanan (SOA) dengan layanan mikro jika memungkinkan. Beban kerja yang mampu berada dalam kondisi stateless akan lebih mampu di-deploy sebagai layanan mikro.

Hasil yang diinginkan: Beban kerja harus dapat didukung, dapat diskalakan, dan digabungkan secara longgar hingga selonggar mungkin.

Saat membuat pilihan tentang cara melakukan segmentasi terhadap beban kerja Anda, seimbangkan manfaat dengan kerumitannya. Hal yang tepat untuk produk baru yang mengejar jadwal peluncuran pertama akan berbeda dengan hal yang dibutuhkan oleh sebuah beban kerja yang dibangun untuk diskalakan dari awal. Saat memfaktor ulang sebuah monolit yang ada, Anda harus mempertimbangkan seberapa baik aplikasi akan mendukung dekomposisi menuju kondisi stateless. Dengan memecah layanan menjadi bagian-bagian yang lebih kecil, tim-tim kecil yang diberi tanggung jawab khusus akan dapat mengembangkan dan mengelolanya. Namun demikian, layanan yang lebih kecil dapat menimbulkan kompleksitas yang semakin tinggi, antara lain peningkatan latensi, debugging yang lebih kompleks, dan peningkatan beban operasional.

Anti-pola umum:

- Death Star layanan mikro adalah situasi saat komponen atomik menjadi sangat saling bergantung sehingga kegagalan salah satu komponen akan menghasilkan kegagalan yang jauh lebih besar, sehingga komponen ini pun menjadi kaku dan rapuh seperti monolit.

Manfaat menerapkan praktik ini:

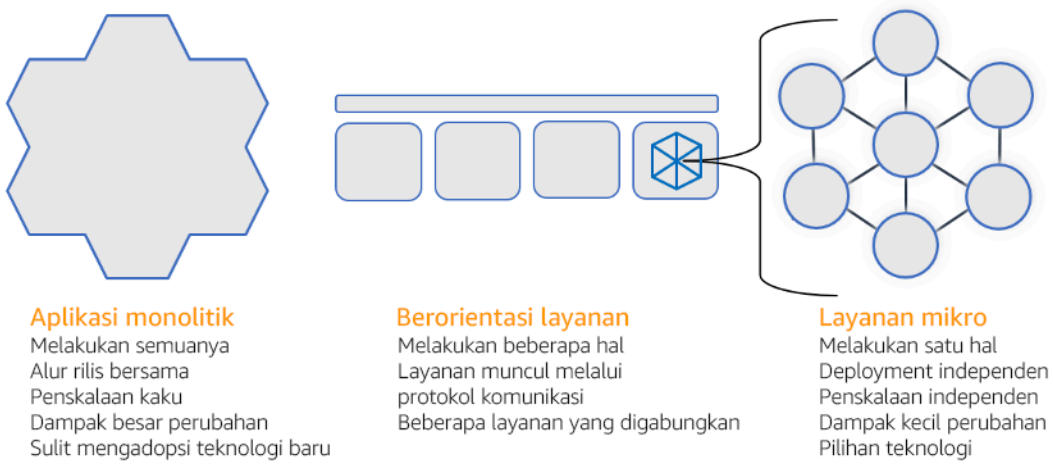
- Segmen-segmen yang lebih spesifik akan menghasilkan ketangkasan, fleksibilitas organisasi, dan skalabilitas yang lebih besar.
- Dampak gangguan layanan yang berkurang.
- Komponen-komponen aplikasi mungkin memiliki persyaratan ketersediaan yang berbeda-beda, yang dapat didukung oleh segmentasi yang lebih kecil (atomic segmentation).
- Tanggung jawab yang ditentukan dengan baik untuk tim yang mendukung beban kerja.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Tinggi

Panduan implementasi

Pilih jenis arsitektur berdasarkan cara beban kerja disegmentasikan. Pilih arsitektur SOA atau layanan mikro (atau dalam beberapa kasus yang jarang terjadi, arsitektur monolitik). Bahkan jika Anda memilih untuk memulai dengan arsitektur monolit, Anda harus memastikan bahwa itu modular dan pada akhirnya dapat berkembang ke SOA atau layanan mikro sebagai skala produk Anda dengan adopsi pengguna. SOAdan layanan mikro menawarkan segmentasi yang lebih kecil, yang lebih disukai sebagai arsitektur modern yang dapat diskalakan dan andal, tetapi ada trade-off yang perlu dipertimbangkan, terutama ketika menerapkan arsitektur layanan mikro.

Salah satu tarik ulur utama adalah Anda sekarang memiliki sebuah arsitektur komputasi terdistribusi yang dapat mempersulit dalam memenuhi kebutuhan latensi pengguna dan ada kerumitan tambahan dalam proses debugging dan penelusuran interaksi pengguna. Anda dapat menggunakan AWS X-Ray untuk membantu Anda memecahkan masalah ini. Efek lain yang perlu dipertimbangkan adalah peningkatan kompleksitas operasional seiring meningkatnya jumlah aplikasi yang Anda kelola, yang memerlukan deployment terhadap banyak komponen independensi.



Arsitektur monolitik yang berorientasi layanan, dan arsitektur layanan mikro

Langkah-langkah implementasi

- Tentukan arsitektur yang sesuai untuk memfaktor ulang atau membangun aplikasi Anda. SOAdan layanan mikro menawarkan segmentasi yang lebih kecil, yang lebih disukai sebagai arsitektur modern yang dapat diskalakan dan andal. SOAdapat menjadi kompromi yang baik untuk mencapai segmentasi yang lebih kecil sambil menghindari beberapa kompleksitas layanan mikro. Untuk detail selengkapnya, lihat Kompromi Layanan Mikro.
- Jika dapat diterima beban kerja dan didukung organisasi, maka Anda harus menggunakan sebuah arsitektur layanan mikro untuk mencapai ketangkasan dan keandalan terbaik. Untuk detail selengkapnya, lihat Menerapkan Layanan Mikro di AWS.
- Pertimbangkan untuk mengikuti pola Strangler Fig berikut untuk memfaktorkan ulang monolit menjadi komponen yang lebih kecil. Hal ini melibatkan penggantian komponen aplikasi tertentu secara bertahap dengan aplikasi dan layanan baru. AWS Migration Hub Refactor Spaces bertindak sebagai titik awal untuk memfaktor ulang secara bertahap. Untuk mendapatkan detail selengkapnya, lihat Memigrasikan beban kerja lama di on-premise dengan lancar menggunakan pola strangler.
- Menerapkan layanan mikro mungkin memerlukan mekanisme penemuan layanan untuk memungkinkan layanan terdistribusi ini berkomunikasi satu sama lain. AWS App Meshdapat digunakan dengan arsitektur berorientasi layanan untuk memberikan penemuan dan akses layanan yang andal. AWS Cloud Mapjuga dapat digunakan untuk penemuan layanan DNS berbasis dinamis.
- Jika Anda bermigrasi dari monolit ke, Amazon SOA MQ dapat membantu menjembatani kesenjangan sebagai bus layanan saat mendesain ulang aplikasi lama di cloud.
- Untuk monolit yang ada dengan satu basis data bersama, pilihlah cara mengatur ulang data menjadi segmen yang lebih kecil. Segmentasi ini dapat didasarkan pada unit bisnis, pola akses, atau struktur data. Pada titik ini dalam proses refactoring, Anda harus memilih untuk bergerak maju dengan jenis database relasional atau non-relasional (No). SQL Untuk detail selengkapnya, lihat Dari SQL ke No SQL.

Tingkat upaya untuk rencana implementasi: Tinggi

Sumber daya

Praktik-praktik terbaik terkait:

- REL03-BP02 Bangun layanan yang berfokus pada domain dan fungsionalitas bisnis khusus

Dokumen terkait:

- Amazon API Gateway: Mengonfigurasi REST API Menggunakan Buka API
- Apa itu Arsitektur Berorientasi Layanan?
- Konteks Terikat (pola sentral di Desain yang Didorong Domain)
- Menerapkan Layanan Mikro pada AWS
- Kompensasi Layanan Mikro
- Layanan mikro - definisi dari istilah arsitektur baru ini
- Microservices pada AWS
- Apa itu AWS App Mesh?

Contoh terkait:

- Lokakarya Modernisasi Aplikasi Iteratif

Video terkait:

- Memberikan Keunggulan dengan Layanan Mikro AWS

REL03-BP02 Bangun layanan yang berfokus pada domain dan fungsionalitas bisnis khusus

Arsitektur berorientasi layanan (SOA) menetapkan layanan dengan fungsi yang digambarkan dengan baik berdasarkan kebutuhan bisnis. Layanan mikro menggunakan model domain dan konteks yang dibatasi untuk menarik batas-batas layanan di sepanjang batas konteks bisnis. Berfokus pada domain dan fungsionalitas bisnis dapat membantu tim untuk menentukan persyaratan keandalan sendiri untuk layanan mereka. Konteks yang dibatasi mengisolasi dan memisahkan logika bisnis, sehingga memungkinkan tim memiliki penalaran yang lebih baik tentang bagaimana menangani kegagalan.

Hasil yang diinginkan: Para rekayasawan dan pemangku kepentingan bisnis bersama-sama menetapkan konteks yang dibatasi dan menggunakannya untuk merancang sebuah sistem sebagai layanan yang memenuhi fungsi-fungsi bisnis tertentu. Tim-tim ini menggunakan praktik-praktik yang

telah lazim seperti event storming untuk menentukan persyaratan. Aplikasi-aplikasi baru dirancang sebagai batasan-batasan layanan yang ditetapkan dengan baik dan penggabungan longgar. Monolit yang ada didekomposisi menjadi konteks terikat dan desain sistem bergerak menuju arsitektur SOA atau layanan mikro. Ketika arsitektur monolit difaktorkan ulang, pendekatan-pendekatan lazim yang ditetapkan seperti konteks gelembung dan pola penguraian monolit diterapkan.

Layanan-layanan yang berorientasi domain dijalankan sebagai satu atau beberapa proses yang statusnya tidak sama. Layanan-layanan tersebut secara independen memberikan respons terhadap fluktuasi permintaan yang terjadi dan menangani skenario kesalahan dengan berpatokan pada persyaratan khusus domain.

Anti-pola umum:

- Tim dibentuk berdasarkan domain-domain teknis tertentu seperti UI dan UX, perangkat lunak perantara (middleware), atau basis data, tidak dibentuk berdasarkan domain bisnis tertentu.
- Aplikasi melibatkan tanggung jawab domain. Layanan-layanan yang mencakup konteks yang dibatasi bisa lebih sulit untuk dipelihara, memerlukan upaya pengujian yang lebih besar, dan memerlukan banyak tim domain untuk berpartisipasi dalam pembaruan perangkat lunak.
- Dependensi domain, seperti pustaka entitas domain, dibagikan di seluruh layanan sehingga adanya perubahan terhadap satu domain layanan mengharuskan dilakukannya perubahan pada domain layanan lainnya
- Kontrak layanan dan logika bisnis tidak mengekspresikan entitas dalam bahasa domain yang umum dan konsisten, sehingga dapat menghasilkan lapisan-lapisan terjemahan yang membuat sistem semakin rumit dan upaya debugging semakin meningkat.

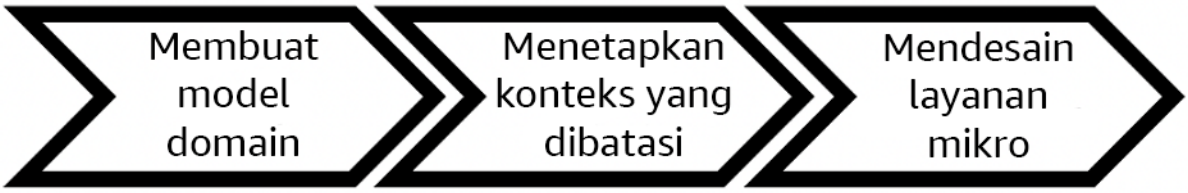
Manfaat menerapkan praktik terbaik ini: Aplikasi dirancang sebagai layanan independen yang dibatasi oleh domain bisnis dan menggunakan bahasa bisnis yang umum. Layanan-layanan dapat diuji dan dapat di-deploy secara independen. Layanan-layanan memenuhi persyaratan ketahanan khusus domain untuk domain yang diterapkan.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Tinggi

Panduan implementasi

Desain berbasis domain (DDD) adalah pendekatan dasar perancangan dan pembangunan perangkat lunak berdasarkan domain bisnis. Menggunakan kerangka kerja yang ada sekarang akan memudahkan Anda dalam membuat layanan yang berfokus pada domain bisnis. Saat

bekerja dengan aplikasi monolitik yang ada, Anda dapat memanfaatkan pola-pola penguraian yang menyediakan teknik-teknik yang sudah lazim dan ditetapkan untuk melakukan modernisasi terhadap aplikasi hingga menjadi layanan.



Desain berbasis domain

Langkah-langkah implementasi

- Tim dapat mengadakan event storming untuk mengidentifikasi peristiwa, perintah, agregat, dan domain secara cepat dalam format catatan tempel ringan.
- Setelah entitas dan fungsi domain dibentuk dalam sebuah konteks domain, Anda dapat membagi domain Anda menjadi layanan menggunakan konteks terikat, di mana entitas yang berbagi fitur dan atribut serupa dikelompokkan bersama. Dengan model yang dibagi-bagi ke dalam konteks, muncul sebuah templat untuk membatasi layanan mikro.
  - Misalnya, entitas-entitas yang dalam situs web Amazon.com dapat meliputi paket, pengantaran, jadwal, harga, diskon, dan mata uang.
  - Paket, pengantaran, dan jadwal dikelompokkan ke dalam konteks pengiriman, sedangkan harga, diskon, dan mata uang dikelompokkan ke dalam konteks harga.
- Mengurai monolit menjadi layanan mikro menguraikan pola untuk memfaktor ulang layanan mikro. Menggunakan pola-pola penguraian berdasarkan kemampuan bisnis, subdomain, atau transaksi selaras dengan pendekatan berbasis domain.
- Teknik taktikal seperti konteks gelembung akan memungkinkan Anda memasukkan DDD di dalam aplikasi yang ada atau aplikasi warisan tanpa penulisan ulang di awal dan komitmen penuh terhadap DDD. Dalam sebuah pendekatan konteks gelembung, sebuah konteks terikat kecil dibuat dengan menggunakan pemetaan dan koordinasi layanan, atau lapisan anti-korupsi, yang melindungi model domain yang baru didefinisikan dari pengaruh eksternal.

Setelah tim melakukan analisis domain dan menentukan entitas serta kontrak layanan, mereka dapat memanfaatkan layanan AWS untuk menerapkan desain berbasis domain mereka sebagai layanan berbasis cloud.

- Mulailah pengembangan Anda dengan menentukan pengujian yang menggunakan aturan-aturan bisnis domain Anda. Pengembangan berbasis pengujian (TDD) dan pengembangan berbasis perilaku (BDD) akan membantu tim dalam menjaga layanan tetap berfokus pada pemecahan masalah-masalah bisnis.
- Pilih layanan AWS yang paling sesuai dengan persyaratan domain bisnis dan arsitektur layanan mikro Anda:
  - Nirserver AWS akan memungkinkan tim Anda untuk fokus pada logika domain tertentu, bukan pada pengelolaan server dan infrastruktur.
  - Kontainer di AWS menyederhanakan pengelolaan infrastruktur Anda, sehingga Anda dapat fokus pada persyaratan domain Anda.
  - Basis data yang dibuat khusus dapat membantu Anda mencocokkan persyaratan domain Anda dengan jenis basis data yang paling sesuai.
- Membangun arsitektur heksagonal di AWS menguraikan kerangka kerja untuk membangun logika bisnis menjadi layanan yang bekerja mundur dari domain bisnis untuk memenuhi persyaratan fungsional dan kemudian melampirkan adaptor integrasi. Pola-pola yang memisahkan detail antarmuka dari logika bisnis dengan layanan-layanan AWS akan membantu tim untuk berfokus pada fungsionalitas domain dan meningkatkan kualitas perangkat lunak.

Sumber daya

Praktik-praktik terbaik terkait:

- REL03-BP01 Pilih cara mengelompokkan beban kerja Anda
- REL03-BP03 Memberikan kontrak layanan per API

Dokumen terkait:

- Layanan mikro AWS
- Mengimplementasikan Layanan Mikro di AWS
- Cara memecah Monolit menjadi Layanan-layanan Mikro
- Mulai Menggunakan DDD di Tengah-Tengah Sistem Warisan
- Desain Berbasis Domain: Mengatasi Kompleksitas di Dalam Inti Perangkat Lunak
- Membangun arsitektur heksagonal di AWS
- Menguraikan monolit menjadi layanan mikro

- Event Storming
- Pesan Antara Konteks-Konteks Terikat
- Layanan mikro
- Pengembangan berbasis pengujian
- Pengembangan berbasis perilaku

Contoh terkait:

- Merancang Layanan Mikro Cloud-Native di AWS (dari DDD/EventStormingWorkshop)

Alat terkait:

- Basis Data AWS Cloud
- Nirserver di AWS
- Kontainer di AWS

REL03-BP03 Memberikan kontrak layanan per API

Kontrak layanan adalah perjanjian terdokumentasi antara API produsen dan konsumen yang didefinisikan dalam definisi yang dapat dibaca mesinAPI. Strategi pembuatan versi kontrak memungkinkan konsumen untuk terus menggunakan yang ada API dan memigrasikan aplikasi mereka ke yang lebih baru API ketika mereka siap. Deployment oleh produsen dapat terjadi kapan saja, selama kontrak dipatuhi. Tim layanan dapat menggunakan tumpukan teknologi pilihan mereka untuk memenuhi API kontrak.

Hasil yang diinginkan: Aplikasi yang dibangun dengan arsitektur berorientasi layanan atau layanan mikro dapat beroperasi secara independen sambil memiliki ketergantungan runtime terintegrasi. Perubahan yang diterapkan ke API konsumen atau produsen tidak mengganggu stabilitas sistem secara keseluruhan ketika kedua belah pihak mengikuti kontrak bersamaAPI. Komponen yang berkomunikasi melalui layanan APIs dapat melakukan rilis fungsional independen, peningkatan ke dependensi runtime, atau gagal ke situs pemulihan bencana (DR) dengan sedikit atau tanpa dampak satu sama lain. Selain itu, layanan-layanan diskret dapat menyesuaikan skala secara independen dengan menyerap permintaan sumber daya tanpa mengharuskan layanan lain untuk menyesuaikan skala (menskalakan) secara serempak.

Anti-pola umum:

- Membuat layanan APIs tanpa skema yang diketik dengan kuat. Hal ini mengakibatkan hal APIs itu tidak dapat digunakan untuk menghasilkan API binding dan payload yang tidak dapat divalidasi secara terprogram.
- Tidak mengadopsi strategi pembuatan versi, yang memaksa API konsumen untuk memperbarui dan merilis atau gagal ketika kontrak layanan berkembang.
- Pesan-pesan kesalahan yang membocorkan detail implementasi layanan yang mendasari, bukan menggambarkan kegagalan integrasi dalam bahasa dan konteks domain.
- Tidak menggunakan API kontrak untuk mengembangkan kasus uji dan API implementasi tiruan untuk memungkinkan pengujian independen komponen layanan.

Manfaat membangun praktik terbaik ini: Sistem terdistribusi yang terdiri dari komponen yang berkomunikasi melalui kontrak API layanan dapat meningkatkan keandalan. Pengembang dapat menangkap potensi masalah di awal proses pengembangan dengan pemeriksaan tipe selama kompilasi untuk memverifikasi bahwa permintaan dan tanggapan mengikuti API kontrak dan bidang wajib ada. APIkontrak menyediakan antarmuka pendokumentasian diri yang jelas untuk APIs dan penyedia interoperabilitas yang lebih baik antara sistem yang berbeda dan bahasa pemrograman.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Sedang

Panduan implementasi

Setelah Anda mengidentifikasi domain bisnis dan menentukan segmentasi beban kerja Anda, Anda dapat mengembangkan layanan Anda. APIs Pertama, tentukan kontrak layanan yang dapat dibaca mesin untukAPIs, dan kemudian terapkan strategi pembuatan versi. API Ketika Anda siap untuk mengintegrasikan layanan melalui protokol umum seperti, REST GraphQL, atau peristiwa asinkron, Anda dapat menggabungkan AWS layanan ke dalam arsitektur Anda untuk mengintegrasikan komponen Anda dengan kontrak yang diketik dengan kuat. API

AWS layanan untuk API kontras layanan

Gabungkan AWS layanan termasuk Amazon API Gateway AWS AppSync, dan Amazon EventBridge ke dalam arsitektur Anda untuk menggunakan kontrak API layanan dalam aplikasi Anda. Amazon API Gateway membantu Anda berintegrasi dengan AWS layanan asli langsung dan layanan web lainnya. APIGateway mendukung APIspesifikasi dan pembuatan versi Terbuka. AWS AppSync adalah titik akhir GraphQL terkelola yang Anda konfigurasikan dengan mendefinisikan skema GraphQL untuk menentukan antarmuka layanan untuk kueri, mutasi, dan langganan. Amazon EventBridge menggunakan skema acara untuk menentukan peristiwa dan menghasilkan binding kode untuk acara Anda.

Langkah-langkah implementasi

- Pertama, tentukan kontrak untuk AndaAPI. Kontrak akan mengekspresikan kemampuan dan API juga mendefinisikan objek dan bidang data yang diketik dengan kuat untuk API input dan output.
- Saat Anda mengonfigurasi APIs di API Gateway, Anda dapat mengimpor dan mengekspor API Spesifikasi Terbuka untuk titik akhir Anda.
  - Mengimpor API definisi Terbuka menyederhanakan pembuatan Anda API dan dapat diintegrasikan dengan AWS infrastruktur sebagai alat kode seperti dan. AWS Serverless Application ModelAWS Cloud Development Kit (AWS CDK)
  - Mengekspor API definisi menyederhanakan integrasi dengan alat API pengujian dan memberikan spesifikasi integrasi kepada konsumen layanan.
- Anda dapat menentukan dan mengelola APIs GraphQL AWS AppSync dengan mendefinisikan file skema GraphQL untuk menghasilkan antarmuka kontrak Anda dan menyederhanakan interaksi dengan model kompleks, beberapa tabel database, atau layanan lama. REST
- AWS Amplifyproyek yang terintegrasi dengan AWS AppSync menghasilkan file JavaScript kueri yang diketik kuat untuk digunakan dalam aplikasi Anda serta pustaka klien AWS AppSync GraphQL untuk tabel Amazon DynamoDB.
- Saat Anda menggunakan peristiwa layanan dari Amazon EventBridge, peristiwa mematuhi skema yang sudah ada di registri skema atau yang Anda tentukan dengan Spesifikasi TerbukaAPI. Dengan sebuah skema yang ditentukan dalam registri tersebut, Anda juga dapat menghasilkan pengikatan klien (client binding) dari kontrak skema tersebut untuk mengintegrasikan kode Anda dengan peristiwa.
- Memperluas atau versi AndaAPI. Memperluas API adalah opsi yang lebih sederhana saat menambahkan bidang yang dapat dikonfigurasi dengan bidang opsional atau nilai default untuk bidang wajib.
  - JSONkontrak berbasis untuk protokol seperti dan REST GraphQL dapat menjadi cocok untuk perpanjangan kontrak.
  - XMLKontrak berbasis untuk protokol seperti SOAP harus diuji dengan konsumen jasa untuk menentukan kelayakan perpanjangan kontrak.
- Saat membuat versiAPI, pertimbangkan untuk menerapkan versi proxy di mana fasad digunakan untuk mendukung versi sehingga logika dapat dipertahankan dalam satu basis kode.
  - Dengan API Gateway Anda dapat menggunakan pemetaan permintaan dan respons untuk menyederhanakan penyerapan perubahan kontrak dengan membuat fasad untuk memberikan nilai default untuk bidang baru atau untuk menghapus bidang yang dihapus dari permintaan atau

respons. Dengan pendekatan ini, layanan-layanan yang mendasari dapat mempertahankan satu basis kode tunggal.

Sumber daya

Praktik-praktik terbaik terkait:

- REL03-BP01 Pilih cara mengelompokkan beban kerja Anda
- REL03-BP02 Bangun layanan yang berfokus pada domain dan fungsionalitas bisnis khusus
- REL04-BP02 Menerapkan dependensi yang digabungkan secara longgar
- REL05-BP03 Kontrol dan batasi panggilan coba lagi
- REL05-BP05 Mengatur batas waktu klien

Dokumen terkait:

- Apa Itu API (Antarmuka Pemrograman Aplikasi)?
- Menerapkan Layanan Mikro pada AWS
- Kompensasi Layanan Mikro
- Layanan mikro - definisi dari istilah arsitektur baru ini
- Microservices pada AWS
- Bekerja dengan ekstensi API Gateway untuk Buka API
- Terbuka API -Spesifikasi
- GraphQL: Skema dan Jenis
- EventBridge Ikatan kode Amazon

Contoh terkait:

- Amazon API Gateway: Mengonfigurasi REST API Menggunakan Buka API
- Amazon API Gateway ke aplikasi Amazon CRUD DynamoDB menggunakan Open API
- Pola integrasi aplikasi modern di era tanpa server: Integrasi Layanan API Gateway
- Menerapkan versi API Gateway berbasis header dengan Amazon CloudFront
- AWS AppSync: Membangun aplikasi klien

Video terkait:

- Menggunakan Open API in AWS SAM untuk mengelola API Gateway

Alat terkait:

- APIGerbang Amazon
- AWS AppSync
- Amazon EventBridge

Merancang interaksi di dalam sistem terdistribusi untuk mencegah kegagalan

Sistem terdistribusi mengandalkan jaringan komunikasi untuk membuat interkoneksi komponen, seperti server atau layanan. Beban kerja Anda harus beroperasi secara andal terlepas latensi atau hilangnya data yang terjadi di jaringan-jaringan ini. Komponen dari sistem terdistribusi harus beroperasi dengan cara yang tidak secara negatif memengaruhi beban kerja atau komponen-komponen lain. Berbagai praktik terbaik ini mencegah kegagalan dan meningkatkan waktu rata-rata antara kegagalan (MTBF).

Praktik terbaik

- REL04-BP01 Mengidentifikasi jenis sistem terdistribusi yang Anda perlukan
- REL04-BP02 Menerapkan dependensi yang digabungkan secara longgar
- REL04-BP03 Lakukan pekerjaan konstan
- REL04-BP04 Buat operasi yang bermutasi menjadi idempoten

REL04-BP01 Mengidentifikasi jenis sistem terdistribusi yang Anda perlukan

Sistem terdistribusi bisa bersifat sinkron, asinkron, atau batch. Sistem selaras harus memproses permintaan secepat mungkin dan berkomunikasi satu sama lain dengan membuat panggilan permintaan dan respons yang selaras dengan menggunakan protokol HTTP/S, REST, atau protokol panggilan prosedur jarak jauh (RPC). Sistem tidak selaras berkomunikasi satu sama lain dengan bertukar data secara asinkron melalui sebuah layanan perantara tanpa melakukan penggabungan (coupling) terhadap masing-masing sistem. Sistem batch menerima data input dalam jumlah besar,

menjalankan proses-proses data otomatis tanpa campur tangan manusia, dan menghasilkan data output.

Hasil yang diinginkan: Merancang desain sebuah beban kerja yang berinteraksi secara efektif dengan dependensi selaras, tidak selaras, dan batch.

Anti-pola umum:

- Beban kerja menunggu respons dari dependensinya tanpa batas waktu, yang dapat menyebabkan klien beban kerja mengalami habis waktu, tanpa mengetahui apakah permintaannya telah diterima atau tidak.
- Beban kerja menggunakan sebuah rantai sistem dependen yang memanggil satu sama lain dengan selaras. Hal ini mengharuskan setiap sistem untuk tersedia dan berhasil memproses sebuah permintaan sebelum seluruh rantai sistem dapat berhasil, sehingga menyebabkan perilaku dan ketersediaan secara keseluruhan menjadi rapuh.
- Beban kerja berkomunikasi dengan dependensinya secara tak selaras dan mengandalkan konsep pengiriman pesan yang dijamin persis satu kali, padahal sering kali pesan duplikat masih memungkinkan untuk diterima.
- Beban kerja tidak menggunakan alat-alat penjadwalan batch yang sesuai dan memungkinkan pelaksanaan pekerjaan batch yang sama secara bersamaan.

Manfaat menerapkan praktik terbaik ini: Sudah menjadi hal yang umum untuk beban kerja tertentu untuk menerapkan satu atau beberapa gaya komunikasi antara selaras, tak selaras, dan batch. Praktik terbaik ini akan membantu Anda untuk mengidentifikasi kompromi-kompromi berbeda yang terkait dengan setiap gaya komunikasi untuk membuat beban kerja Anda dapat memberikan toleransi terhadap gangguan pada dependensinya.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Tinggi

Panduan implementasi

Bagian-bagian berikutnya berisi panduan implementasi umum dan spesifik untuk masing-masing jenis dependensi.

Panduan umum

- Pastikan bahwa tujuan tingkat layanan (SLO) kinerja dan keandalan yang ditawarkan oleh dependensi Anda memenuhi persyaratan performa dan keandalan beban kerja Anda.

- Gunakan layanan observabilitas AWS untuk memantau waktu respons dan tingkat kesalahan untuk memastikan bahwa dependensi Anda sedang menyediakan layanan pada tingkat yang dibutuhkan oleh beban kerja Anda.
- Identifikasi potensi-potensi tantangan yang mungkin dihadapi beban kerja Anda saat berkomunikasi dengan dependensinya. Sistem terdistribusi datang dengan berbagai tantangan yang dapat meningkatkan kompleksitas arsitektur, beban operasional, dan biaya. Tantangan-tantangan yang biasanya dihadapi mencakup latensi, gangguan jaringan, kehilangan data, penskalaan, dan jeda replikasi data.
- Terapkan penanganan dan pencatatan log kesalahan yang kuat untuk membantu Anda memecahkan masalah saat dependensi Anda mengalami masalah.

Dependensi sinkron

Dalam komunikasi yang selaras, beban kerja Anda mengirimkan permintaan ke dependensinya dan memblokir operasi yang menunggu sebuah respons. Ketika dependensinya menerima permintaan, dependensi tersebut akan mencoba menanganinya sesegera mungkin dan mengembalikan respons ke beban kerja Anda. Tantangan besar pada komunikasi selaras adalah jenis komunikasi ini menyebabkan terjadinya penggabungan temporal, yang mengharuskan beban kerja Anda dan dependensinya untuk tersedia pada saat yang bersamaan. Ketika beban kerja Anda perlu berkomunikasi secara selaras dengan dependensinya, pertimbangkan panduan berikut:

- Beban kerja Anda seharusnya tidak bergantung pada beberapa dependensi selaras untuk melakukan satu fungsi tunggal. Rangkaian dependensi ini akan meningkatkan kerapuhan secara keseluruhan karena semua dependensi di jalurnya harus tersedia agar permintaan berhasil diselesaikan.
- Ketika sebuah dependensi berada dalam kondisi tidak sehat atau tidak tersedia, tentukan penanganan kesalahan dan strategi coba lagi. Hindari penggunaan perilaku bimodal. Perilaku bimodal adalah ketika beban kerja Anda menunjukkan perilaku yang berbeda dalam mode normal dan mode kegagalan. Untuk mendapatkan detail selengkapnya tentang perilaku bimodal, silakan lihat REL11-BP05 Menggunakan stabilitas statis untuk mencegah perilaku bimodal.
- Harap diingat bahwa gagal cepat (fail fast) lebih baik daripada membuat beban kerja Anda menunggu. Misalnya, Panduan Pengembang AWS Lambda menjelaskan cara menangani percobaan ulang dan kegagalan saat Anda menginvokasi fungsi Lambda.
- Tetapkan batas waktu saat beban kerja Anda memanggil dependensinya. Teknik ini menghindari waktu tunggu yang terlalu lama atau waktu tunggu respons tanpa batas. Untuk melihat

pembahasan yang bermanfaat mengenai topik ini, silakan lihat Menyetel pengaturan permintaan AWS Java SDK HTTP untuk aplikasi Amazon DynamoDB yang sadar latensi.

- Minimalkan jumlah panggilan yang dilakukan dari beban kerja Anda ke dependensinya untuk memenuhi satu permintaan tunggal. Memiliki banyak panggilan (chatty) di antara keduanya dapat meningkatkan penggabungan dan latensi.

Dependensi tidak selaras

Untuk memisahkan beban kerja Anda dari dependensinya secara sementara, keduanya harus berkomunikasi secara tidak selaras. Jika menggunakan pendekatan tidak selaras, beban kerja Anda dapat melanjutkan pemrosesan lain tanpa harus menunggu dependensinya, atau rangkaian dependensinya, untuk mengirimkan sebuah respons.

Ketika beban kerja Anda perlu berkomunikasi secara tidak selaras dengan dependensinya, pertimbangkan panduan berikut:

- Tentukan apakah akan menggunakan perpesanan atau streaming peristiwa berdasarkan kasus penggunaan dan kebutuhan Anda. Layanan perpesanan akan memungkinkan beban kerja Anda untuk berkomunikasi dengan dependensinya dengan mengirim dan menerima pesan melalui broker pesan. Streaming acara akan memungkinkan beban kerja Anda dan dependensinya untuk menggunakan layanan streaming untuk mempublikasikan dan berlangganan acara, disampaikan sebagai aliran data berkelanjutan, yang perlu diproses sesegera mungkin.

- Perpesanan dan streaming peristiwa menangani pesan secara berbeda sehingga Anda perlu mengambil keputusan-keputusan kompromi berdasarkan:
  - Prioritas pesan: broker pesan dapat memproses pesan prioritas tinggi lebih dahulu dari pesan normal. Dalam streaming peristiwa, semua pesan memiliki prioritas yang sama.
  - Konsumsi pesan: broker pesan memastikan bahwa konsumen menerima pesan. Pemakai streaming peristiwa harus terus melacak pesan terakhir yang telah mereka baca.
  - Pengurutan pesan: dengan pesan, Anda akan menerima pesan dalam urutan yang tepat yang dikirim dan tidak dijamin kecuali Anda menggunakan pendekatan first-in-first-out (FIFO). Streaming peristiwa selalu akan mempertahankan urutan sesuai urutan data ketika dibuat.
  - Penghapusan pesan: dengan layanan pengiriman pesan, konsumen harus menghapus pesan setelah memprosesnya. Layanan streaming peristiwa menambahkan pesan tersebut ke sebuah aliran dan tetap berada di sana sampai periode penyimpanan pesan berakhir. Kebijakan penghapusan ini menjadikan streaming peristiwa cocok untuk pemutaran ulang pesan.

- Tentukan bagaimana beban kerja Anda mengetahui kapan dependensinya menyelesaikan pekerjaan. Sebagai contoh, saat beban kerja Anda menginvokasi fungsi Lambda asinkron, Lambda akan menempatkan peristiwa dalam antrean dan mengembalikan respons sukses tanpa menyertakan informasi tambahan. Setelah pemrosesan selesai, fungsi Lambda dapat mengirim hasilnya ke sebuah tujuan, dapat dikonfigurasi berdasarkan keberhasilan atau kegagalan.
- Bangun beban kerja Anda untuk menangani pesan duplikat dengan memanfaatkan idempotensi. Idempotensi adalah ketika hasil beban kerja Anda tidak berubah bahkan jika beban kerja Anda dihasilkan lebih dari sekali untuk pesan yang sama. Penting untuk menunjukkan bahwa layanan pengiriman pesan atau streaming akan mengirim ulang pesan jika terjadi kegagalan jaringan atau jika ada pengakuan bahwa pesan belum diterima.
- Jika beban kerja Anda tidak mendapatkan respons dari dependensinya, maka permintaan perlu dikirimkan kembali. Pertimbangkan untuk membatasi jumlah percobaan ulang untuk menghemat sumber daya CPU, memori, dan jaringan beban kerja Anda untuk menangani premintaan-permintaan lainnya. Dokumentasi AWS Lambda menunjukkan cara menangani kesalahan untuk panggilan asinkron.
- Manfaatkan alat-alat observabilitas, debugging, dan pelacakan yang sesuai untuk mengelola dan mengoperasikan komunikasi tidak selaras dari beban kerja Anda dengan dependensinya. Anda dapat menggunakan Amazon CloudWatch untuk memantau layanan perpesanan dan streaming acara. Anda juga dapat menginstrumentasikan beban kerja Anda dengan AWS X-Ray untuk dengan cepat mendapatkan wawasan untuk pemecahan masalah.

Dependensi Batch

Sistem batch mengambil data input, memulai serangkaian pekerjaan untuk memprosesnya, dan kemudian menghasilkan beberapa data output, tanpa intervensi manual. Tergantung ukuran data, pekerjaan dapat berjalan dari hitungan menit hingga, dalam beberapa kasus, beberapa hari. Ketika beban kerja Anda berkomunikasi dengan dependensinya, pertimbangkan untuk menggunakan panduan berikut:

- Tentukan rentang periode ketika beban kerja Anda harus menjalankan tugas batch. Beban kerja Anda dapat mengatur pola pengulangan untuk menginvokasi sebuah sistem batch, misalnya, setiap jam atau pada akhir setiap bulan.
- Tentukan lokasi input data dan output data yang diproses. Pilih sebuah layanan penyimpanan, seperti Amazon Simple Storage Service (Amazon S3), Amazon Elastic File System (Amazon EFS), dan Amazon FSx for Lustre, yang memungkinkan beban kerja Anda untuk membaca dan menulis file dalam skala besar.

- Jika beban kerja Anda perlu menginvokasi beberapa pekerjaan batch, Anda dapat memanfaatkan AWS Step Functions untuk menyederhanakan orkestrasi pekerjaan batch yang berjalan di AWS atau on-premise. Proyek sampel ini mendemonstrasikan orkestrasi pekerjaan batch dengan menggunakan Step Functions, AWS Batch, dan Lambda.
- Lakukan pemantauan terhadap pekerjaan batch untuk mencari kelainan, seperti pekerjaan yang membutuhkan waktu lebih lama dari yang seharusnya. Anda dapat menggunakan alat-alat seperti Wawasan Kontainer CloudWatch untuk memantau lingkungan dan pekerjaan AWS Batch. Dalam keadaan ini, beban kerja Anda akan menghentikan pekerjaan berikutnya dari awal dan memberitahukan pengecualian kepada staf yang relevan.

Sumber daya

Dokumen terkait:

- Operasi AWS Cloud: Pemantauan dan Observabilitas
- Amazon Builders' Library: Tantangan dengan sistem terdistribusi
- REL11-BP05 Menggunakan stabilitas statis untuk mencegah perilaku bimodal
- Panduan Pengembang AWS Lambda: Penanganan kesalahan dan percobaan ulang otomatis di AWS Lambda
- Menyetel pengaturan permintaan AWS Java SDK HTTP untuk aplikasi Amazon DynamoDB yang sadar latensi
- Perpesanan AWS
- Apa itu data streaming?
- Panduan Pengembang AWS Lambda: Panggilan asinkron
- Pertanyaan Umum tentang Amazon Simple Queue Service: Antrean FIFO
- Panduan Pengembang Amazon Kinesis Data Streams: Menangani Rekaman Duplikat
- Panduan Pengembang Amazon Simple Queue Service: Metrik CloudWatch yang Tersedia untuk Amazon SQS
- Panduan Pengembang Amazon Kinesis Data Streams: Memantau Layanan Amazon Kinesis Data Streams dengan Amazon CloudWatch
- Panduan Pengembang AWS X-Ray: Konsep AWS X-Ray
- Sampel AWS di GitHub: AWS Step menggunakan Aplikasi Orkestrator Kompleks
- Panduan Pengguna AWS Batch: Wawasan Kontainer CloudWatch AWS Batch

Video terkait:

- AWS Summit SF 2022 - Observabilitas tumpukan penuh (full-stack) dan pemantauan aplikasi dengan AWS (COP310)

Alat terkait:

- Amazon CloudWatch
- Log Amazon CloudWatch
- AWS X-Ray
- Amazon Simple Storage Service (Amazon S3)
- Amazon Elastic File System (Amazon EFS)
- Amazon FSx for Lustre
- AWS Step Functions
- AWS Batch

REL04-BP02 Menerapkan dependensi yang digabungkan secara longgar

Dependensi seperti sistem pengantrean, sistem streaming, alur kerja, dan penyeimbang beban digabungkan secara longgar. Penggabungan longgar membantu memisahkan perilaku suatu komponen dari komponen lainnya yang bergantung pada komponen tersebut, sehingga meningkatkan ketahanan dan ketangkasan.

Memisahkan dependensi, seperti sistem antrean, sistem streaming, dan alur kerja, membantu meminimalkan dampak perubahan atau kegagalan pada suatu sistem. Pemisahan ini akan mengisolasi perilaku komponen dari mempengaruhi orang lain yang bergantung padanya, meningkatkan ketahanan dan kelincahan.

Dalam sistem penggabungan erat (tightly coupled), perubahan pada satu komponen dapat menyebabkan perubahan pada komponen lain yang bergantung padanya, yang mengakibatkan penurunan performa di semua komponen. Penggabungan longgar menghilangkan dependensi ini, sehingga komponen-komponen yang bergantung hanya perlu mengetahui antarmuka versi terbaru dan yang dipublikasikan. Mengimplementasikan penggabungan longgar antar dependensi akan memisahkan kegagalan pada salah satu dependensi agar tidak memengaruhi dependensi yang lain.

Penggabungan longgar akan memungkinkan Anda untuk mengubah kode atau menambahkan fitur ke sebuah komponen sekaligus meminimalkan risiko pada komponen lain yang bergantung pada

komponen tersebut. Hal ini juga memungkinkan ketahanan hingga tingkatan terkecil (granular) pada tingkat komponen sehingga Anda dapat menambahkan skala (scale-out) atau bahkan mengubah implementasi yang mendasari dependensi.

Agar makin meningkatkan ketahanan melalui penggabungan longgar, buatlah interaksi-interaksi komponen tak selaras, jika memungkinkan. Model ini cocok untuk interaksi apa pun yang tidak memerlukan respons langsung dan di mana pengakuan bahwa permintaan telah terdaftar sudah dianggap cukup. Ini melibatkan satu komponen yang menghasilkan peristiwa dan komponen-komponen lain yang menggunakannya. Kedua komponen tidak terintegrasi melalui point-to-point interaksi langsung tetapi biasanya melalui lapisan penyimpanan tahan lama menengah, seperti SQS antrian Amazon, platform data streaming seperti Amazon Kinesis, atau. AWS Step Functions

Gambar 4: Dependensi seperti sistem pengantrean dan penyeimbang beban digabungkan dengan longgar

Amazon SQS mengantri dan hanya AWS Step Functions dua cara untuk menambahkan lapisan perantara untuk kopling longgar. Arsitektur berbasis peristiwa juga dapat dibangun menggunakan AWS Cloud Amazon EventBridge, yang dapat mengabstraksi klien (produsen acara) dari layanan yang mereka andalkan (konsumen acara). Amazon Simple Notification Service (AmazonSNS) adalah solusi efektif ketika Anda membutuhkan throughput tinggi, berbasis push, perpesanan. many-to-many Menggunakan SNS topik Amazon, sistem penerbit Anda dapat menyebarkan pesan ke sejumlah besar titik akhir pelanggan untuk pemrosesan paralel.

Meskipun antrean menawarkan sejumlah manfaat, di sebagian besar sistem waktu nyata yang keras, permintaan yang lebih lama dari waktu ambang batas (sering kali dalam hitungan detik) harus dianggap basi (klien telah menyerah dan sudah tidak menunggu respons lagi), dan tidak diproses. Dengan begitu, permintaan yang lebih baru (dan kemungkinan masih valid) dapat diproses sebagai gantinya.

Hasil yang diinginkan: Menerapkan dependensi yang digabungkan dengan longgar akan memungkinkan Anda untuk meminimalkan peluang kegagalan ke tingkat komponen, dan akan membantu Anda untuk mendiagnosis dan menyelesaikan masalah. Cara ini juga dapat menyederhanakan siklus pengembangan, sehingga memungkinkan tim untuk menerapkan perubahan-perubahan pada tingkat modular tanpa memengaruhi performa komponen-komponen lain yang bergantung padanya. Pendekatan ini memberikan kemampuan untuk menambahkan skala (scale-out) pada tingkat komponen berdasarkan kebutuhan sumber daya, serta pemanfaatan komponen yang berkontribusi terhadap efektivitas biaya.

Anti-pola umum:

- Melakukan deployment beban kerja monolitik.
- Langsung memanggil APIs antara tingkatan beban kerja tanpa kemampuan failover atau pemrosesan permintaan asinkron.
- Penggabungan erat dengan menggunakan data bersama. Sistem-sistem yang digabungkan dengan longgar sebaiknya tidak berbagi data melalui basis data bersama atau bentuk penyimpanan data yang digabungkan secara erat, yang dapat menimbulkan kembali penggabungan erat dan akan menghambat skalabilitas.
- Mengabaikan tekanan balik. Beban kerja Anda harus memiliki kemampuan untuk memperlambat atau menghentikan data yang masuk ketika sebuah komponen tidak dapat memprosesnya dengan kecepatan yang sama.

Manfaat menerapkan praktik terbaik ini: Penggabungan longgar akan membantu Anda untuk memisahkan perilaku suatu komponen dari komponen lainnya yang bergantung pada komponen tersebut, sehingga akan meningkatkan ketahanan dan ketangkasan. Kegagalan di salah satu komponen dipisahkan dari komponen lain.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Tinggi

Panduan implementasi

Implementasikan dependensi yang digabungkan dengan longgar. Ada berbagai solusi yang memungkinkan Anda untuk membangun aplikasi yang digabungkan dengan longgar. Ini termasuk layanan untuk menerapkan antrian yang dikelola sepenuhnya, alur kerja otomatis, bereaksi terhadap peristiwa, dan APIs antara lain yang dapat membantu mengisolasi perilaku komponen dari komponen lain, dan dengan demikian meningkatkan ketahanan dan kelincahan.

- Bangun arsitektur berbasis peristiwa: Amazon EventBridge membantu Anda membangun arsitektur berbasis peristiwa yang digabungkan dan didistribusikan secara longgar.
- Menerapkan antrian dalam sistem terdistribusi: Anda dapat menggunakan Amazon Simple Queue Service (AmazonSQS) untuk mengintegrasikan dan memisahkan sistem terdistribusi.
- Kontainerisasi komponen sebagai layanan mikro: Layanan mikro memungkinkan tim untuk membangun aplikasi yang terdiri dari komponen independen kecil yang berkomunikasi melalui definisi yang terdefinisi dengan baik. APIs Amazon Elastic Container Service (AmazonECS), dan Amazon Elastic Kubernetes Service (EKSAmazon) dapat membantu Anda memulai lebih cepat dengan kontainer.

- Kelola alur kerja dengan Step Functions: Step Functions membantu Anda mengoordinasikan beberapa AWS layanan ke dalam alur kerja yang fleksibel.
- Manfaatkan arsitektur perpesanan berlangganan publikasi (pub/sub): Amazon Simple Notification Service (AmazonSNS) menyediakan pengiriman pesan dari penerbit ke pelanggan (juga dikenal sebagai produsen dan konsumen).

Langkah-langkah implementasi

- Komponen dalam sebuah arsitektur yang didorong peristiwa dimulai oleh peristiwa. Peristiwa adalah tindakan-tindakan yang terjadi dalam sebuah sistem, seperti pengguna menambahkan item ke keranjang. Ketika suatu tindakan berhasil, sebuah peristiwa dihasilkan, yang akan menggerakkan komponen berikutnya dalam sistem tersebut.
  - Membangun Aplikasi Event Driven dengan Amazon EventBridge
  - AWS re:invent 2022 - Merancang Integrasi Berbasis Acara menggunakan Amazon EventBridge
- Sistem olah pesan terdistribusi memiliki tiga bagian utama yang perlu diimplementasikan untuk sebuah arsitektur berbasis antrean. Mereka termasuk komponen sistem terdistribusi, antrian yang digunakan untuk decoupling (didistribusikan di SQS server Amazon), dan pesan dalam antrian. Sistem seperti ini biasanya memiliki produsen yang memulai pesan ke dalam antrean, dan konsumen yang menerima pesan dari antrean tersebut. Antrian menyimpan pesan di beberapa SQS server Amazon untuk redundansi.
  - SQSArsitektur Amazon dasar
  - Kirim Pesan Antara Aplikasi Terdistribusi dengan Amazon Simple Queue Service
- Layanan mikro, jika dimanfaatkan dengan baik, akan meningkatkan pemeliharaan dan mendongkrak skalabilitas, karena komponen-komponen yang digabungkan dengan longgar dikelola oleh tim independen. Hal ini juga akan memungkinkan isolasi perilaku ke satu komponen jika terjadi perubahan.
  - Menerapkan Layanan Mikro pada AWS
  - Mari Merancang! Merancang arsitektur layanan mikro dengan kontainer
- Dengan AWS Step Functions Anda dapat membangun aplikasi terdistribusi, mengotomatiskan proses, mengatur layanan mikro, antara lain. Melakukan orkestrasi beberapa komponen ke dalam sebuah alur kerja otomatis akan memungkinkan Anda untuk memisahkan dependensi dalam aplikasi Anda.
  - Buat Alur Kerja Tanpa Server dengan dan AWS Step FunctionsAWS Lambda
  - Memulai dengan AWS Step Functions

Sumber daya

Dokumen terkait:

- AmazonEC2: Memastikan Idempotensi
- Amazon Builders' Library: Tantangan dengan sistem terdistribusi
- Amazon Builders' Library: Keandalan, kerja konstan, dan pilihan yang tepat
- Apa itu Amazon EventBridge?
- Apa Itu Amazon Simple Queue Service?
- Putus dengan monolit Anda
- Mengatur Layanan Mikro Berbasis Antrian dengan dan Amazon AWS Step Functions SQS
- SQSArsitektur Amazon dasar
- Arsitektur Berbasis Antrian

Video terkait:

- AWS KTT New York 2019: Pengantar Arsitektur Berbasis Acara dan Amazon EventBridge (05) MAD2
- AWS RE: Invent 2018: Tutup Loop dan Membuka Pikiran: Cara Mengendalikan Sistem, Besar dan Kecil ARC337 (termasuk kopling longgar, kerja konstan, stabilitas statis)
- AWS re:invent 2019: Pindah ke arsitektur berbasis acara (08) SVS3
- AWS re:invent 2019: Aplikasi berbasis peristiwa tanpa server yang dapat diskalakan menggunakan Amazon dan Lambda SQS
- AWS re:invent 2022 - Merancang integrasi berbasis acara menggunakan Amazon EventBridge
- AWS RE: Invent 2017: Elastic Load Balancing Deep Dive dan Praktik Terbaik

REL04-BP03 Lakukan pekerjaan konstan

Sistem dapat gagal mengalami kegagalan saat ada perubahan besar dan cepat pada beban. Misalnya, jika beban kerja Anda sedang melakukan pemeriksaan kondisi yang memantau kondisi dari ribuan server, beban kerja Anda harus mengirimkan payload berukuran sama (snapshot penuh berisi status saat ini) setiap saat. Saat tidak ada server yang gagal, atau semuanya gagal, sistem pemeriksaan kondisi melakukan tugas konstan tanpa perubahan besar dan cepat.

Misalnya, jika sistem pemeriksaan kondisi sedang memantau 100.000 server, dengan tingkat kegagalan server normal yang ringan, maka beban yang ditanggung kecil. Namun demikian, jika ada sebuah peristiwa besar yang membuat separuh server menjadi tidak sehat, maka sistem pemeriksaan kondisi akan kewalahan untuk memperbarui sistem notifikasi dan menyampaikan status ke kliennya. Jadi alih-alih sistem pemeriksaan kondisi harus mengirim snapshot lengkap dari keadaan saat ini setiap kali. 100.000 status kesehatan server, masing-masing diwakili oleh satu bit, dan itu hanya akan menjadi muatan sebesar 12,5 KB. Saat tidak ada server yang gagal, atau semuanya gagal, sistem pemeriksaan kondisi akan melakukan tugas konstan, dan perubahan yang besar dan cepat bukanlah ancaman untuk stabilitas sistem. Seperti inilah Amazon Route 53 menangani pemeriksaan kondisi untuk titik akhir (seperti alamat IP) untuk menentukan bagaimana pengguna akhir dirutekan ke sana.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Rendah

Panduan implementasi

- Lakukan tugas konstan sehingga sistem tidak gagal saat terdapat perubahan beban yang besar dan cepat.
- Implementasikan dependensi yang digabungkan dengan longgar. Dependensi seperti sistem pengantrean, sistem streaming, alur kerja, dan penyeimbang beban digabungkan secara longgar. Penggabungan longgar membantu memisahkan perilaku suatu komponen dari komponen lainnya yang bergantung pada komponen tersebut, sehingga meningkatkan ketahanan dan ketangkasan.
  - Amazon Builders' Library: Keandalan, kerja konstan, dan pilihan yang tepat
  - AWS Re: Invent 2018: Tutup Loop dan Membuka Pikiran: Cara Mengendalikan Sistem, Besar dan Kecil ARC337 (termasuk pekerjaan konstan)
    - Untuk contoh sistem pemeriksaan kondisi yang memantau 100.000 server, lakukan rekayasa beban kerja sehingga ukuran payload tetap sama berapa pun jumlah keberhasilan atau kegagalan yang terjadi.

Sumber daya

Dokumen terkait:

- AmazonEC2: Memastikan Idempotensi
- Amazon Builders' Library: Tantangan dengan sistem terdistribusi
- Amazon Builders' Library: Keandalan, kerja konstan, dan pilihan yang tepat

Video terkait:

- AWS KTT New York 2019: Pengantar Arsitektur yang Digerakkan oleh Acara dan Amazon EventBridge (05) MAD2
- AWS Re: Invent 2018: Tutup Loop dan Membuka Pikiran: Cara Mengendalikan Sistem, Besar dan Kecil ARC337 (termasuk pekerjaan konstan)
- AWS RE: Invent 2018: Tutup Loop dan Membuka Pikiran: Cara Mengendalikan Sistem, Besar dan Kecil ARC337 (termasuk kopling longgar, kerja konstan, stabilitas statis)
- AWS re:invent 2019: Pindah ke arsitektur berbasis acara (08) SVS3

REL04-BP04 Buat operasi yang bermutasi menjadi idempoten

Layanan idempoten menjamin setiap permintaan diproses tepat satu kali, sehingga pembuatan beberapa permintaan yang identik memiliki efek yang sama seperti membuat satu permintaan. Hal ini memudahkan klien untuk mengimplementasikan percobaan ulang permintaan tanpa khawatir permintaan tersebut akan diproses lebih dari sekali dan menyebabkan kesalahan. Untuk melakukan ini, klien dapat mengeluarkan permintaan API dengan token idempotensi, yang digunakan setiap kali permintaan diulang. API layanan idempoten menggunakan token untuk mengembalikan respons yang identik dengan respons yang dikembalikan saat pertama kali permintaan diselesaikan, bahkan jika status dasar sistem telah berubah.

Dalam sebuah sistem terdistribusi, cukup simpel untuk melakukan tindakan paling banyak satu kali (klien hanya membuat satu permintaan), atau setidaknya satu kali (tetap mengirimkan permintaan sampai klien mendapatkan konfirmasi berhasil). Lebih sulit untuk menjamin suatu tindakan dilakukan tepat satu kali, sehingga membuat beberapa permintaan yang identik memiliki efek yang sama seperti membuat satu permintaan. Menggunakan token idempotensi di API, layanan dapat menerima sebuah permintaan yang bermutasi satu kali atau lebih tanpa perlu membuat data ganda atau efek samping.

Hasil yang diinginkan: Anda memiliki pendekatan yang konsisten, terdokumentasi dengan baik, dan diadopsi secara luas untuk memastikan idempotensi di semua komponen dan layanan.

Anti-pola umum:

- Anda menerapkan idempotensi secara sembarangan, bahkan ketika tidak diperlukan.
- Anda memperkenalkan logika yang terlalu kompleks untuk menerapkan idempotensi.

- Anda menggunakan stempel waktu sebagai kunci untuk idempotensi. Hal ini dapat menyebabkan ketidakakuratan karena perbedaan waktu atau karena banyak klien yang menggunakan stempel waktu yang sama untuk menerapkan perubahan.
- Anda menyimpan seluruh payload untuk idempotensi. Dalam pendekatan ini, Anda menyimpan payload data lengkap untuk setiap permintaan dan menimpanya pada setiap permintaan baru. Hal ini dapat menurunkan kinerja dan memengaruhi skalabilitas.
- Anda menghasilkan kunci secara tidak konsisten di seluruh layanan. Tanpa kunci yang konsisten, layanan mungkin gagal mengenali permintaan duplikat, yang mengakibatkan hasil yang tidak diinginkan.

Manfaat menjalankan praktik terbaik ini:

- Skalabilitas yang lebih besar: Sistem dapat menangani percobaan ulang permintaan dan permintaan duplikat tanpa harus menjalankan logika tambahan atau manajemen status yang kompleks.
- Keandalan yang meningkat: Idempotensi membantu layanan menangani beberapa permintaan identik secara konsisten, yang mengurangi risiko efek samping yang tidak diinginkan atau data duplikat. Hal ini terutama penting dalam sistem terdistribusi, yang sering mengalami kegagalan jaringan dan memerlukan percobaan ulang.
- Konsistensi data yang lebih baik: Karena permintaan yang sama menghasilkan respons yang sama, idempotensi membantu menjaga konsistensi data di seluruh sistem terdistribusi. Hal ini penting untuk menjaga integritas transaksi dan operasi.
- Penanganan kesalahan: Token idempotensi membuat penanganan kesalahan lebih mudah. Jika klien tidak menerima respons karena masalah, klien dapat secara aman mengirim ulang permintaan dengan token idempotensi yang sama.
- Transparansi operasional: Idempotensi memungkinkan pemantauan dan pencatatan log yang lebih baik. Layanan dapat membuat log permintaan dengan token idempotensi mereka, sehingga memudahkan proses pelacakan dan debug masalah.
- Kontrak API yang disederhanakan: Hal ini dapat menyederhanakan kontrak antara klien dan sistem sisi server dan mengurangi kekhawatiran akan kesalahan dalam pemrosesan data.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Sedang

Panduan implementasi

Dalam sebuah sistem terdistribusi, melakukan tindakan maksimal satu kali (klien hanya membuat satu permintaan) atau minimal satu kali (klien terus mengirimkan permintaan sampai klien mendapatkan konfirmasi berhasil) cukup mudah. Namun, sulit untuk menerapkan perilaku tepat satu kali. Untuk mencapai hal ini, klien Anda harus membuat dan memberikan token idempotensi untuk setiap permintaan.

Dengan menggunakan token idempotensi, layanan dapat membedakan antara permintaan baru dan permintaan berulang. Ketika layanan menerima permintaan dengan token idempotensi, layanan tersebut memeriksa apakah token sudah pernah digunakan. Jika token sudah pernah digunakan, layanan mengambil dan mengembalikan respons yang disimpan. Jika token masih baru, layanan memproses permintaan tersebut, menyimpan respons beserta token, lalu mengembalikan respons. Mekanisme ini membuat semua respons idempoten, yang meningkatkan keandalan dan konsistensi sistem terdistribusi.

Idempotensi juga merupakan perilaku penting dari arsitektur berbasis peristiwa. Arsitektur ini biasanya didukung oleh antrean pesan seperti Amazon SQS, Amazon MQ, Amazon Kinesis Streams, atau Amazon Managed Streaming for Apache Kafka (MSK). Dalam beberapa situasi, pesan yang diterbitkan hanya sekali dapat dikirim lebih dari sekali secara tidak sengaja. Ketika penerbit membuat dan menyertakan token idempotensi dalam pesan, penerbit tersebut meminta agar pemrosesan pesan duplikat yang diterima tidak menghasilkan tindakan berulang untuk pesan yang sama. Konsumen harus melacak setiap token yang diterima dan mengabaikan pesan yang berisi token duplikat.

Layanan dan konsumen juga harus meneruskan token idempotensi yang diterima ke layanan hilir apa pun yang dipanggil. Setiap layanan hilir dalam rantai pemrosesan juga harus memastikan bahwa idempotensi diimplementasikan untuk menghindari efek samping berupa pemrosesan pesan lebih dari sekali.

Langkah-langkah implementasi

1. Identifikasi operasi idempoten

   Tentukan operasi mana yang membutuhkan idempotensi. Hal ini biasanya termasuk metode HTTP POST, PUT, dan DELETE dan operasi basis data insert, update, atau delete. Operasi yang tidak mengubah status, seperti kueri hanya-baca, biasanya tidak memerlukan idempotensi kecuali jika memiliki efek samping.
2. Gunakan pengidentifikasi unik

Sertakan token unik di setiap permintaan operasi idempoten yang dikirim oleh pengirim, baik secara langsung dalam permintaan maupun sebagai bagian dari metadatanya (misalnya, header HTTP). Hal ini memungkinkan penerima mengenali dan menangani permintaan atau operasi duplikat. Pengidentifikasi yang biasa digunakan untuk token termasuk Universally Unique Identifiers (UUID) dan K-Sortable Unique Identifiers (KSUID).

3. Lacak dan kelola status

   Pertahankan status setiap operasi atau permintaan dalam beban kerja Anda. Hal ini dapat dicapai dengan menyimpan token idempotensi dan status yang sesuai (seperti tertunda, selesai, atau gagal) dalam basis data, cache, atau penyimpanan persisten lainnya. Informasi status ini memungkinkan beban kerja mengidentifikasi dan menangani permintaan atau operasi duplikat.

   Pertahankan konsistensi dan atomisitas dengan mekanisme kontrol konkurensi yang sesuai jika diperlukan, seperti kunci, transaksi, atau kontrol konkurensi optimis. Hal ini termasuk proses merekam token idempoten dan menjalankan semua operasi bermutasi yang terkait dengan pemrosesan permintaan. Hal ini membantu mencegah "race condition" dan memverifikasi bahwa operasi idempoten berjalan dengan benar.

   Hapus token idempotensi lama secara teratur dari penyimpanan data untuk mengelola penyimpanan dan kinerja. Jika sistem penyimpanan Anda mendukungnya, pertimbangkan untuk menggunakan stempel waktu kedaluwarsa untuk data (sering dikenal sebagai nilai time to live atau TTL). Kemungkinan penggunaan ulang token idempotensi berkurang seiring waktu.

   Opsi penyimpanan AWS umum yang biasanya digunakan untuk menyimpan token idempotensi dan status terkait mencakup:

   - Amazon DynamoDB: DynamoDB adalah layanan basis data NoSQL yang menyediakan kinerja latensi rendah dan ketersediaan tinggi, yang membuatnya sangat cocok untuk penyimpanan data terkait idempotensi. Model data nilai-kunci dan dokumen DynamoDB memungkinkan penyimpanan dan pengambilan token idempotensi dan informasi status terkait secara efisien. DynamoDB juga dapat menghapus token idempotensi secara otomatis jika aplikasi Anda menetapkan nilai TTL saat menyisipkannya.
   - Amazon ElastiCache: ElastiCache dapat menyimpan token idempotensi dengan throughput tinggi, latensi rendah, dan hemat biaya. ElastiCache (Redis) dan ElastiCache (Memcached) juga dapat menghapus token idempotensi secara otomatis jika aplikasi Anda menetapkan nilai TTL saat menyisipkannya.

- Amazon Relational Database Service (RDS): Anda dapat menggunakan Amazon RDS untuk menyimpan token idempotensi dan informasi status terkait, terutama jika aplikasi Anda sudah menggunakan basis data relasional untuk tujuan lain.
- Amazon Simple Storage Service (S3): Amazon S3 adalah layanan penyimpanan objek yang sangat mudah diskalakan dan tahan lama yang dapat digunakan untuk menyimpan token idempotensi dan metadata terkait. Kemampuan penentuan versi S3 dapat berguna terutama untuk pemeliharaan status operasi idempoten. Pilihan layanan penyimpanan biasanya tergantung pada faktor-faktor seperti volume data terkait idempotensi, karakteristik kinerja yang diperlukan, kebutuhan akan daya tahan dan ketersediaan, dan bagaimana mekanisme idempotensi terintegrasi dengan arsitektur beban kerja secara keseluruhan.

4. Implementasikan operasi idempoten

   Rancang komponen API dan beban kerja Anda agar idempoten. Gabungkan pemeriksaan idempotensi ke dalam komponen beban kerja Anda. Sebelum Anda memproses permintaan atau melakukan suatu operasi, periksa apakah pengidentifikasi uniknya sudah diproses. Jika sudah, kembalikan hasil sebelumnya, bukan menjalankan operasi kembali. Misalnya, jika klien mengirim permintaan untuk membuat pengguna, periksa apakah sudah ada pengguna dengan pengidentifikasi unik yang sama. Jika pengguna sudah ada, sistem seharusnya mengembalikan informasi pengguna yang sudah ada tersebut, bukan membuat yang baru. Demikian pula, jika konsumen antrean menerima pesan dengan token idempotensi duplikat, konsumen ini harus mengabaikan pesan tersebut.

   Buat rangkaian pengujian komprehensif yang memvalidasi idempotensi permintaan. Pengujian tersebut harus mencakup berbagai skenario, seperti permintaan berhasil, permintaan gagal, dan permintaan duplikat.

   Jika beban kerja Anda memanfaatkan fungsi AWS Lambda, pertimbangkan Powertools for AWS Lambda. Powertools for AWS Lambda adalah toolkit developer yang membantu mengimplementasikan praktik terbaik nirserver dan meningkatkan kecepatan developer saat Anda bekerja dengan fungsi AWS Lambda. Secara khusus, alat ini menyediakan utilitas untuk mengubah fungsi Lambda Anda menjadi operasi idempoten yang aman untuk dicoba ulang.

5. Komunikasikan idempotensi dengan jelas

   Dokumentasikan komponen API dan beban kerja Anda untuk mengomunikasikan dengan jelas sifat operasi yang idempoten. Hal ini membantu klien memahami perilaku yang diharapkan dan cara berinteraksi dengan beban kerja Anda secara andal.

6. Pantau dan audit

Implementasikan mekanisme pemantauan dan audit untuk mendeteksi masalah apa pun yang terkait dengan idempotensi respons, seperti variasi respons yang tidak terduga atau penanganan permintaan duplikat yang berlebihan. Hal ini dapat membantu Anda mendeteksi dan menyelidiki masalah atau perilaku yang tidak terduga dalam beban kerja Anda.

Sumber daya

Praktik-praktik terbaik terkait:

- REL05-BP03 Mengontrol dan membatasi panggilan percobaan ulang
- REL06-BP01 Memantau semua komponen untuk beban kerja (Pembuatan)
- REL06-BP03 Mengirimkan notifikasi (Pemrosesan dan pembuatan alarm waktu nyata)
- REL08-BP02 Integrasikan pengujian fungsional sebagai bagian dari deployment Anda

Dokumen terkait:

- Amazon Builders' Library: Membuat pengulangan aman dengan API idempoten
- Amazon Builders' Library: Tantangan dengan sistem terdistribusi
- Amazon Builders' Library: Keandalan, kerja konstan, dan secangkir kopi nikmat
- Amazon Elastic Container Service: Memastikan idempotensi
- Bagaimana cara membuat fungsi Lambda saya idempoten?
- Memastikan idempotensi dalam permintaan API Amazon EC2

Video terkait:

- Membangun Aplikasi Terdistribusi dengan Arsitektur yang Didorong Peristiwa - AWS Online Tech Talks
- AWS re:invent 2023 - Membangun aplikasi generasi berikutnya dengan arsitektur yang didorong peristiwa
- AWS re:Invent 2023 - Pola integrasi tingkat lanjut & kompromi untuk sistem yang digabungkan dengan metode penggabungan longgar
- AWS re:Invent 2023 - Pola yang didorong peristiwa tingkat lanjut dengan Amazon EventBridge
- AWS re:Invent 2018: Close Loops and Opening Minds: Cara Mengontrol Sistem, ARC337 Besar dan Kecil (mencakup penggabungan longgar, kerja konstan, stabilitas statis)

- AWS re:Invent 2019: Beralih ke arsitektur berbasis peristiwa (SVS308)

Alat terkait:

- Idempotensi dengan AWS Lambda Powertools (Java)
- Idempotensi dengan AWS Lambda Powertools (Python)
- Halaman GitHub AWS Lambda Powertools

Mendesain interaksi dalam sistem terdistribusi untuk mitigasi atau bertahan dari kegagalan

Sistem terdistribusi mengandalkan jaringan komunikasi untuk membuat interkoneksi komponen (seperti server atau layanan). Beban kerja Anda harus beroperasi secara andal terlepas latensi atau hilangnya data pada jaringan-jaringan ini. Komponen dari sistem terdistribusi harus beroperasi dengan cara yang tidak secara negatif memengaruhi beban kerja atau komponen-komponen lain. Berbagai praktik terbaik ini memungkinkan beban kerja bertahan dari stres atau kegagalan, lebih cepat pulih darinya, dan memitigasi dampak gangguan tersebut. Hasilnya yakni peningkatan dalam waktu rata-rata untuk pemulihan (MTTR).

Berbagai praktik terbaik ini mencegah kegagalan dan meningkatkan waktu rata-rata antara kegagalan (MTBF).

Praktik terbaik

- REL05-BP01 Mengimplementasikan degradasi yang tepat (graceful degradation) untuk mengubah dependensi keras yang berlaku menjadi dependensi lunak
- REL05-BP02 Membatasi (throttling) permintaan
- REL05-BP03 Kontrol dan batasi panggilan coba lagi
- REL05-BP04 Melakukan gagal cepat (fail fast) dan membatasi antrean
- REL05-BP05 Mengatur batas waktu klien
- REL05-BP06 Membuat sistem tanpa kewarganegaraan jika memungkinkan
- REL05-BP07 Menerapkan tuas darurat

REL05-BP01 Mengimplementasikan degradasi yang tepat (graceful degradation) untuk mengubah dependensi keras yang berlaku menjadi dependensi lunak

Komponen aplikasi harus terus menjalankan fungsi intinya bahkan jika dependensi menjadi tidak tersedia. Komponen mungkin menyajikan data yang sedikit basi, data alternatif, atau bahkan tidak menyajikan data sama sekali. Hal ini memastikan fungsi sistem secara keseluruhan hanya terhambat secara minimum oleh kegagalan lokal sekaligus memberikan nilai bisnis utama.

Hasil yang diinginkan: Saat dependensi sebuah komponen tidak optimum, komponen tersebut masih dapat berfungsi, meskipun terbatas atau terdegradasi. Mode-mode kegagalan komponen harus dipandang sebagai operasi normal. Alur kerja harus dirancang dengan desain sedemikian rupa sehingga kegagalan tersebut tidak menyebabkan kegagalan total atau setidaknya hanya menyebabkan keadaan yang dapat diprediksi dan dapat dipulihkan.

Anti-pola umum:

- Tidak mengidentifikasi fungsi bisnis inti yang dibutuhkan. Tidak menguji bahwa komponen berfungsi bahkan selama kegagalan dependensi.
- Tidak menyajikan data jika terjadi kesalahan atau ketika hanya ada satu dari beberapa dependensi yang tidak tersedia dan hasil sebagian masih dapat dikembalikan.
- Menciptakan sebuah keadaan yang tidak konsisten ketika transaksi mengalami gagal sebagian.
- Tidak memiliki cara alternatif untuk mengakses tempat penyimpanan parameter pusat.
- Membatalkan atau mengosongkan status lokal sebagai akibat dari penyegaran yang gagal tanpa mempertimbangkan konsekuensi yang ditimbulkan oleh tindakan tersebut.

Manfaat menerapkan praktik terbaik ini: Degradasi bertahap (graceful degradation) akan meningkatkan ketersediaan sistem secara keseluruhan dan mempertahankan fungsionalitas dari fungsi-fungsi yang paling penting, bahkan selama terjadi kegagalan.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Tinggi

Panduan implementasi

Menerapkan degradasi yang tepat akan membantu Anda meminimalkan dampak kegagalan dependensi yang terjadi pada fungsi komponen. Idealnya, sebuah komponen mendeteksi kegagalan-

kegagalan dependensi dan menanganinya dengan cara yang berdampak minim pada pelanggan atau komponen lain.

Merancang arsitektur untuk degradasi yang tepat berarti mempertimbangkan potensi mode kegagalan selama desain dependensi. Untuk setiap mode kegagalan, miliki cara untuk menghadirkan sebagian besar atau setidaknya fungsionalitas yang paling penting dari komponen kepada pemanggil atau pelanggan. Pertimbangan-pertimbangan ini dapat menjadi persyaratan tambahan yang dapat diuji dan diverifikasi. Idealnya, sebuah komponen harus mampu menjalankan fungsi intinya dengan cara yang dapat diterima bahkan ketika satu atau beberapa dependensi gagal.

Ini bukan hanya pembahasan teknis, melainkan juga pembahasan bisnis. Semua persyaratan bisnis penting dan harus dipenuhi, jika memungkinkan. Namun demikian, menanyakan apa yang seharusnya terjadi ketika tidak semua persyaratan tersebut dapat dipenuhi adalah hal yang wajar. Suatu sistem dapat dirancang agar tersedia dan konsisten, tetapi dalam keadaan yang mengharuskan salah satu persyaratan untuk dikorbankan, mana yang lebih penting? Untuk pemrosesan pembayaran, jawabannya mungkin adalah konsistensi. Untuk aplikasi waktu nyata, jawabannya mungkin adalah ketersediaan. Untuk sebuah situs web yang digunakan langsung oleh pelanggan, jawabannya mungkin tergantung pada ekspektasi pelanggan.

Seberapa pentingnya, ini tergantung persyaratan komponen dan apa yang seharusnya dianggap sebagai fungsi intinya. Misalnya:

- Situs web ecommerce mungkin akan menampilkan data dari berbagai sistem, misalnya rekomendasi yang dipersonalisasi, produk dengan peringkat tertinggi, dan status pesanan pelanggan di halaman arahan. Ketika salah satu sistem hulu gagal, masih masuk akal untuk menampilkan semua daripada menampilkan halaman kesalahan kepada pelanggan.
- Sebuah komponen yang menjalankan penulisan batch masih dapat melanjutkan pemrosesan batch jika salah satu operasi mengalami kegagalan. Implementasi mekanisme percobaan ulang harus sederhana. Hal ini dapat dilakukan dengan mengembalikan informasi tentang operasi-operasi yang berhasil, yang telah gagal, dan mengapa operasi-operasi tersebut gagal ke pemanggil, atau dengan menempatkan permintaan yang gagal ke dalam antrean surat mati untuk mengimplementasikan percobaan ulang tidak selaras. Informasi tentang operasi-operasi yang gagal juga harus dibuatkan log.
- Sebuah sistem yang memproses transaksi harus memastikan bahwa semua pembaruan individual dijalankan atau tidak sama sekali. Untuk transaksi-transaksi terdistribusi, pola saga dapat digunakan untuk kembali ke operasi sebelumnya jika operasi selanjutnya dari transaksi yang sama mengalami kegagalan. Di sini, fungsi intinya adalah menjaga konsistensi.

- Sistem-sistem time-critical harus mampu menangani dependensi yang tidak memberikan respons secara tepat waktu. Dalam kasus-kasus ini, pola pemutus sirkuit dapat digunakan. Ketika respons dari sebuah dependensi mulai mencapai batas waktu, sistem dapat beralih ke keadaan ditutup di mana tidak ada panggilan tambahan yang dibuat.
- Sebuah aplikasi dapat membaca parameter dari tempat penyimpanan parameter. Membuat citra kontainer dengan serangkaian parameter default akan membantu agar apabila tempat penyimpanan parameter tidak tersedia citra tersebut dapat digunakan.

Perlu diperhatikan bahwa jalur-jalur yang diambil jika terjadi kegagalan komponen perlu diuji dan harus jauh lebih sederhana daripada jalur-jalur utama. Umumnya, strategi fallback harus dihindari.

Langkah-langkah implementasi

Identifikasi dependensi eksternal dan internal. Pertimbangkan jenis-jenis kegagalan yang bisa terjadi di dalamnya. Pikirkan tentang cara-cara yang dapat meminimalkan dampak negatif terhadap pelanggan serta sistem hulu dan hilir selama kegagalan-kegagalan tersebut.

Berikut ini adalah daftar dependensi dan cara melakukan degradasi yang tepat ketika dependensi mengalami kegagalan:

1. Kegagalan dependensi parsial: Sebuah komponen dapat melakukan beberapa permintaan ke sistem-sistem hilir, baik beberapa permintaan ke satu sistem atau satu permintaan ke beberapa sistem. Tergantung konteks bisnis, mungkin ada berbagai cara penanganan yang sesuai (untuk detail lebih lanjut, silakan lihat contoh-contoh sebelumnya dalam Panduan implementasi).
2. Sistem hilir tidak dapat memproses permintaan karena beban tinggi: Jika permintaan ke sistem hilir terus-menerus gagal, sebaiknya Anda tidak mencoba lagi. Tindakan ini dapat menciptakan beban tambahan pada sistem yang sudah mengalami kelebihan beban dan mempersulit pemulihan. Pola pemutus sirkuit dapat digunakan di sini, yang memantau kegagalan panggilan ke sistem hilir. Jika ada banyak panggilan yang mengalami kegagalan, permintaan akan berhenti dikirimkan ke sistem hilir dan hanya sesekali panggilan dibiarkan masuk untuk menguji apakah sistem hilir sudah tersedia kembali.
3. Gudang parameter tidak tersedia: Untuk mengubah tempat penyimpanan parameter, caching dependensi lunak atau sane default yang disertakan di dalam image kontainer atau mesin dapat digunakan. Perlu diperhatikan bahwa default ini harus selalu diperbarui dan disertakan dalam rangkaian pengujian.
4. Layanan pemantauan atau dependensi non-fungsional lainnya tidak tersedia: Jika sebuah komponen sebentar-sebentar tidak dapat mengirim log, metrik, atau jejak ke layanan pemantauan

pusat, langkah terbaiknya sering kali adalah tetap menjalankan fungsi-fungsi bisnis seperti biasa. Diam-diam tidak membuat log atau mendorong metrik dalam waktu yang lama sering kali tidak dapat diterima. Selain itu, beberapa kasus penggunaan mungkin akan memerlukan entri audit lengkap untuk memenuhi persyaratan-persyaratan kepatuhan.

5. Instans primer basis data relasional mungkin tidak tersedia: Amazon Relational Database Service, seperti hampir semua database relasional, hanya dapat memiliki satu contoh penulis utama. Hal ini akan menciptakan satu titik kegagalan untuk beban kerja tulis dan menjadikan penskalaan menjadi lebih sulit. Hal ini dapat diatasi sebagiannya dengan menggunakan konfigurasi Multi-AZ untuk mendapatkan ketersediaan tinggi atau Amazon Aurora Nirserver untuk mendapatkan penskalaan yang lebih baik. Untuk persyaratan-persyaratan ketersediaan yang sangat tinggi, ada baiknya untuk tidak bergantung pada penulis utama sama sekali. Untuk kueri yang hanya membaca, replika baca dapat digunakan, yang memberikan redundansi dan kemampuan untuk melakukan penambahan skala (scale-out), bukan hanya scale-up. Tulis dapat di-buffer, misalnya dalam antrean Amazon Simple Queue Service, sehingga permintaan tulis dari pelanggan masih dapat diterima bahkan jika penulis utama tidak tersedia untuk sementara.

Sumber daya

Dokumen terkait:

- Amazon API Gateway: Membatasi Permintaan API untuk Peningkatan Throughput
- CircuitBreaker (rangkuman Pemutus Sirkuit dari buku “Release It!”)
- Percobaan Ulang Kesalahan dan Penundaan Eksponensial di AWS
- Michael Nygard “Release It! Rancang dan Lakukan Deployment Perangkat Lunak yang Siap Diproduksi”
- Amazon Builders' Library: Menghindari fallback dalam sistem terdistribusi
- Amazon Builders' Library: Menghindari backlog antrean yang tidak dapat diatasi
- Amazon Builders' Library: Tantangan dan strategi caching
- Amazon Builders' Library: Batas waktu, percobaan ulang, dan penundaan dengan jitter

Video terkait:

- Percobaan ulang, penundaan, dan jitter: AWS re:Invent 2019: Memperkenalkan Amazon Builders' Library (DOP328)

REL05-BP02 Membatasi (throttling) permintaan

Batasi permintaan untuk memitigasi kehabisan sumber daya karena peningkatan permintaan yang tidak terduga. Permintaan di bawah tingkat throttling akan diproses, sedangkan permintaan di atas batas yang ditentukan akan ditolak dengan memunculkan pesan bahwa permintaan telah dibatasi.

Hasil yang diinginkan: Lonjakan volume dalam jumlah besar baik dari peningkatan lalu lintas pelanggan yang naik tiba-tiba, serangan membanjir, atau banjir percobaan ulang akan diminimalkan dengan throttling permintaan, sehingga beban kerja dapat melanjutkan pemrosesan volume permintaan normal yang didukung.

Anti-pola umum:

- Throttling titik akhir API tidak diimplementasikan atau dibiarkan pada nilai default tanpa mempertimbangkan volume yang diharapkan.
- Titik akhir API tidak diberi uji beban atau batas throttling tidak diuji.
- Lakukan throttling terhadap angka permintaan tanpa mempertimbangkan ukuran atau kompleksitas permintaan.
- Melakukan uji laju permintaan maksimum atau uji ukuran permintaan maksimum, tetapi tidak menguji keduanya bersama-sama.
- Sumber daya tidak disediakan untuk batas yang sama yang ditetapkan dalam pengujian.
- Rencana penggunaan belum dikonfigurasi atau dipertimbangkan untuk konsumen API aplikasi ke aplikasi (A2A).
- Tidak ada konfigurasi pengaturan konkurensi maksimum pada konsumen antrean yang mengalami penskalaan horizontal.
- Pembatasan tingkat untuk setiap alamat IP belum diimplementasikan.

Manfaat menerapkan praktik terbaik ini: Beban kerja yang menetapkan batas throttling dapat beroperasi secara normal dan berhasil memproses beban permintaan yang diterima saat terjadi lonjakan volume yang tidak terduga. Lonjakan permintaan yang terjadi tiba-tiba atau secara terus menerus pada API dan antrean akan dibatasi (throttling) dan tidak menghabiskan sumber daya pemrosesan permintaan. Batas angka permintaan akan membatasi (throttling) setiap pengirim permintaan sehingga volume lalu lintas yang tinggi dari satu alamat IP atau konsumen API tidak akan menghabiskan sumber daya atau berimbas pada konsumen yang lain.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Tinggi

Panduan implementasi

Layanan-layanan harus dirancang untuk memproses kapasitas permintaan yang diketahui; kapasitas ini dapat ditetapkan melalui pengujian beban. Jika laju kedatangan permintaan sudah melampaui batas, maka respons yang tepat menandakan bahwa permintaan telah mengalami throttling. Hal ini akan memungkinkan konsumen untuk menangani kesalahan dan mencoba ulang di lain waktu.

Saat layanan-layanan Anda memerlukan implementasi throttling, pertimbangkan untuk mengimplementasikan algoritme bucket token, yang menghitung satu token sebagai satu permintaan. Token diisi ulang dengan laju throttling per detik dan dikosongkan secara tidak selaras dengan satu token per permintaan.

Algoritme bucket token.

Amazon API Gateway mengimplementasikan algoritme bucket token sesuai dengan batas yang dimiliki akun dan wilayah dan dapat dikonfigurasi untuk setiap klien dengan rencana penggunaan. Selain itu, Amazon Simple Queue Service (Amazon SQS) dan Amazon Kinesis dapat menyangga permintaan untuk memperlancar laju permintaan, dan memungkinkan laju throttling yang lebih tinggi untuk permintaan yang dapat ditangani. Terakhir, Anda dapat menerapkan pembatasan laju dengan AWS WAF untuk melakukan throttling terhadap konsumen API tertentu yang menghasilkan beban yang luar biasa tinggi.

Langkah-langkah implementasi

Anda dapat mengonfigurasi API Gateway dengan batas throttling untuk API Anda dan menampilkan kesalahan `429 Too Many Requests` saat batas terlampaui. Anda dapat menggunakan AWS WAF

dengan titik akhir AWS AppSync dan API Gateway Anda untuk mengaktifkan pembatasan laju untuk setiap alamat IP. Selain itu, apabila sistem Anda dapat memberikan toleransi terhadap pemrosesan tidak selaras, Anda dapat memasukkan pesan ke dalam antrean atau aliran guna mempercepat respons terhadap klien layanan, yang memungkinkan Anda untuk melakukan lonjakan ke tingkat throttling yang lebih tinggi.

Dengan pemrosesan asinkron, ketika Anda telah mengonfigurasi Amazon SQS sebagai sumber peristiwa untuk AWS Lambda, Anda dapat mengonfigurasi konkurensi maksimum untuk menghindari tingkat peristiwa yang tinggi dari penggunaan kuota eksekusi konkuren akun yang tersedia yang diperlukan untuk layanan lain di beban kerja atau akun Anda.

Meskipun API Gateway menyediakan implementasi bucket token yang dikelola, apabila Anda tidak dapat menggunakan API Gateway, Anda dapat memanfaatkan implementasi sumber terbuka bahasa khusus (lihat contoh terkait di Sumber Daya) bucket token untuk layanan Anda.

- Memahami dan mengonfigurasi batas throttling API Gateway di level akun per wilayah, API per tahap, dan kunci API per level paket penggunaan.
- Menerapkan aturan pembatasan laju AWS WAF ke API Gateway dan titik akhir AWS AppSync untuk melindungi dari banjir dan memblokir IP berbahaya. Aturan-aturan pembatas laju juga dapat dikonfigurasi pada kunci API AWS AppSync untuk konsumen A2A.
- Pertimbangkan apakah Anda memerlukan kontrol throttling yang lebih besar daripada pembatasan laju untuk API AWS AppSync, dan jika demikian, konfigurasikan API Gateway di depan titik akhir AWS AppSync Anda.
- Saat antrean Amazon SQS disiapkan sebagai pemicu bagi konsumen antrean Lambda, tetapkan konkurensi maksimum ke nilai yang cukup diproses untuk memenuhi tujuan tingkat layanan Anda tetapi tidak menggunakan batas konkurensi yang memengaruhi fungsi Lambda lainnya. Pertimbangkan untuk menetapkan konkurensi cadangan pada fungsi Lambda lain di akun dan wilayah yang sama saat Anda menggunakan antrean dengan Lambda.
- Gunakan API Gateway dengan integrasi layanan native ke Amazon SQS atau Kinesis untuk melakukan buffering terhadap permintaan.
- Jika Anda tidak dapat menggunakan API Gateway, lihat pustaka bahasa khusus untuk mengimplementasikan algoritme bucket token untuk beban kerja Anda. Periksa bagian contoh dan lakukan riset sendiri untuk menemukan pustaka yang sesuai.
- Uji batas yang ingin Anda tetapkan, atau yang ingin Anda izinkan untuk ditingkatkan, dan dokumentasikan batas-batas yang diuji.

- Jangan tingkatkan batas melebihi apa yang Anda tetapkan dalam pengujian. Saat meningkatkan batas, pastikan bahwa sumber daya yang disediakan sudah setara atau lebih besar daripada yang ada dalam skenario pengujian sebelum menerapkan peningkatan.

Sumber daya

Praktik-praktik terbaik terkait:

- REL04-BP03 Lakukan pekerjaan konstan
- REL05-BP03 Kontrol dan batasi panggilan coba lagi

Dokumen terkait:

- Amazon API Gateway: Membatasi Permintaan API untuk Peningkatan Throughput
- AWS WAF: Pernyataan aturan berbasis laju
- Memperkenalkan konkurensi maksimum AWS Lambda saat menggunakan Amazon SQS sebagai sumber peristiwa
- AWS Lambda: Konkurensi Maksimum

Contoh terkait:

- Tiga aturan berbasis laju AWS WAF yang paling penting
- Java Bucket4j
- Bucket token Python
- Bucket token Node
- Pembatasan Tingkat Threading Sistem .NET

Video terkait:

- Mengimplementasikan praktik terbaik keamanan API GraphQL dengan AWS AppSync

Alat terkait:

- Amazon API Gateway
- AWS AppSync

- Amazon SQS
- Amazon Kinesis
- AWS WAF
- Ruang Tunggu Virtual di AWS

REL05-BP03 Kontrol dan batasi panggilan coba lagi

Gunakan penundaan eksponensial untuk mencoba ulang permintaan dengan interval yang makin lama antara setiap percobaan ulang. Terapkan jitter antara percobaan ulang untuk mengacak interval percobaan ulang. Batasi jumlah percobaan ulang maksimum.

Hasil yang diinginkan: Komponen khas dalam sistem perangkat lunak terdistribusi termasuk server, penyeimbang beban, database, dan server. DNS Selama operasi normal, komponen-komponen ini dapat merespons permintaan yang memiliki kesalahan yang bersifat sementara atau terbatas, dan juga kesalahan yang persisten terlepas dari percobaan ulang. Ketika klien membuat permintaan ke layanan, permintaan tersebut mengonsumsi sumber daya termasuk memori, thread, koneksi, port, atau sumber daya terbatas lainnya. Mengontrol dan membatasi percobaan ulang adalah strategi untuk melepaskan dan meminimalkan konsumsi sumber daya sehingga komponen sistem yang ada di bawah tekanan tidak kewalahan.

Ketika permintaan klien mengalami batas waktu atau menerima respons kesalahan, mereka harus menentukan apakah akan mencoba lagi atau tidak. Jika mereka mencoba lagi, mereka melakukannya dengan penundaan eksponensial dengan jitter dan nilai coba ulang maksimum. Karena itu, layanan dan proses backend mendapat kelonggaran beban dan waktu untuk pulih secara mandiri, sehingga hal itu akan mengakibatkan terjadinya pemulihan yang lebih cepat dan pelayanan permintaan yang berhasil.

Anti-pola umum:

- Mengimplementasikan percobaan ulang tanpa menambahkan penundaan eksponensial, jitter, dan nilai coba ulang maksimum. Penundaan dan jitter membantu menghindari lonjakan lalu lintas semu yang disebabkan percobaan ulang yang dikoordinasikan secara tidak sengaja pada interval umum.
- Menerapkan percobaan ulang tanpa menguji efeknya atau dengan asumsi percobaan ulang sudah dibangun ke dalam skenario coba lagi SDK tanpa pengujian.
- Tidak memahami kode kesalahan yang dipublikasikan dari dependensi, yang menyebabkan percobaan ulang semua kesalahan, termasuk kesalahan dengan penyebab yang dengan jelas

menunjukkan tidak adanya izin, kesalahan konfigurasi, atau kondisi-kondisi lain yang jelas tidak akan terselesaikan tanpa intervensi manual.

- Tidak menangani praktik-praktik observabilitas, termasuk pemantauan dan peringatan tentang kegagalan layanan berulang sehingga masalah yang mendasari dapat diketahui dan diatasi.
- Mengembangkan mekanisme-mekanisme percobaan ulang kustom saat kemampuan coba ulang bawaan atau pihak ketiga sudah mencukupi.
- Melakukan percobaan ulang pada beberapa lapisan tumpukan aplikasi dengan cara yang makin memperparah upaya-upaya percobaan ulang sehingga makin menyita sumber daya yang sedang berada dalam badai percobaan ulang. Pastikan bahwa Anda memahami bagaimana kesalahan-kesalahan ini memengaruhi aplikasi Anda dan dependensi yang Anda andalkan, dan kemudian terapkan percobaan ulang hanya pada satu tingkat.
- Melakukan percobaan ulang pada panggilan layanan yang tidak idempoten, sehingga menyebabkan efek samping yang tidak terduga seperti hasil-hasil ganda.

Manfaat menerapkan praktik terbaik ini: Percobaan ulang akan membantu klien memperoleh hasil yang diinginkan ketika permintaan mengalami kegagalan, tetapi juga dapat menyita lebih banyak waktu server untuk mendapatkan respons berhasil yang mereka inginkan. Ketika kegagalan jarang terjadi atau sementara, percobaan ulang dapat berfungsi dengan baik. Ketika kegagalan disebabkan oleh kelebihan beban sumber daya, melakukan percobaan ulang dapat memperburuk keadaan. Menambahkan penundaan eksponensial dengan jitter ke percobaan ulang klien memungkinkan server pulih ketika kegagalan disebabkan oleh kelebihan beban sumber daya. Jitter menghindarkan penyelarasan permintaan menjadi lonjakan, dan penundaan dapat mengurangi eskalasi beban yang disebabkan oleh penambahan percobaan ulang ke beban permintaan normal. Terakhir, penting bagi Anda untuk mengonfigurasi jumlah coba ulang maksimum atau waktu yang telah berlalu untuk menghindari terciptanya backlog yang dapat menghasilkan kegagalan yang bersifat metastabil.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Tinggi

Panduan implementasi

Mengontrol dan membatasi panggilan percobaan ulang. Gunakan penundaan eksponensial untuk percobaan ulang setelah interval yang makin lama. Masukkan jitter untuk mengacak interval percobaan ulang dan batasi jumlah percobaan ulang maksimum.

Beberapa AWS SDKs mengimplementasikan percobaan ulang dan backoff eksponensial secara default. Gunakan AWS implementasi bawaan ini jika berlaku dalam beban kerja Anda.

Implementasikan logika serupa dalam beban kerja Anda saat memanggil layanan yang bersifat idempoten dan apabila percobaan ulang meningkatkan ketersediaan klien Anda. Tentukan batas waktu dan kapan harus berhenti melakukan percobaan ulang berdasarkan kasus penggunaan Anda. Buat dan latih skenario pengujian untuk kasus-kasus penggunaan percobaan ulang tersebut.

Langkah-langkah implementasi

- Tentukan lapisan optimal dalam yang ada dalam tumpukan aplikasi Anda untuk mengimplementasikan percobaan ulang untuk layanan-layanan yang diandalkan aplikasi Anda.
- Sadarilah SDKs yang ada yang menerapkan strategi coba lagi yang telah terbukti dengan backoff eksponensial dan jitter untuk bahasa pilihan Anda, dan dukung ini daripada menulis implementasi coba ulang Anda sendiri.
- Verifikasi bahwa layanan sudah idempoten sebelum Anda menerapkan percobaan ulang. Setelah percobaan ulang diterapkan, pastikan bahwa keduanya diuji dan latihlah secara rutin dalam lingkungan produksi.
- Saat memanggil AWS layananAPIs, gunakan AWS SDKsdan AWS CLIdan pahami opsi konfigurasi coba lagi. Tentukan apakah konfigurasi default cocok untuk kasus penggunaan Anda, uji, dan lakukan penyesuaian sesuai kebutuhan.

Sumber daya

Praktik-praktik terbaik terkait:

- REL04-BP04 Buat operasi yang bermutasi menjadi idempoten
- REL05-BP02 Membatasi (throttling) permintaan
- REL05-BP04 Melakukan gagal cepat (fail fast) dan membatasi antrean
- REL05-BP05 Mengatur batas waktu klien
- REL11-BP01 Memantau semua komponen beban kerja untuk mendeteksi kegagalan

Dokumen terkait:

- Kesalahan Mencoba Ulang dan Backoff Eksponensial di AWS
- Amazon Builders' Library: Batas waktu, percobaan ulang, dan penundaan dengan jitter
- Penundaan Eksponensial dan Jitter
- Membuat percobaan ulang aman dengan idempoten APIs

Contoh terkait:

- Spring Retry
- Resilience4j Retry

Video terkait:

- Coba lagi, backoff, dan jitter: AWS re:Invent 2019: Memperkenalkan Perpustakaan Pembangun Amazon () DOP328

Alat terkait:

- AWS SDKsdan Alat: Coba lagi perilaku
- AWS Command Line Interface: mencoba AWS CLI lagi

REL05-BP04 Melakukan gagal cepat (fail fast) dan membatasi antrean

Ketika layanan tidak berhasil merespons permintaan, lakukanlah gagal cepat (fail fast). Hal ini memungkinkan pelepasan sumber daya yang terkait dengan permintaan, dan mengizinkan layanan untuk melakukan pemulihan ketika layanan tersebut kehabisan sumber daya. Gagal cepat (fail fast) adalah pola desain perangkat lunak mapan yang dapat dimanfaatkan untuk membangun beban kerja yang sangat andal di cloud. Antrean juga merupakan pola integrasi korporat yang sudah mapan yang dapat memperlancar beban dan memungkinkan klien untuk melepaskan sumber daya ketika pemrosesan tak selaras dapat ditoleransi. Ketika layanan berhasil memberikan respons dalam kondisi normal tetapi gagal ketika laju permintaan terlalu tinggi, gunakan antrean untuk melakukan buffering permintaan. Namun demikian, jangan sampai ada penumpukan backlog antrean panjang yang dapat mengakibatkan diprosesnya permintaan-permintaan yang telah kedaluwarsa dan telah ditinggalkan oleh klien.

Hasil yang diinginkan: Ketika sistem berebut sumber daya, mengalami waktu habis, pengecualian, atau grey failure (kegagalan samar-samar) yang menyebabkan target tingkat layanan tidak dapat dicapai, strategi gagal cepat (fail fast) akan memungkinkan Anda melakukan pemulihan sistem yang lebih cepat. Sistem yang harus menyerap lonjakan lalu lintas dan dapat mengakomodasi pemrosesan tak selaras dapat meningkatkan keandalan dengan memungkinkan klien untuk secara cepat melepaskan permintaan dengan menggunakan antrean untuk melakukan buffering permintaan ke layanan backend. Ketika melakukan buffering permintaan ke antrean, strategi manajemen antrean diimplementasikan untuk menghindari backlog yang terlalu membebani.

Anti-pola umum:

- Mengimplementasikan antrean pesan tetapi tidak mengonfigurasi antrean surat mati (DLQ) atau alarm pada volume DLQ untuk mendeteksi kegagalan yang terjadi pada sistem.
- Tidak mengukur usia pesan dalam antrean, yakni ukuran latensi untuk mengetahui kapan konsumen antrean tertinggal atau mengalami kesalahan yang menyebabkan percobaan ulang.
- Tidak menghapus pesan-pesan yang menumpuk dari antrean, padahal tidak ada gunanya memproses pesan-pesan tersebut jika kebutuhan bisnis sudah tidak lagi ada.
- Mengonfigurasi antrean berbasis first in first out (FIFO), padahal antrean berbasis last in first out (LIFO) lebih memenuhi kebutuhan klien, misalnya ketika pengurutan yang ketat tidak diperlukan dan pemrosesan backlog menunda semua permintaan baru dan sensitif waktu sehingga semua klien merasakan bahwa tingkat layanan gagal dipenuhi.
- Mengekspos antrean internal ke klien, bukan mengekspos API yang mengelola masuknya pekerjaan dan menempatkan permintaan ke dalam antrean internal.
- Menggabungkan terlalu banyak jenis permintaan kerja ke dalam satu antrean tunggal yang dapat memperburuk kondisi backlog dengan menyebarkan permintaan sumber daya di seluruh jenis permintaan.
- Memproses permintaan-permintaan yang kompleks dan sederhana dalam antrean yang sama, sehingga mengabaikan perbedaan kebutuhan pemantauan, batas waktu, dan alokasi sumber daya.
- Tidak memvalidasi input atau menggunakan pernyataan untuk mengimplementasikan mekanisme gagal cepat (fail fast) dalam perangkat lunak yang menaikkan pengecualian ke komponen dengan level lebih tinggi yang dapat menangani kesalahan secara mulus.
- Tidak menghapus sumber daya yang rusak dari perutean permintaan, terutama ketika terjadi kegagalan samar-samar yang menunjukkan keberhasilan sekaligus kegagalan akibat crash dan mulai ulang, kegagalan dependensi intermiten, kapasitas yang menurun, atau hilangnya paket jaringan.

Manfaat menerapkan praktik terbaik ini: Sistem yang gagal cepat (fail fast) lebih mudah untuk di-debug dan diperbaiki, dan sering kali dapat mengekspos masalah-masalah dalam pengodean dan konfigurasi sebelum rilis dipublikasikan ke tahap produksi. Sistem yang menggabungkan strategi antrean yang efektif memberikan ketahanan dan keandalan yang lebih baik terhadap lonjakan lalu lintas dan kondisi gangguan sistem intermiten.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Tinggi

Panduan implementasi

Strategi gagal cepat (fail fast) dapat dikodekan ke dalam solusi perangkat lunak serta dikonfigurasi ke dalam infrastruktur. Selain gagal cepat (fail fast), antrean adalah teknik arsitektur yang sederhana namun ampuh untuk memisahkan komponen-komponen sistem dan memperlancar beban. Amazon CloudWatch menyediakan kemampuan untuk melakukan pemantauan dan memberikan alarm kegagalan. Setelah sistem diketahui mengalami kegagalan, strategi mitigasi dapat diinvokasi, termasuk gagal dan menjauh (fail away) dari sumber daya yang terdampak. Ketika sistem mengimplementasikan antrean dengan Amazon SQS dan teknologi antrean lainnya untuk melancarkan beban, sistem harus mempertimbangkan bagaimana ia akan mengelola backlog antrean, serta kegagalan konsumsi pesan.

Langkah-langkah implementasi

- Terapkan pernyataan terprogram atau metrik spesifik dalam perangkat lunak Anda dan gunakan itu untuk memperingatkan secara eksplisit tentang masalah sistem. Amazon CloudWatch akan membantu Anda membuat metrik dan alarm berdasarkan pola log aplikasi dan instrumentasi SDK.
- Gunakan metrik dan alarm CloudWatch untuk gagal dan menjauh dari sumber daya terdampak yang menambahkan latensi untuk pemrosesan atau berulang kali gagal memproses permintaan.
- Gunakan pemrosesan tak selaras dengan merancang desain API untuk menerima permintaan dan menambahkan permintaan ke antrean internal dengan menggunakan Amazon SQS dan kemudian menanggapi klien penghasil pesan dengan pesan keberhasilan sehingga klien dapat melepaskan sumber daya dan beralih dengan pekerjaan lain sementara konsumen antrean backend memproses permintaan.
- Lakukan pengukuran dan pemantauan terhadap latensi pemrosesan antrean dengan menghasilkan sebuah metrik CloudWatch setiap kali Anda melepaskan sebuah pesan dari antrean dengan membandingkan sekarang dengan stempel waktu pesan.
- Ketika ada kegagalan yang menghambat keberhasilan pemrosesan pesan atau terjadi lonjakan volume lalu lintas sehingga tidak dapat diproses dalam batas perjanjian tingkat layanan, sisihkan lalu lintas yang lebih lama atau berlebih ke antrean spillover. Hal ini akan memungkinkan pemrosesan yang berprioritas pada pekerjaan baru, dan pekerjaan-pekerjaan yang lebih lama ketika kapasitas tersedia. Teknik ini mirip dengan pemrosesan LIFO dan dapat memungkinkan pemrosesan sistem yang normal untuk semua pekerjaan baru.
- Gunakan antrean surat mati atau redrive untuk memindahkan pesan yang tidak dapat diproses dari backlog ke lokasi yang dapat dicari ulang dan selesaikan di lain waktu

- Coba lagi atau, apabila dapat ditoleransi, singkirkan pesan lama dengan membandingkan sekarang dengan stempel waktu pesan dan membuang pesan yang sudah tidak relevan dengan klien yang membuat permintaan.

Sumber daya

Praktik-praktik terbaik terkait:

- REL04-BP02 Menerapkan dependensi yang digabungkan secara longgar
- REL05-BP02 Membatasi (throttling) permintaan
- REL05-BP03 Kontrol dan batasi panggilan coba lagi
- REL06-BP02 Menetapkan dan menghitung metrik (Agregasi)
- REL06-BP07 Memantau pelacakan permintaan menyeluruh melalui sistem Anda

Dokumen terkait:

- Menghindari backlog antrean yang terlalu membebani
- Gagal Cepat (Fail Fast)
- Bagaimana cara mencegah peningkatan backlog pesan dalam antrean Amazon SQS saya?
- Penyeimbangan Beban Elastis: Peralihan Zona
- Pengontrol Pemulihan Aplikasi Amazon: Kontrol perutean untuk failover lalu lintas

Contoh terkait:

- Pola Integrasi Korporat: Saluran Surat Mati

Video terkait:

- AWS re:Invent 2022 - Mengoperasikan aplikasi Multi-AZ dengan ketersediaan tinggi

Alat terkait:

- Amazon SQS
- Amazon MQ

- AWS IoT Core
- Amazon CloudWatch

REL05-BP05 Mengatur batas waktu klien

Atur batas waktu secara tepat pada koneksi dan permintaan, verifikasikan waktu tersebut secara sistematis, dan jangan selalu bergantung pada nilai default karena nilai tersebut mengabaikan hal-hal spesifik tentang beban kerja.

Hasil yang Diinginkan: Batas waktu klien harus mempertimbangkan biaya untuk klien, server, dan beban kerja yang berkaitan dengan proses menunggu permintaan yang memerlukan waktu sangat lama untuk diselesaikan. Karena penyebab batas waktu tidak mungkin diketahui secara pasti, maka klien harus menggunakan pengetahuannya mengenai layanan untuk membangun ekspektasi tentang kemungkinan-kemungkinan yang menjadi penyebab dan batas waktu yang tepat

Koneksi klien mengalami habis waktu berdasarkan nilai yang dikonfigurasi. Setelah mengalami batas waktu, klien mengambil keputusan untuk menunda dan mencobanya lagi atau membuka pemutus sirkuit. Pola-pola ini akan mencegah mengeluarkan permintaan yang dapat memperburuk kondisi kesalahan yang menyebabkannya.

Anti-pola umum:

- Tidak mengetahui batas waktu sistem atau batas waktu default.
- Tidak mengetahui waktu penyelesaian permintaan normal.
- Tidak mengetahui kemungkinan-kemungkinan yang menjadi penyebab permintaan membutuhkan waktu yang terlalu lama untuk diselesaikan, atau biaya untuk klien, layanan, atau kinerja beban kerja yang berkaitan dengan proses tunggu penyelesaian ini.
- Tidak mengetahui adanya kemungkinan jaringan rusak yang menyebabkan permintaan mengalami kegagalan hanya sesaat setelah batas waktu tercapai, dan biaya untuk klien dan kinerja beban kerja karena tidak mengadopsi batas waktu yang lebih singkat.
- Tidak menguji skenario batas waktu, baik untuk koneksi maupun permintaan.
- Mengatur batas waktu terlalu tinggi, yang berimbas pada waktu tunggu menjadi lama dan meningkatkan pemanfaatan sumber daya.
- Mengatur batas waktu terlalu rendah, sehingga mengakibatkan terjadinya kegagalan buatan.
- Mengabaikan pola-pola untuk menangani kesalahan-kesalahan batas waktu untuk panggilan jarak jauh seperti pemutus sirkuit dan percobaan ulang.

- Tidak mempertimbangkan pemantauan untuk angka kesalahan panggilan layanan, target latensi di tingkat layanan, dan outlier latensi. Metrik-metrik ini dapat memberikan wawasan tentang batas waktu yang agresif atau permisif

Manfaat menerapkan praktik terbaik ini: Waktu tunggu panggilan jarak jauh dikonfigurasi dan sistem dirancang untuk menangani batas waktu secara perlahan sehingga sumber daya akan dihemat ketika panggilan jarak jauh memberikan respons yang terlalu lambat dan kesalahan batas waktu ditangani secara perlahan oleh klien layanan.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Tinggi

Panduan implementasi

Atur batas waktu koneksi dan batas waktu permintaan untuk panggilan dependensi layanan apa pun, serta secara umum untuk panggilan apa pun yang dilakukan di seluruh proses. Banyak kerangka kerja yang menawarkan kemampuan batas waktu bawaan, tetapi Anda harus tetap memperhatikan bahwa nilai default bawaan bisa saja tidak terbatas atau lebih tinggi dari nilai yang dapat diterima untuk sasaran layanan Anda. Nilai yang terlalu tinggi dapat mengurangi kegunaan waktu habis karena sumber daya akan terus terpakai saat klien menunggu terjadinya waktu habis. Nilai yang terlalu rendah dapat menyebabkan lalu lintas yang tinggi di backend serta akan meningkatkan latensi karena terlalu banyak permintaan yang dicoba ulang. Dalam beberapa kasus, hal ini dapat menyebabkan penghentian total karena semua permintaan dicoba ulang.

Pertimbangkan hal berikut saat Anda menentukan strategi batas waktu:

- Permintaan mungkin membutuhkan waktu pemrosesan yang lebih lama dari biasanya dikarenakan kontennya, terjadi gangguan pada layanan target, atau kegagalan partisi jaringan.
- Permintaan-permintaan dengan konten yang terlalu mahal dapat mengonsumsi sumber daya server dan klien yang tidak perlu. Dalam hal ini, membatasi waktu dan tidak mencoba ulang permintaan-permintaan tersebut dapat menghemat sumber daya. Layanan juga harus melindungi diri dari konten yang terlalu mahal dengan melakukan throttling dan batas waktu sisi server.
- Permintaan yang memakan waktu terlalu lama karena adanya gangguan layanan dapat diberikan batas waktu dan dicoba ulang. Pertimbangan harus diberikan pada biaya layanan untuk membuat permintaan dan melakuan percobaan ulang, tetapi jika penyebabnya adalah gangguan yang terbatas di suatu tempat, percobaan ulang kemungkinan tidak akan mahal dan akan mengurangi konsumsi sumber daya klien. Batas waktu juga dapat melepaskan sumber daya server, tergantung sifat gangguan.

- Permintaan-permintaan yang membutuhkan waktu penyelesaian yang lama karena permintaan atau respons gagal dikirimkan oleh jaringan dapat diberikan batas waktu dan dicoba ulang. Karena permintaan atau respons tidak dikirimkan, kegagalan akan terjadi, terlepas dari lamanya batas waktu yang dibreikan. Memberikan batas waktu pada kasus ini tidak akan melepaskan sumber daya server, tetapi akan melepaskan sumber daya klien dan meningkatkan kinerja beban kerja.

Manfaatkan pola desain yang sudah mapan seperti percobaan ulang dan pemutus sirkuit untuk menangani batas waktu dengan anggun dan mendukung pendekatan cepat gagal. AWS SDKsdan AWS CLImemungkinkan konfigurasi koneksi dan batas waktu permintaan dan untuk percobaan ulang dengan backoff dan jitter eksponensial. AWS Lambdafungsi mendukung konfigurasi batas waktu, dan dengan AWS Step Functions, Anda dapat membangun pemutus sirkuit kode rendah yang memanfaatkan integrasi pra-bangun dengan layanan dan. AWS SDKs AWS App Mesh Envoy memberikan kemampuan batas waktu dan pemutus sirkuit.

Langkah-langkah implementasi

- Konfigurasikan batas waktu pada panggilan layanan jarak jauh dan manfaatkan fitur batas waktu bahasa bawaan atau pustaka batas waktu sumber terbuka.
- Saat beban kerja Anda melakukan panggilan dengan AWS SDK, tinjau dokumentasi untuk konfigurasi batas waktu khusus bahasa.
  - Python
  - PHP
  - .NET
  - Ruby
  - Java
  - Go
  - Node.js
  - C++
- Saat menggunakan AWS SDKs atau AWS CLI perintah di beban kerja Anda, konfigurasikan nilai batas waktu default dengan menyetel default AWS konfigurasi untuk dan. `connectTimeoutInMillis` `tlsNegotiationTimeoutInMillis`
- Terapkan opsi baris perintah `cli-connect-timeout` dan `cli-read-timeout` untuk mengontrol AWS CLI perintah satu kali ke AWS layanan.

- Lakukan pemantauan panggilan layanan jarak jauh untuk memeriksa batas waktu, dan atur alarm pada kesalahan persisten sehingga Anda dapat menangani skenario kesalahan secara proaktif.
- Menerapkan CloudWatch Metrik dan deteksi CloudWatch anomali pada tingkat kesalahan panggilan, tujuan tingkat layanan untuk latensi, dan latensi outlier untuk memberikan wawasan tentang mengelola batas waktu yang terlalu agresif atau permisif.
- Konfigurasikan batas waktu pada fungsi Lambda.
- APIKlien gateway harus menerapkan percobaan ulang mereka sendiri saat menangani batas waktu. APIGateway mendukung batas waktu integrasi 50 milidetik hingga 29 detik untuk integrasi hilir dan tidak mencoba lagi saat integrasi meminta batas waktu.
- Implementasikan pemutus sirkuit untuk menghindari pembuatan panggilan jarak jauh ketika waktu habis. Buka sirkuit untuk menghindari kegagalan panggilan dan tutup sirkuit saat panggilan memberikan respons secara normal.
- Untuk beban kerja berbasis kontainer, tinjau fitur App Mesh Envoy untuk memanfaatkan batas waktu bawaan dan pemutus sirkuit.
- Gunakan AWS Step Functions untuk membangun pemutus sirkuit kode rendah untuk panggilan layanan jarak jauh, terutama saat memanggil integrasi Step Functions AWS asli SDKs dan didukung untuk menyederhanakan beban kerja Anda.

Sumber daya

Praktik-praktik terbaik terkait:

- REL05-BP03 Kontrol dan batasi panggilan coba lagi
- REL05-BP04 Melakukan gagal cepat (fail fast) dan membatasi antrean
- REL06-BP07 Memantau pelacakan permintaan menyeluruh melalui sistem Anda

Dokumen terkait:

- AWS SDK: Coba Ulang dan Batas Waktu
- Amazon Builders' Library: Batas waktu, percobaan ulang, dan penundaan dengan jitter
- Kuota Amazon API Gateway dan catatan penting
- AWS Command Line Interface: Opsi baris perintah
- AWS SDK for Java 2.x: Konfigurasikan Batas API Waktu

- AWS Botocore menggunakan objek konfigurasi dan Referensi Config
- AWS SDK untuk .NET: Percobaan Ulang dan Batas Waktu
- AWS Lambda: Mengonfigurasi opsi fungsi Lambda

Contoh terkait:

- Menggunakan pola pemutus sirkuit dengan AWS Step Functions dan Amazon DynamoDB
- Martin Fowler: CircuitBreaker

Alat terkait:

- AWS SDKs
- AWS Lambda
- Amazon SQS
- AWS Step Functions
- AWS Command Line Interface

REL05-BP06 Membuat sistem tanpa kewarganegaraan jika memungkinkan

Sistem seharusnya tidak memerlukan state, atau seharusnya mengalihkan state sedemikian rupa sehingga di antara permintaan klien yang berbeda, tidak ada dependensi di penyimpanan data lokal di disk dan memori. Hal ini membuat server dapat diganti sesuai keinginan tanpa menimbulkan dampak ketersediaan.

Ketika pengguna atau layanan berinteraksi dengan sebuah aplikasi, mereka sering kali melakukan serangkaian interaksi yang membentuk sebuah sesi. Sesi adalah data unik bagi para pengguna yang lama berada di antara permintaan ketika mereka menggunakan aplikasi. Aplikasi stateless adalah sebuah aplikasi yang tidak memerlukan pengetahuan tentang interaksi sebelumnya dan tidak menyimpan informasi sesi.

Setelah dirancang untuk menjadi stateless, Anda kemudian dapat menggunakan layanan komputasi tanpa server, seperti atau. AWS Lambda AWS Fargate

Selain penggantian server, manfaat lain dari aplikasi stateless adalah mereka dapat menskalakan secara horizontal karena salah satu sumber daya komputasi yang tersedia (seperti EC2 instance dan AWS Lambda fungsi) dapat melayani permintaan apa pun.

Manfaat menerapkan praktik terbaik ini: Sistem yang dirancang untuk menjadi stateless lebih mudah beradaptasi dengan penskalaan horizontal, sehingga akan memungkinkan Anda untuk menambah atau menghapus kapasitas berdasarkan lalu lintas dan permintaan yang berfluktuasi. Sistem ini juga memiliki sifat yang tahan terhadap kegagalan-kegagalan yang terjadi dan memberikan fleksibilitas serta ketangkasan dalam pengembangan aplikasi.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Sedang

Panduan implementasi

Jadikan aplikasi Anda stateless. Aplikasi stateless memungkinkan penskalaan horizontal dan toleran terhadap kegagalan yang dialami simpul individual. Lakukan analisis terhadap dan pahami komponen-komponen aplikasi Anda yang mempertahankan state di dalam arsitektur. Hal ini akan membantu Anda menilai dampak potensial dari melakukan transisi ke desain stateless. Arsitektur stateless memisahkan data pengguna dan memindahkan data sesi. Hal ini dapat memberikan fleksibilitas untuk menskalakan setiap komponen secara independen untuk memenuhi permintaan beban kerja yang beragam dan mengoptimalkan pemanfaatan sumber daya.

Langkah-langkah implementasi

- Identifikasi dan pahami komponen-komponen stateful yang ada dalam aplikasi Anda.
- Pisahkan data dengan memisahkan dan mengelola data pengguna dari logika aplikasi inti.
  - Amazon Cognito dapat memisahkan data pengguna dari kode aplikasi dengan menggunakan fitur, seperti kolam identitas, kumpulan pengguna, dan Amazon Cognito Sync.
  - Anda dapat menggunakan AWS Secrets Manager untuk memisahkan data pengguna dengan menyimpan rahasia di lokasi yang aman dan tersentralisasi. Ini berarti kode aplikasi tidak perlu menyimpan rahasia, sehingga membuatnya menjadi lebih aman.
  - Pertimbangkan untuk menggunakan Amazon S3 untuk menyimpan data besar dan tidak terstruktur, seperti gambar dan dokumen. Aplikasi Anda dapat mengambil data tersebut saat diperlukan, sehingga akan menghilangkan kebutuhan untuk menyimpannya di dalam memori.
  - Gunakan Amazon DynamoDB untuk menyimpan informasi seperti profil pengguna. Aplikasi Anda dapat melakukan kueri terhadap data tersebut hampir dalam waktu nyata.
- Pindahkan data sesi ke basis data, cache, atau file eksternal.
  - Amazon ElastiCache, Amazon DynamoDB, Amazon Elastic File System (Amazon), dan EFS Amazon MemoryDB adalah contoh layanan yang dapat Anda AWS gunakan untuk membongkar data sesi.

- Rancang sebuah arsitektur stateless setelah Anda mengidentifikasi state dan data pengguna mana yang perlu dipertahankan dengan solusi penyimpanan pilihan Anda.

Sumber daya

Praktik-praktik terbaik terkait:

- REL11-BP03 Mengotomatiskan penyembuhan pada semua lapisan

Dokumen terkait:

- Amazon Builders' Library: Menghindari fallback dalam sistem terdistribusi
- Amazon Builders' Library: Menghindari backlog antrean yang tidak dapat diatasi
- Amazon Builders' Library: Tantangan dan strategi caching
- Praktik Terbaik untuk Tingkat Web Stateless di AWS

REL05-BP07 Menerapkan tuas darurat

Tuas darurat adalah proses cepat yang dapat memitigasi dampak ketersediaan pada beban kerja.

Tuas darurat bekerja dengan cara menonaktifkan, melakukan throttling, atau mengubah perilaku komponen atau dependensi dengan menggunakan mekanisme yang diketahui dan sudah diuji. Hal ini dapat mengurangi gangguan beban kerja yang disebabkan oleh kelelahan sumber daya karena peningkatan permintaan yang terjadi secara tidak terduga dan mengurangi dampak kegagalan pada komponen non-kritis yang ada dalam beban kerja Anda.

Hasil yang diinginkan: Dengan menerapkan tuas-tuas darurat, Anda dapat menetapkan proses yang diketahui baik untuk menjaga ketersediaan komponen penting dalam beban kerja Anda. Beban kerja akan mengalami degradasi secara perlahan (graceful degradation) dan terus menjalankan fungsi-fungsi kritis bisnisnya selama tuas darurat masih dalam keadaan aktif. Untuk detail lebih lanjut tentang degradasi anggun, lihat REL05-BP01 Menerapkan degradasi anggun untuk mengubah dependensi keras yang berlaku menjadi dependensi lunak.

Anti-pola umum:

- Kegagalan dependensi non-kritis akan berdampak pada ketersediaan beban kerja inti Anda.
- Tidak menguji atau memverifikasi perilaku komponen-komponen kritis saat terjadi gangguan komponen non-kritis.

- Tidak ada kriteria yang jelas dan deterministik yang ditentukan untuk pengaktifan atau penonaktifan sebuah tuas darurat.

Manfaat menerapkan praktik terbaik ini: Menerapkan tuas darurat dapat meningkatkan ketersediaan komponen penting dalam beban kerja Anda dengan menyediakan resolver Anda proses-proses yang telah ditetapkan untuk menanggapi lonjakan permintaan yang tidak terduga atau kegagalan dependensi non-kritis.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Sedang

Panduan implementasi

- Identifikasi komponen-komponen kritis yang ada dalam beban kerja Anda.
- Buat agar rancangan dan arsitek komponen kritis dalam beban kerja Anda dapat menahan kegagalan komponen non-kritis.
- Lakukan pengujian untuk memvalidasi perilaku komponen-komponen kritis Anda saat terjadi kegagalan komponen non-kritis.
- Tentukan dan pantau metrik atau pemicu yang relevan untuk memulai prosedur tuas darurat.
- Tentukan prosedur (manual atau otomatis) yang mencakup tuas darurat.

Langkah-langkah implementasi

- Identifikasi komponen-komponen kritis bagi bisnis yang ada dalam beban kerja Anda.
  - Setiap komponen teknis yang ada dalam beban kerja Anda harus dipetakan ke fungsi bisnisnya yang relevan dan diberi peringkat sebagai komponen kritis atau non-kritis. Untuk contoh fungsionalitas kritis dan non-kritis yang ada di Amazon, silakan lihat Kapan Saja Bisa Menjadi Hari yang Prima: Bagaimana Penelusuran Pencarian Amazon.com Menggunakan Perekayasaan Chaos untuk Menangani Lebih dari 84K Permintaan Per Detik.
  - Ini merupakan keputusan teknis sekaligus keputusan bisnis, dan berbeda-beda berdasarkan organisasi dan beban kerja.
- Buat agar rancangan dan arsitek komponen kritis dalam beban kerja Anda dapat menahan kegagalan komponen non-kritis.
  - Saat melakukan analisis dependensi, pertimbangkan semua mode kegagalan yang dapat terjadi, dan pastikan bahwa mekanisme tuas darurat Anda memberikan fungsionalitas kritis pada komponen-komponen hilir.

- Lakukan pengujian untuk melakukan validasi terhadap perilaku komponen kritis Anda saat tuas darurat Anda diaktifkan.
  - Hindari perilaku bimodal. Untuk detail lebih lanjut, lihat REL11-BP05 Gunakan stabilitas statis untuk mencegah perilaku bimodal.
- Tentukan, pantau, dan munculkan peringatan pada metrik-metrik yang relevan untuk memulai prosedur tuas darurat.
  - Beban kerja Anda menentukan metrik yang tepat untuk dipantau. Beberapa contoh metrik adalah latensi atau jumlah permintaan yang gagal ke sebuah dependensi.
- Tentukan prosedur yang mencakup tuas darurat, bisa manual atau otomatis.
  - Ini mungkin termasuk mekanisme seperti memuat pelepasan, permintaan throttling, atau melaksanakan degradasi yang tepat.

Sumber daya

Praktik-praktik terbaik terkait:

- REL05-BP01 Menerapkan degradasi anggun untuk mengubah dependensi keras yang berlaku menjadi dependensi lunak
- REL05-BP02 Permintaan Throttle
- REL11-BP05 Gunakan stabilitas statis untuk mencegah perilaku bimodal

Dokumen terkait:

- Melakukan otomatisasi deployment secara aman dan otonom
- Kapan Saja Bisa Menjadi Hari yang Prima: Bagaimana Penelusuran Pencarian Amazon.com Menggunakan Perekayasaan Chaos untuk Menangani Lebih dari 84K Permintaan Per Detik

Video terkait:

- AWS Re: Invent 2020: Keandalan, konsistensi, dan kepercayaan diri melalui kekekalan

Manajemen perubahan

Perubahan beban kerja atau lingkungannya harus diantisipasi dan diakomodasi guna mencapai operasi beban kerja yang andal. Perubahan mencakup semua yang terjadi ke beban kerja seperti lonjakan permintaan, serta perubahan dari dalam, seperti deployment fitur dan patch keamanan.

Praktik terbaik untuk manajemen perubahan dijelaskan dalam bagian-bagian berikut ini.

Topik

Memantau sumber daya beban kerja

Log dan metrik merupakan alat yang luar biasa untuk mendapatkan wawasan mengenai kondisi beban kerja Anda. Anda dapat mengonfigurasikan beban kerja Anda untuk memantau log dan metrik serta mengirimkan notifikasi ketika ambang batas terlampaui atau ada peristiwa signifikan yang terjadi. Pemantauan memungkinkan beban kerja Anda untuk mengenali ketika ambang batas kinerja rendah terlampaui atau ada kegagalan yang terjadi, sehingga pemulihan dapat terjadi secara otomatis untuk menanggapinya.

Pemantauan sangat penting untuk memastikan Anda memenuhi persyaratan ketersediaan. Pemantauan harus mendeteksi kegagalan secara efektif. Mode kegagalan terburuk adalah kegagalan “senyap”, yaitu saat fungsionalitas tidak lagi bekerja, tetapi kegagalan itu tidak dapat dideteksi secara langsung. Pelanggan mengetahuinya lebih dulu dari Anda. Salah satu alasan utama pemantauan adalah untuk memperingatkan saat ada masalah. Peringatan harus dipisahkan dari sistem sebanyak mungkin. Jika gangguan layanan menghapus kemampuan untuk memberikan peringatan, Anda akan mengalami gangguan lebih lama.

Di AWS, kami melengkapi aplikasi pada berbagai tingkat. Kami mencatat latensi, tingkat kesalahan, dan ketersediaan untuk semua permintaan, dependensi, serta ketersediaan untuk operasi utama yang ada dalam proses. Kami juga mencatat metrik operasi yang berhasil. Dengan begitu, kami dapat melihat potensi masalah sebelum terjadi. Kami tidak hanya mengamati latensi rata-rata. Kami bahkan lebih fokus pada outlier latensi, seperti persentil 99,9 dan 99,99. Karena jika satu permintaan dari 1.000 atau 10.000 lambat, hal ini masih termasuk pengalaman yang buruk. Selain itu, meski rata-

ratanya dapat diterima, jika ada satu dari 100 permintaan yang menyebabkan latensi ekstrem, maka hal ini nantinya dapat menjadi masalah seiring dengan meningkatnya lalu lintas Anda.

Pemantauan di AWS terdiri dari empat fase berbeda:

1. Pembuatan — Memantau semua komponen untuk beban kerja
2. Agregasi — Menentukan dan mengalkulasi metrik
3. Pemberian peringatan dan pemrosesan waktu nyata — Mengirimkan notifikasi dan mengotomatiskan respons
4. Penyimpanan dan Analitik

Praktik terbaik

- REL06-BP01 Memantau semua komponen untuk beban kerja (Pembuatan)
- REL06-BP02 Menetapkan dan menghitung metrik (Agregasi)
- REL06-BP03 Mengirimkan notifikasi (Pemrosesan dan pembuatan alarm waktu nyata)
- REL06-BP04 Mengotomatiskan respons (Peringatan dan pemrosesan waktu nyata)
- REL06-BP05 Menganalisis log
- REL06-BP06 Meninjau cakupan dan metrik pemantauan secara berkala
- REL06-BP07 Memantau pelacakan permintaan menyeluruh melalui sistem Anda

REL06-BP01 Memantau semua komponen untuk beban kerja (Pembuatan)

Pantau komponen beban kerja dengan Amazon CloudWatch atau alat pihak ketiga. Pantau layanan AWS dengan Dasbor AWS Health.

Semua komponen beban kerja Anda harus dipantau, termasuk front-end, logika bisnis, dan tingkat penyimpanan. Tetapkan metrik utama, jelaskan cara mengekstraknya dari log (jika diperlukan), dan tetapkan ambang batas untuk menginvokasi peristiwa alarm yang sesuai. Pastikan bahwa metrik-metrik relevan dengan indikator kinerja utama (KPI) beban kerja Anda, dan gunakan metrik dan log untuk mengidentifikasi tanda-tanda peringatan dini terjadinya penurunan layanan. Contohnya, sebuah metrik yang terkait dengan hasil bisnis, seperti jumlah pesanan yang berhasil diproses per menit, dapat menunjukkan masalah beban kerja lebih cepat dari metrik teknis, seperti metrik Pemanfaatan CPU. Gunakan Dasbor AWS Health untuk mendapatkan tampilan yang dipersonalisasi tentang kinerja dan ketersediaan layanan AWS yang mendasari sumber daya AWS Anda.

Pemantauan di cloud menawarkan peluang-peluang baru. Sebagian besar penyedia cloud telah mengembangkan hook yang dapat disesuaikan dan dapat menghadirkan wawasan untuk membantu Anda memantau beberapa lapisan beban kerja Anda. Layanan-layanan AWS seperti Amazon CloudWatch menerapkan algoritme statis dan machine learning untuk terus menganalisis metrik sistem dan aplikasi, menentukan garis dasar normal dan anomali permukaan dengan sedikit campur tangan pengguna. Algoritma deteksi anomali bertanggung jawab atas perubahan-perubahan musiman dan tren metrik.

AWS menyediakan banyak informasi pemantauan dan log yang bisa digunakan untuk menentukan metrik khusus beban kerja, proses perubahan permintaan, dan untuk mengadopsi teknik machine learning, terlepas dari keahlian ML.

Selain itu, pantau semua titik akhir eksternal Anda untuk memastikan bahwa mereka tidak bergantung pada implementasi dasar Anda. Pemantauan aktif ini dapat dilakukan dengan transaksi sintetis (kadang-kadang disebut sebagai canary pengguna, tetapi jangan disamakan dengan deployment canary) yang secara berkala menjalankan sejumlah tindakan yang cocok dengan tugas-tugas umum yang dilakukan oleh klien dari beban kerja. Buat tugas-tugas ini berdurasi singkat dan pastikan untuk tidak membebani beban kerja Anda saat melakukan pengujian. Amazon CloudWatch Synthetics memungkinkan Anda membuat canary sintetis untuk Anda melakukan pemantauan terhadap titik akhir dan API Anda. Anda juga dapat menggabungkan simpul-simpul klien canary sintetis dengan konsol AWS X-Ray untuk mengidentifikasi canary sintetis mana yang mengalami masalah berupa error, fault, atau tingkat throttling untuk jangka waktu yang dipilih.

Hasil yang Diinginkan:

Kumpulkan dan gunakan metrik-metrik kritis dari semua komponen beban kerja untuk memastikan keandalan beban kerja dan pengalaman pengguna yang optimal. Dengan mendeteksi bahwa sebuah beban kerja tidak mencapai hasil bisnis, Anda dapat dengan cepat mengumumkan terjadinya sebuah bencana dan kemudian segera melakukan pemulihan dari insiden.

Anti-pola umum:

- Melakukan pemantauan hanya untuk antarmuka eksternal beban kerja Anda.
- Tidak membuat metrik khusus beban kerja dan hanya mengandalkan metrik-metrik yang diberikan kepada Anda oleh layanan AWS yang digunakan oleh beban kerja Anda.
- Hanya dengan menggunakan metrik teknis yang ada di beban kerja Anda dan tidak memantau metrik apa pun yang terkait dengan KPI non-teknis yang menerima kontribusi dari beban kerja Anda.

- Mengandalkan lalu lintas produksi dan pemeriksaan kondisi sederhana untuk melakukan pemantauan dan evaluasi terhadap status (state) beban kerja.

Manfaat menerapkan praktik terbaik ini: Dengan memantau semua tingkatan di beban kerja Anda akan memungkinkan Anda untuk dapat lebih cepat mengantisipasi dan menyelesaikan masalah di komponen dalam beban kerja.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Tinggi

Panduan implementasi

1. Aktifkan pencatatan log jika tersedia. Data pemantauan harus diperoleh dari semua komponen beban kerja. Aktifkan pencatatan log tambahan, seperti S3 Access Logs, dan mengizinkan beban kerja Anda untuk mencatat log khusus beban kerja. Kumpulkan metrik untuk CPU, I/O jaringan, dan rata-rata I/O diska dari layanan-layanan seperti Amazon ECS, Amazon EKS, Amazon EC2, Penyeimbangan Beban Elastis, AWS Auto Scaling, dan Amazon EMR. Lihat Layanan AWS yang Memublikasikan Metrik CloudWatch untuk mendapatkan daftar layanan AWS yang memublikasikan metrik ke CloudWatch.
2. Tinjau semua metrik default dan telusuri celah pengumpulan data apa pun. Setiap layanan menghasilkan metrik default. Dengan mengumpulkan metrik default, Anda dapat lebih memahami dependensi yang terjadi antar komponen beban kerja dan bagaimana keandalan dan kinerja komponen memengaruhi beban kerja tersebut. Anda juga dapat membuat dan menerbitkan metrik Anda sendiri ke CloudWatch dengan menggunakan AWS CLI atau API.
3. Lakukan evaluasi terhadap semua metrik untuk menentukan mana yang harus dibuatkan peringatan untuk setiap layanan AWS di beban kerja Anda. Anda dapat memilih subset metrik yang memiliki dampak besar terhadap keandalan beban kerja. Berfokus pada metrik-metrik dan ambang batas kritis akan memungkinkan Anda untuk menyempurnakan jumlah peringatan dan dapat membantu Anda meminimalkan positif palsu.
4. Tetapkan peringatan dan proses pemulihan beban kerja Anda setelah peringatan diinvokasi. Dengan menetapkan peringatan, Anda dapat dengan cepat memberikan notifikasi, melakukan eskalasi, dan mengikuti langkah-langkah yang diperlukan untuk melakukan pemulihan dari insiden dan memenuhi Sasaran Waktu Pemulihan (RTO) yang Anda tentukan. Anda dapat menggunakan Amazon CloudWatch Alarms untuk menginvokasi alur kerja otomatis dan memulai prosedur pemulihan berdasarkan ambang batas yang ditentukan.
5. Jelajahi penggunaan transaksi sintetis untuk mengumpulkan data yang relevan tentang state beban kerja. Pemantauan sintetis mengikuti rute yang sama dan menjalankan tindakan-tindakan

yang sama seperti seorang pelanggan, sehingga memungkinkan Anda untuk terus memverifikasi pengalaman pelanggan Anda bahkan saat Anda tidak memiliki lalu lintas pelanggan apa pun pada beban kerja Anda. Dengan menggunakan transaksi sintetis, Anda dapat menemukan masalah-masalah sebelum para pelanggan Anda menemukannya.

Sumber daya

Praktik-praktik terbaik terkait:

- REL11-BP03 Melakukan otomatisasi pemulihan di semua lapisan

Dokumen terkait:

- Memulai Dasbor AWS Health Anda – Kesehatan akun Anda
- Layanan AWS yang Menerbitkan Metrik CloudWatch
- Pencatatan Log Akses Untuk Penyeimbang Beban Jaringan Anda
- Pencatatan Log Akses untuk penyeimbang beban aplikasi Anda
- Mengakses Log Amazon CloudWatch untuk AWS Lambda
- Pencatatan Log Akses Server Amazon S3
- Mengaktifkan Pencatatan Log Akses untuk Penyeimbang Beban Klasik Anda
- Mengekspor data log ke Amazon S3
- Menginstal agen CloudWatch pada instans Amazon EC2
- Memublikasikan Metrik Kustom
- Menggunakan Dasbor Amazon CloudWatch
- Menggunakan Metrik Amazon CloudWatch
- Menggunakan Canary (Amazon CloudWatch Synthetics)
- Apa yang dimaksud dengan Log Amazon CloudWatch?

Panduan pengguna:

- Membuat jejak
- Memantau memori dan metrik disk untuk instans Linux Amazon EC2
- Menggunakan CloudWatch Logs dengan instans kontainer

- Log Alur VPC
- Apa itu Amazon DevOps Guru?
- Apa itu AWS X-Ray?

Blog terkait:

- Melakukan Debugging dengan Amazon CloudWatch Synthetics dan AWS X-Ray

Contoh terkait:

- Amazon Builders' Library: Instrumentasi sistem terdistribusi untuk visibilitas operasional
- Lokakarya observabilitas

REL06-BP02 Menetapkan dan menghitung metrik (Agregasi)

Kumpulkan metrik dan log dari komponen beban kerja Anda dan hitung metrik agregat yang relevan dari metrik dan log tersebut. Metrik ini memberikan observabilitas yang luas dan mendalam terhadap beban kerja Anda dan dapat meningkatkan postur ketahanan Anda secara signifikan.

Observabilitas bukan hanya sekadar mengumpulkan metrik dari komponen beban kerja dan dapat melihat serta memberikan peringatan tentangnya. Tujuannya adalah memiliki pemahaman holistik atas perilaku beban kerja Anda. Informasi perilaku ini berasal dari semua komponen dalam beban kerja Anda, yang mencakup layanan cloud yang diandalkannya, log yang dibuat dengan baik, dan metrik. Data ini memberi Anda pengawasan atas perilaku beban kerja Anda secara keseluruhan, serta pemahaman tentang interaksi setiap komponen dengan setiap unit kerja pada tingkat detail yang terperinci.

Hasil yang diinginkan:

- Anda mengumpulkan log dari komponen beban kerja dan dependensi layanan AWS Anda, dan Anda menerbitkannya ke lokasi pusat yang memudahkannya diakses dan diproses.
- Log Anda berisi stempel waktu dengan akurasi dan fidelitas tinggi.
- Log Anda berisi informasi yang relevan tentang konteks pemrosesan, seperti pengidentifikasi jejak, pengidentifikasi pengguna atau akun, dan alamat IP jarak jauh.
- Anda membuat metrik agregat dari log yang merepresentasikan perilaku beban kerja Anda dari perspektif tingkat tinggi.

- Anda dapat melakukan kueri pada log agregat Anda untuk mendapatkan wawasan yang mendalam dan relevan tentang beban kerja Anda dan mengidentifikasi masalah yang sedang atau mungkin terjadi.

Anti-pola umum:

- Anda tidak mengumpulkan log atau metrik yang relevan dari instans komputasi yang dijalankan beban kerja Anda atau layanan cloud yang digunakan beban kerja Anda.
- Anda mengabaikan kumpulan log dan metrik yang terkait dengan indikator kinerja utama (KPI) bisnis Anda.
- Anda menganalisis telemetri terkait beban kerja secara terpisah tanpa agregasi dan korelasi.
- Anda membiarkan metrik dan log kedaluwarsa terlalu cepat, sehingga menghambat analisis tren dan identifikasi masalah berulang.

Manfaat menerapkan praktik terbaik ini: Anda dapat mendeteksi lebih banyak anomali serta mengorelasikan peristiwa dan metrik di antara berbagai komponen beban kerja Anda. Anda dapat membuat wawasan dari komponen beban kerja Anda berdasarkan informasi yang terdapat dalam log yang sering kali tidak tersedia dalam metrik saja. Anda dapat menentukan penyebab kegagalan lebih cepat dengan melakukan kueri pada log Anda dalam skala besar.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Tinggi

Panduan implementasi

Identifikasi sumber data telemetri yang relevan dengan beban kerja Anda dan komponennya. Data ini tidak hanya berasal dari komponen yang menerbitkan metrik, seperti sistem operasi (OS) dan runtime aplikasi seperti Java, tetapi juga dari log aplikasi dan layanan cloud. Misalnya, server web biasanya mencatat log setiap permintaan dengan informasi terperinci seperti stempel waktu, latensi pemrosesan, ID pengguna, alamat IP jarak jauh, jalur, dan string kueri. Tingkat detail dalam log ini membantu Anda melakukan kueri terperinci dan menghasilkan metrik yang mungkin tidak tersedia sebelumnya.

Kumpulkan metrik dan log menggunakan alat dan proses yang sesuai. Log yang dihasilkan oleh aplikasi yang berjalan di instans Amazon EC2 dapat dikumpulkan oleh agen seperti Agen Amazon CloudWatch dan diterbitkan ke layanan penyimpanan pusat seperti Log Amazon CloudWatch. Layanan komputasi yang dikelola AWS seperti AWS Lambda dan Amazon Elastic Container Service menerbitkan log ke Log CloudWatch untuk Anda secara otomatis. Aktifkan pengumpulan log untuk

layanan penyimpanan dan pemrosesan AWS yang digunakan oleh beban kerja Anda seperti Amazon CloudFront, Amazon S3, Elastic Load Balancing, dan Amazon API Gateway.

Perkaya data telemetri Anda dengan dimensi yang dapat membantu Anda melihat pola perilaku secara lebih jelas dan mengisolasi masalah yang terkorelasi dengan grup komponen terkait. Setelah ditambahkan, Anda dapat mengamati perilaku komponen pada tingkat detail yang lebih baik, mendeteksi kegagalan yang terkorelasi, dan mengambil langkah-langkah perbaikan yang tepat. Contoh dimensi yang berguna termasuk Zona Ketersediaan, ID instans EC2, dan tugas kontainer atau ID Pod.

Setelah mengumpulkan metrik dan log, Anda dapat menulis kueri dan membuat metrik agregat dari metrik dan log yang memberikan wawasan berguna tentang perilaku normal dan anomali. Misalnya, Anda dapat menggunakan Wawasan Log Amazon CloudWatch untuk memperoleh metrik kustom dari log aplikasi, Wawasan Metrik Amazon CloudWatch untuk melakukan kueri pada metrik Anda dalam skala besar, Wawasan Kontainer Amazon CloudWatch untuk mengumpulkan, menggabungkan, serta meringkas metrik dan log dari aplikasi dan layanan mikro kontainer Anda, atau Wawasan Lambda Amazon CloudWatch jika Anda menggunakan fungsi AWS Lambda. Untuk membuat metrik tingkat kesalahan agregat, Anda dapat menambah counter setiap kali respons kesalahan atau pesan ditemukan di log komponen Anda, atau menghitung nilai agregat dari metrik tingkat kesalahan yang ada. Anda dapat menggunakan data ini untuk menghasilkan histogram yang menunjukkan perilaku turunan, seperti permintaan atau proses berkinerja terburuk. Anda juga dapat memindai data ini secara real time untuk mencari pola anomali menggunakan solusi seperti deteksi anomali Log CloudWatch. Wawasan ini dapat ditempatkan di dasbor agar tetap tertata sesuai dengan kebutuhan dan preferensi Anda.

Melakukan kueri pada log dapat membantu Anda memahami bagaimana permintaan spesifik ditangani oleh komponen beban kerja Anda dan mengungkapkan pola permintaan atau konteks lain yang berdampak pada ketahanan beban kerja Anda. Sebaiknya Anda meneliti dan menyiapkan kueri terlebih dahulu, berdasarkan pengetahuan Anda tentang bagaimana aplikasi Anda dan komponen lainnya berperilaku, sehingga Anda dapat lebih mudah menjalankannya sesuai kebutuhan. Misalnya, dengan Wawasan Log CloudWatch, Anda dapat secara interaktif mencari dan menganalisis data log Anda yang disimpan di Log CloudWatch. Anda juga dapat menggunakan Amazon Athena untuk melakukan kueri pada log dari berbagai sumber, termasuk banyak layanan AWS, pada skala petabyte.

Saat Anda menentukan kebijakan penyimpanan log, pertimbangkan nilai log historis. Log historis dapat membantu mengidentifikasi penggunaan jangka panjang dan pola perilaku, regresi, serta peningkatan pada kinerja beban kerja Anda. Log yang dihapus secara permanen tidak dapat

dianalisis nantinya. Namun, nilai log historis cenderung berkurang seiring berjalannya waktu. Pilih kebijakan yang menyeimbangkan kebutuhan Anda dengan tepat serta sesuai dengan persyaratan hukum atau kontrak apa pun yang mungkin berlaku untuk Anda.

Langkah-langkah implementasi

1. Pilih mekanisme pengumpulan, penyimpanan, analisis, dan tampilan untuk data observabilitas Anda.
2. Instal dan konfigurasikan pengumpul metrik dan log pada komponen yang sesuai dari beban kerja Anda (misalnya, pada instans Amazon EC2 dan dalam kontainer sidecar). Konfigurasikan pengumpul ini agar dimulai ulang secara otomatis jika tiba-tiba berhenti. Aktifkan buffering disk atau memori untuk pengumpul ini sehingga kegagalan penerbitan sementara tidak memengaruhi aplikasi Anda atau mengakibatkan hilangnya data.
3. Aktifkan pembuatan log pada layanan AWS yang Anda gunakan sebagai bagian dari beban kerja Anda, dan teruskan log tersebut ke layanan penyimpanan yang Anda pilih jika diperlukan. Lihat panduan pengguna atau developer layanan masing-masing untuk petunjuk yang mendetail.
4. Tentukan metrik operasional yang relevan dengan beban kerja Anda yang didasarkan pada data telemetri Anda. Hal ini dapat didasarkan pada metrik langsung yang dihasilkan dari komponen beban kerja Anda, yang dapat mencakup metrik terkait KPI bisnis, atau hasil perhitungan agregat seperti jumlah, laju, persentil, atau histogram. Hitung metrik ini menggunakan penganalisis log Anda, dan letakkan di dasbor yang sesuai.
5. Siapkan kueri log yang sesuai untuk menganalisis komponen beban kerja, permintaan, atau perilaku transaksi sesuai kebutuhan.
6. Tentukan dan aktifkan kebijakan penyimpanan log untuk log komponen Anda. Hapus log secara berkala ketika melebihi batas waktu yang diizinkan kebijakan.

Sumber daya

Praktik-praktik terbaik terkait:

- REL06-BP01 Memantau semua komponen untuk beban kerja (Pembuatan)
- REL06-BP03 Mengirimkan notifikasi (Pemrosesan dan pembuatan alarm waktu nyata)
- REL06-BP04 Mengotomatiskan respons (Pemrosesan dan pembuatan alarm waktu nyata)
- REL06-BP05 Menganalisis log
- REL06-BP06 Meninjau cakupan dan metrik pemantauan secara berkala
- REL06-BP07 Memantau pelacakan permintaan menyeluruh melalui sistem Anda

Dokumentasi terkait:

- Cara kerja Amazon CloudWatch
- Amazon Managed Prometheus
- Amazon Managed Grafana
- Menganalisis data log dengan Wawasan Log CloudWatch
- Wawasan Lambda Amazon CloudWatch
- Wawasan Kontainer Amazon CloudWatch
- Buat kueri metrik Anda dengan Wawasan Metrik CloudWatch
- AWS Distro for Open Telemetry
- Kueri Contoh Wawasan Log Amazon CloudWatch
- Melakukan Debugging dengan Amazon CloudWatch Synthetics dan AWS X-Ray
- Mencari dan Menyaring Data Log
- Mengirim Log Langsung ke Amazon S3
- Amazon Builders' Library: Instrumentasi sistem terdistribusi untuk visibilitas operasional

Lokakarya terkait:

- One Observability Workshop

Alat terkait:

- AWS Distro for OpenTelemetry (GitHub)

REL06-BP03 Mengirimkan notifikasi (Pemrosesan dan pembuatan alarm waktu nyata)

Ketika organisasi mendeteksi potensi masalah, mereka mengirimkan notifikasi dan peringatan waktu nyata kepada personel dan sistem yang sesuai untuk merespons masalah ini dengan cepat dan efektif.

Hasil yang diinginkan: Respons yang cepat terhadap peristiwa operasional dapat terjadi melalui konfigurasi alarm yang relevan berdasarkan metrik layanan dan aplikasi. Ketika ambang batas

alarm dilanggar, personel dan sistem yang sesuai mendapatkan notifikasi sehingga mereka dapat mengatasi masalah-masalah yang mendasarinya.

Anti-pola umum:

- Mengonfigurasi alarm dengan ambang batas yang terlalu tinggi, akan mengakibatkan kegagalan untuk mengirim notifikasi-notifikasi penting.
- Mengonfigurasi alarm dengan ambang batas yang terlalu rendah, akan menyebabkan tidak adanya tindakan atas notifikasi-notifikasi penting karena kebisingan notifikasi yang berlebihan.
- Tidak memperbarui alarm dan ambang batasnya saat penggunaan berubah.
- Untuk alarm yang paling sesuai untuk ditangani melalui tindakan otomatis, mengirim notifikasi ke personel alih-alih membuat tindakan otomatis, akan menyebabkan terjadinya pengiriman notifikasi yang berlebihan.

Manfaat menerapkan praktik terbaik ini: Mengirimkan notifikasi dan pemberitahuan waktu nyata kepada personel dan sistem yang sesuai akan memungkinkan dilakukannya deteksi dini terhadap masalah dan memungkinkan respons yang cepat terhadap insiden operasional.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Tinggi

Panduan implementasi

Beban kerja harus dilengkapi dengan pemrosesan dan peringatan alarm waktu nyata untuk meningkatkan pendeteksian masalah yang dapat memengaruhi ketersediaan aplikasi dan berfungsi sebagai pemicu respons otomatis. Organisasi dapat melakukan pemrosesan dan peringatan alarm waktu nyata dengan menciptakan peringatan dengan metrik yang ditentukan untuk menerima notifikasi setiap kali peristiwa signifikan terjadi atau sebuah metrik melebihi ambang batas.

Amazon CloudWatch akan memungkinkan Anda untuk membuat metrik dan alarm komposit dengan menggunakan alarm CloudWatch berdasarkan ambang batas statis, deteksi anomali, dan kriteria lainnya. Untuk detail selengkapnya mengenai jenis alarm yang dapat Anda konfigurasikan dengan menggunakan CloudWatch, silakan lihat bagian alarm pada dokumentasi CloudWatch.

Anda dapat membuat konsep tampilan metrik dan peringatan yang disesuaikan dari sumber daya AWS Anda untuk tim Anda dengan menggunakan dasbor CloudWatch. Halaman beranda yang dapat disesuaikan di konsol CloudWatch dapat memungkinkan Anda memantau sumber daya yang ada di beberapa Wilayah dalam satu tampilan.

Peringatan alarm dapat melakukan satu atau beberapa tindakan, seperti mengirimkan notifikasi ke topik Amazon SNS, pelaksanaan tindakan Amazon EC2 atau tindakan Amazon EC2 Auto Scaling, atau pembuatan OpsItem atau insiden di AWS Systems Manager.

Amazon CloudWatch menggunakan Amazon SNS untuk mengirimkan notifikasi ketika alarm berubah status, menyediakan pengiriman pesan dari penerbit (produsen) ke pelanggan (konsumen). Untuk detail selengkapnya tentang cara mengatur notifikasi Amazon SNS, silakan lihat Mengonfigurasi Amazon SNS.

CloudWatch mengirimkan peristiwa-peristiwa EventBridge ketika sebuah alarm CloudWatch dibuat, diperbarui, dihapus, atau status alarmnya berubah. Anda dapat menggunakan EventBridge dengan peristiwa ini untuk membuat aturan yang melakukan tindakan, seperti memberi tahu Anda setiap kali status alarm berubah atau secara otomatis memicu peristiwa di akun Anda menggunakan otomatisasi Systems Manager.

Terus dapatkan informasi dengan AWS Health. AWS Health adalah sumber informasi otoritatif tentang kondisi sumber daya AWS Cloud Anda. Gunakan AWS Health untuk mendapatkan notifikasi tentang peristiwa layanan yang dikonfirmasi sehingga Anda dapat dengan cepat mengambil langkah-langkah untuk memitigasi dampak apa pun. Buat notifikasi peristiwa AWS Health sesuai keperluan yang dikirim ke saluran email dan obrolan melalui Notifikasi Pengguna AWS serta integrasikan secara programatis dengan alat pemantauan dan peringatan Anda melalui Amazon EventBridge. Jika Anda menggunakan AWS Organizations, agregasikan peristiwa AWS Health di seluruh akun.

Kapan Anda harus menggunakan EventBridge atau Amazon SNS?

Baik EventBridge maupun Amazon SNS dapat digunakan untuk mengembangkan aplikasi berbasis peristiwa, dan Anda bisa memilihnya berdasarkan kebutuhan spesifik Anda.

Anda disarankan untuk menggunakan Amazon EventBridge jika Anda ingin membuat sebuah aplikasi yang bereaksi terhadap peristiwa-peristiwa dari aplikasi, aplikasi SaaS, dan layanan AWS Anda sendiri. EventBridge adalah satu-satunya layanan berbasis peristiwa yang terintegrasi langsung dengan mitra SaaS pihak ketiga. EventBridge juga secara otomatis dapat menyerap peristiwa dari lebih dari 200 layanan AWS tanpa mengharuskan pengembang untuk membuat sumber daya apa pun di akun mereka.

EventBridge menggunakan sebuah struktur berbasis JSON yang ditentukan untuk peristiwa, dan dapat membantu Anda untuk membuat aturan-aturan yang diterapkan di seluruh badan peristiwa untuk memilih peristiwa tempat di mana target. akan diteruskan EventBridge saat ini mendukung lebih dari 20 layanan AWS sebagai target, termasuk AWS Lambda, Amazon SQS, Amazon SNS, Amazon Kinesis Data Streams, dan Amazon Data Firehose.

Amazon SNS direkomendasikan untuk aplikasi-aplikasi yang membutuhkan fan out tinggi (ribuan atau jutaan titik akhir). Pola umum yang kita lihat adalah bahwa pelanggan menggunakan Amazon SNS sebagai target aturan mereka untuk memfilter peristiwa-peristiwa yang mereka butuhkan, dan untuk melakukan fan out ke beberapa titik akhir.

Pesan tidak terstruktur dan dapat dalam format apa pun. Amazon SNS mendukung penerusan pesan ke enam jenis target yang berbeda, termasuk Lambda, Amazon SQS, titik akhir HTTP/S, SMS, push seluler, dan email. Amazon SNS biasanya memiliki latensi di bawah 30 milidetik. Berbagai layanan AWS mengirimkan pesan Amazon SNS dengan mengonfigurasi layanan untuk melakukannya (lebih dari 30, termasuk Amazon EC2, Amazon S3, dan Amazon RDS).

Langkah-langkah implementasi

1. Buat sebuah alarm dengan menggunakan alarm Amazon CloudWatch.
    a. Sebuah alarm metrik memantau metrik CloudWatch tunggal atau ekspresi yang bergantung pada metrik CloudWatch. Alarm memulai satu atau beberapa tindakan berdasarkan nilai metrik atau ekspresi dibandingkan dengan ambang batas selama interval waktu tertentu. Tindakan itu dapat berupa pengiriman sebuah notifikasi ke topik Amazon SNS, melaksanakan tindakan Amazon EC2 atau tindakan Amazon EC2 Auto Scaling, atau membuat OpsItem atau insiden di AWS Systems Manager.
    b. Sebuah alarm gabungan terdiri dari ekspresi aturan yang mempertimbangkan kondisi alarm dari alarm-alarm lain yang telah Anda buat. Alarm gabungan hanya memasuki status alarm jika semua kondisi aturan terpenuhi. Alarm yang ditentukan dalam ekspresi aturan suatu alarm komposit dapat mencakup alarm-alarm metrik dan alarm gabungan tambahan. Alarm gabungan dapat mengirimkan notifikasi Amazon SNS ketika statusnya berubah, dan dapat membuat OpsItems atau insiden Systems Manager ketika statusnya beralih ke status alarm, tetapi alarm tersebut tidak dapat melakukan tindakan-tindakan Amazon EC2 atau Penskalaan Otomatis (Auto Scaling).
2. Mengatur notifikasi Amazon SNS. Saat membuat sebuah alarm CloudWatch, Anda dapat menyertakan sebuah topik Amazon SNS untuk mengirimkan sebuah notifikasi saat status alarm berubah.
3. Buat aturan di EventBridge yang cocok dengan alarm-alarm CloudWatch yang ditentukan. Setiap aturan mendukung beberapa target, termasuk fungsi Lambda. Misalnya, Anda dapat menentukan sebuah alarm yang dimulai saat ruang diska yang tersedia hampir habis, yang memicu sebuah fungsi Lambda melalui sebuah aturan EventBridge, untuk mengosongkan ruang. Untuk detail lebih lanjut tentang target-target EventBridge, lihat target EventBridge.

Sumber daya

Praktik terbaik Well-Architected terkait:

- REL06-BP01 Memantau semua komponen untuk beban kerja (Pembuatan)
- REL06-BP02 Menetapkan dan menghitung metrik (Agregasi)
- REL12-BP01 Menggunakan playbook untuk menyelidiki kegagalan

Dokumen terkait:

- Amazon CloudWatch
- Wawasan Log CloudWatch
- Menggunakan alarm Amazon CloudWatch
- Menggunakan dasbor Amazon CloudWatch
- Menggunakan metrik Amazon CloudWatch
- Menyiapkan notifikasi Amazon SNS
- Deteksi anomali CloudWatch
- Perlindungan data CloudWatch Logs
- Amazon EventBridge
- Amazon Simple Notification Service

Video terkait:

- Video observabilitas reinvent 2022
- AWS re:Invent 2022 - Praktik terbaik observabilitas di Amazon

Contoh terkait:

- One Observability Workshop
- Amazon EventBridge ke AWS Lambda dengan kontrol umpan balik oleh Alarm Amazon CloudWatch

REL06-BP04 Mengotomatiskan respons (Peringatan dan pemrosesan waktu nyata)

Gunakan otomatisasi untuk melakukan tindakan ketika peristiwa terdeteksi, misalnya, mengganti komponen yang rusak.

Pemrosesan alarm waktu nyata secara otomatis diimplementasikan agar sistem dapat mengambil tindakan-tindakan korektif dengan cepat dan berupaya mencegah terjadinya kegagalan atau penurunan layanan ketika alarm terpicu. Respons otomatis terhadap alarm dapat mencakup penggantian komponen yang mengalami kegagalan, penyesuaian kapasitas komputasi, pengalihan lalu lintas ke host yang dalam kondisi sehat, zona ketersediaan, atau wilayah lain, dan pemberitahuan operator.

Hasil yang diinginkan: Alarm waktu nyata diidentifikasi, dan pemrosesan alarm otomatis diatur untuk menginvokasi tindakan-tindakan yang tepat yang diambil untuk mempertahankan tujuan tingkat layanan dan perjanjian tingkat layanan (SLA). Otomatisasi dapat berupa banyak hal, dari aktivitas pemulihan diri sebuah komponen hingga failover seluruh situs.

Anti-pola umum:

- Tidak memiliki inventaris atau katalog alarm waktu nyata utama yang jelas.
- Tidak ada respons otomatis terhadap alarm-alarm kritis (misalnya, penskalaan otomatis berjalan ketika komputasi hampir habis).
- Tindakan respons alarm yang kontradiktif.
- Tidak ada prosedur operasi standar (SOP) yang harus diikuti operator ketika mereka menerima pemberitahuan peringatan.
- Tidak memantau perubahan konfigurasi, padahal perubahan konfigurasi yang tidak terdeteksi dapat menyebabkan waktu henti terhadap beban kerja.
- Tidak memiliki strategi untuk membatalkan perubahan konfigurasi yang tidak diinginkan.

Manfaat menerapkan praktik terbaik ini: Melakukan otomatisasi atas pemrosesan alarm dapat meningkatkan ketahanan sistem. Sistem mengambil tindakan-tindakan korektif secara otomatis, sehingga akan mengurangi aktivitas manual yang memberi peluang adanya intervensi manusia yang rawan menyebabkan kesalahan. Operasi beban kerja memenuhi tujuan-tujuan ketersediaan, dan mengurangi gangguan layanan.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Sedang

Panduan implementasi

Untuk mengelola peringatan secara efektif dan mengotomatiskan responsnya, Anda harus mengkategorikan peringatan berdasarkan tingkat kekritisan dan dampaknya, mendokumentasikan prosedur respons, dan merencanakan respons sebelum menentukan peringkat tugas.

Identifikasi tugas yang membutuhkan tindakan-tindakan tertentu (sering kali diperinci dalam runbook), dan periksa semua runbook dan playbook untuk menentukan tugas mana yang dapat diotomatisasi. Jika tindakan dapat ditentukan, tindakan tersebut sering kali dapat diotomatisasi. Jika tindakan tidak dapat diotomatisasi, maka Anda harus mendokumentasikan langkah-langkah pelaksanaannya dalam SOP dan latih operator untuk melakukannya. Terus cari peluang otomatisasi pada proses-proses yang masih dilakukan secara manual agar Anda dapat membuat dan menerapkan rencana untuk mengotomatiskan respons peringatan.

Langkah-langkah implementasi

1. Buat inventaris alarm: Untuk mendapatkan daftar semua alarm, Anda dapat menggunakan AWS CLI dengan menggunakan perintah `describe-alarms` Amazon CloudWatch. Bergantung pada berapa banyak alarm yang telah Anda atur, Anda mungkin harus menggunakan pagination untuk mengambil subset alarm untuk setiap panggilan, atau sebagai alternatif, Anda dapat menggunakan SDK AWS untuk mendapatkan alarm dengan menggunakan panggilan API.
2. Dokumentasikan semua tindakan alarm: Perbarui runbook dengan semua alarm dan tindakannya, terlepas dari apakah runbook itu manual atau otomatis. AWS Systems Manager menyediakan runbook yang telah ditentukan sebelumnya. Untuk informasi lebih lanjut tentang runbook, lihat Bekerja dengan runbook. Untuk detail tentang cara melihat konten runbook, silakan lihat Menampilkan konten runbook.
3. Mengatur dan mengelola tindakan alarm: Untuk alarm apa pun yang memerlukan sebuah tindakan, tentukan tindakan otomatis dengan menggunakan CloudWatch SDK. Misalnya, Anda dapat mengubah status instans Amazon EC2 Anda secara otomatis berdasarkan sebuah alarm CloudWatch dengan membuat dan mengaktifkan tindakan pada alarm atau menonaktifkan tindakan pada alarm.

   Anda dapat juga menggunakan Amazon EventBridge untuk memberikan respons secara otomatis terhadap peristiwa sistem, seperti terjadinya masalah ketersediaan aplikasi atau perubahan sumber daya. Anda dapat membuat aturan untuk menunjukkan peristiwa yang sesuai kepentingan Anda, dan tindakan yang akan diambil ketika peristiwa sesuai dengan aturan. Tindakan-tindakan yang dapat dimulai secara otomatis termasuk menginvokasi fungsi AWS Lambda, menginvokasi

`Run Command` Amazon EC2, menyampaikan peristiwa ke Amazon Kinesis Data Streams, dan melihat Mengotomatiskan Amazon EC2 menggunakan EventBridge.

4. Prosedur Operasi Standar (SOP): Berdasarkan komponen-komponen aplikasi Anda, AWS Resilience Hub merekomendasikan beberapa templat SOP. Anda dapat menggunakan SOP ini untuk mendokumentasikan semua proses yang harus diikuti operator saat ada peringatan yang muncul. Anda juga dapat membuat konsep SOP berdasarkan rekomendasi dari Resilience Hub, di mana Anda memerlukan aplikasi Resilience Hub yang memiliki kebijakan ketahanan terkait, serta penilaian ketahanan historis terhadap aplikasi tersebut. Rekomendasi untuk SOP Anda berasal dari penilaian ketahanan.

   Resilience Hub bekerja dengan Systems Manager untuk melakukan otomatisasi pada langkah-langkah yang diuraikan dalam SOP Anda dengan menyediakan sejumlah dokumen SSM yang dapat Anda gunakan sebagai landasan untuk SOP tersebut. Misalnya, Resilience Hub mungkin akan merekomendasikan SOP untuk menambahkan ruang diska berdasarkan dokumen otomatisasi SSM yang sudah ada.

5. Lakukan tindakan otomatis menggunakan Amazon DevOps Guru: Anda dapat menggunakan Amazon DevOps Guru untuk melakukan pemantauan secara otomatis terhadap sumber daya aplikasi untuk mengetahui perilaku anomali dan memberikan rekomendasi terarah untuk mempercepat waktu perbaikan serta identifikasi masalah. Dengan DevOps Guru, Anda dapat memantau aliran data operasional dalam waktu dekat dari berbagai sumber, antara lain metrik Amazon CloudWatch, AWS Config, AWS CloudFormation, dan AWS X-Ray. Anda juga dapat menggunakan DevOps Guru untuk secara otomatis membuat OpsItems di OpsCenter dan mengirim peristiwa-peristiwa ke EventBridge untuk otomatisasi tambahan.

Sumber daya

Praktik-praktik terbaik terkait:

- REL06-BP01 Memantau semua komponen untuk beban kerja (Pembuatan)
- REL06-BP02 Menetapkan dan menghitung metrik (Agregasi)
- REL06-BP03 Mengirimkan notifikasi (Pemrosesan dan pembuatan alarm waktu nyata)
- REL08-BP01 Menggunakan runbook untuk aktivitas standar seperti deployment

Dokumen terkait:

- Otomatisasi AWS Systems Manager

- Membuat Aturan EventBridge yang Memicu Peristiwa dari Sumber Daya AWS
- Lokakarya Satu Observabilitas
- Amazon Builders' Library: Instrumentasi sistem terdistribusi untuk visibilitas operasional
- Apa itu Amazon DevOps Guru?
- Bekerja dengan Dokumen Otomatisasi (Playbooks)

Video terkait:

- AWS re:Invent 2022 - Praktik terbaik observabilitas di Amazon
- AWS re:Invent 2020: Otomatiskan apa pun dengan AWS Systems Manager
- Pengantar tentang AWS Resilience Hub
- Buat Sistem Tiket Kustom untuk Notifikasi Amazon DevOps Guru
- Aktifkan Agregasi Wawasan Multi-Akun dengan Amazon DevOps Guru

Contoh terkait:

- Lokakarya Amazon CloudWatch and Systems Manager

REL06-BP05 Menganalisis log

Kumpulkan riwayat metrik dan file log dan analisis ini untuk mendapatkan wawasan beban kerja dan tren lebih luas.

Wawasan Log Amazon CloudWatch mendukung bahasa kueri sederhana namun berdayaguna yang dapat Anda gunakan untuk melakukan analisis data log. Log Amazon CloudWatch juga mendukung berlangganan yang memungkinkan data mengalir dengan mulus ke Amazon S3 tempat Anda dapat menggunakan data atau Amazon Athena untuk melakukan kueri data. Amazon CloudWatch Logs juga mendukung pembuatan kueri dengan berbagai macam format. Lihat SerDes dan Format Data yang Didukung di Panduan Pengguna Amazon Athena untuk mendapatkan informasi selengkapnya. Untuk melakukan analisis set file log yang sangat besar, Anda dapat menjalankan klaster Amazon EMR untuk melakukan analisis skala petabita.

Ada sejumlah alat yang disediakan oleh Partner AWS dan pihak ketiga yang dapat memungkinkan agregasi, pemrosesan, penyimpanan, dan analitik. Alat-alat ini antara lain New Relic, Splunk, Loggly, Logstash, CloudHealth, dan Nagios. Namun demikian, log generasi di luar sistem dan log aplikasi bersifat unik untuk setiap penyedia cloud, dan sering kali bersifat unik untuk masing-masing layanan.

Bagian proses pemantauan yang sering kali tidak diperhatikan adalah manajemen data. Anda harus menentukan persyaratan-persyaratan retensi untuk memantau data, kemudian Anda harus menerapkan kebijakan siklus hidup yang sesuai. Amazon S3 mendukung manajemen siklus hidup pada tingkat bucket S3. Manajemen siklus hidup ini dapat diterapkan secara berbeda ke jalur-jalur yang berbeda yang ada dalam bucket tersebut. Menjelang akhir siklus hidup, Anda dapat melakukan transisi data ke Amazon Glacier untuk mendapatkan penyimpanan jangka panjang, dan kemudian akan menjadi kedaluwarsa setelah akhir jangka waktu retensi tercapai. Kelas penyimpanan S3 Intelligent-Tiering didesain untuk mengoptimalkan biaya dengan secara otomatis memindahkan data ke tingkat akses yang paling hemat biaya, tanpa memengaruhi performa atau menimbulkan biaya operasional tambahan.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Sedang

Panduan implementasi

- Wawasan Log CloudWatch akan memungkinkan Anda secara interaktif mencari dan menganalisis data log Anda di Log Amazon CloudWatch.
  - Melakukan Analisis Data Log dengan Wawasan Log CloudWatch
  - Kueri Contoh Wawasan Log Amazon CloudWatch
- Gunakan Log Amazon CloudWatch untuk mengirim log ke Amazon S3 tempat Anda dapat menggunakan data atau Amazon Athena tempat Anda dapat melakukan kueri data.
  - Bagaimana cara menganalisis log akses server Amazon S3 menggunakan Athena?
    - Buat kebijakan siklus hidup S3 untuk bucket log akses server Anda. Konfigurasikan kebijakan siklus hidup untuk secara berkala menghapus file log. Hal ini mengurangi jumlah data yang dianalisis oleh Athena untuk setiap kueri.
      - Bagaimana Cara Membuat Kebijakan Siklus Hidup untuk Bucket S3?

Sumber daya

Dokumen terkait:

- Kueri Contoh Wawasan Log Amazon CloudWatch
- Melakukan Analisis Data Log dengan Wawasan Log CloudWatch
- Melakukan Debugging dengan Amazon CloudWatch Synthetics dan AWS X-Ray
- Bagaimana Cara Membuat Kebijakan Siklus Hidup untuk Bucket S3?
- Bagaimana cara menganalisis log akses server Amazon S3 menggunakan Athena?

- Lokakarya Satu Observabilitas
- Amazon Builders' Library: Instrumentasi sistem terdistribusi untuk visibilitas operasional

REL06-BP06 Meninjau cakupan dan metrik pemantauan secara berkala

Tinjau secara berkala bagaimana pemantauan beban kerja diterapkan, dan sesuaikan peninjauan Anda seiring perkembangan beban kerja Anda dan arsitekturnya. Pengauditan rutin atas pemantauan Anda membantu mengurangi risiko indikator masalah yang terlewatkan atau diabaikan dan lebih membantu beban kerja Anda memenuhi sasaran ketersediaannya.

Pemantauan yang efektif bergantung pada metrik bisnis utama, yang berkembang seiring dengan perubahan prioritas bisnis Anda. Proses peninjauan pemantauan Anda harus menekankan indikator tingkat layanan (SLI) dan menggabungkan wawasan dari infrastruktur, aplikasi, klien, dan pengguna Anda.

Hasil yang diinginkan: Anda memiliki strategi pemantauan yang efektif yang ditinjau dan diperbarui secara berkala, serta setelah setiap peristiwa atau perubahan signifikan. Anda memverifikasi bahwa indikator kesehatan aplikasi utama masih relevan seiring dengan perkembangan beban kerja dan kebutuhan bisnis Anda.

Anti-pola umum:

- Anda hanya mengumpulkan metrik default.
- Anda membuat strategi pemantauan, tetapi Anda tidak pernah meninjaunya.
- Anda tidak membahas pemantauan ketika ada deployment perubahan besar.
- Anda memercayai metrik yang sudah usang untuk menentukan kesehatan beban kerja.
- Tim operasi Anda kewalahan dengan peringatan positif palsu karena metrik dan ambang batas yang sudah usang.
- Anda kekurangan observabilitas atas komponen aplikasi yang tidak dipantau.
- Anda hanya fokus pada metrik teknis tingkat rendah dan tidak memasukkan metrik bisnis dalam pemantauan Anda.

Manfaat menjalankan praktik terbaik ini: Ketika Anda secara teratur meninjau pemantauan Anda, Anda dapat mengantisipasi potensi masalah dan memverifikasi bahwa Anda mampu mendeteksinya. Hal ini juga memungkinkan Anda menemukan titik buta yang mungkin Anda lewati dalam peninjauan sebelumnya, sehingga lebih meningkatkan kemampuan Anda untuk mendeteksi masalah.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Sedang

Panduan implementasi

Tinjau metrik dan cakupan pemantauan selama proses peninjauan kesiapan operasional (ORR) Anda. Lakukan peninjauan kesiapan operasional berkala dengan jadwal yang konsisten untuk mengevaluasi apakah ada kesenjangan antara beban kerja Anda saat ini dan pemantauan yang telah Anda konfigurasikan. Tetapkan ritme teratur untuk peninjauan performa operasional dan berbagi pengetahuan untuk meningkatkan kemampuan Anda dalam mencapai kinerja tim operasional yang lebih tinggi. Validasi apakah ambang batas peringatan yang ada masih memadai, dan periksa dalam situasi apa tim operasional menerima peringatan positif palsu atau tidak memantau aspek aplikasi yang harus dipantau.

Kerangka Kerja Analisis Ketahanan memberikan panduan berguna yang dapat membantu Anda menavigasi proses ini. Fokus kerangka kerja ini adalah untuk mengidentifikasi mode kegagalan potensial serta kontrol preventif dan korektif yang dapat Anda gunakan untuk memitigasi dampaknya. Pengetahuan ini dapat membantu Anda mengidentifikasi metrik dan peristiwa yang tepat untuk dipantau dan diwaspadai.

Langkah-langkah implementasi

1. Jadwalkan dan lakukan peninjauan dasbor beban kerja secara teratur. Anda mungkin memiliki berbagai ritme untuk kedalaman inspeksi Anda.
2. Inspeksi apakah ada kecenderungan dalam metrik-metrik tersebut. Bandingkan nilai-nilai metrik dengan nilai-nilai historis untuk melihat apakah ada tren yang mungkin mengindikasikan bahwa sesuatu perlu diselidiki. Contohnya antara lain: peningkatan latensi, penurunan fungsi bisnis utama, dan peningkatan respons-respons kegagalan.
3. Periksa pencilan dan anomali dalam metrik Anda, yang dapat tertutupi oleh rata-rata atau median. Perhatikan nilai tertinggi dan terendah selama jangka waktu tersebut, dan selidiki penyebab observasi yang jauh di luar batas normal. Seiring Anda terus menghilangkan penyebab ini, Anda dapat memperketat batas metrik yang diharapkan sebagai respons terhadap peningkatan konsistensi kinerja beban kerja Anda.
4. Cari perubahan-perubahan mendadak dalam perilaku. Perubahan cepat yang terjadi dalam jumlah atau arah metrik dapat menandakan telah ada perubahan dalam aplikasi, atau ada faktor eksternal yang mungkin mengharuskan Anda menambahkan metrik-metrik lainnya untuk dilacak.
5. Tinjau apakah strategi pemantauan saat ini tetap relevan untuk aplikasi. Berdasarkan analisis insiden sebelumnya (atau Kerangka Kerja Analisis Ketahanan), nilai apakah ada aspek tambahan dari aplikasi yang harus dimasukkan ke dalam cakupan pemantauan.

6. Tinjau metrik Pemantauan Pengguna Nyata (RUM) Anda untuk menentukan apakah ada kesenjangan dalam cakupan fungsionalitas aplikasi.
7. Tinjau proses manajemen perubahan Anda. Perbarui prosedur Anda jika perlu untuk menyertakan langkah analisis pemantauan yang harus dilakukan sebelum Anda menyetujui perubahan.
8. Terapkan tinjauan pemantauan sebagai bagian dari tinjauan kesiapan operasional Anda dan koreksi proses kesalahan.

Sumber daya

Praktik-praktik terbaik terkait

- REL06-BP01 Memantau semua komponen untuk beban kerja (Pembuatan)
- REL06-BP02 Menetapkan dan menghitung metrik (Agregasi)
- REL06-BP07 Memantau pelacakan permintaan menyeluruh melalui sistem Anda
- REL12-BP02 Menjalankan analisis setelah insiden
- REL12-BP06 Mengadakan game day secara rutin

Dokumen terkait:

- Mengapa Anda harus mengembangkan koreksi kesalahan (COE)
- Menggunakan Dasbor Amazon CloudWatch
- Membangun dasbor untuk visibilitas operasional
- Pola Ketahanan Multi-AZ Tingkat Lanjut - Kegagalan abu-abu
- Kueri Contoh Wawasan Log Amazon CloudWatch
- Melakukan Debugging dengan Amazon CloudWatch Synthetics dan AWS X-Ray
- Lokakarya Satu Observabilitas
- Amazon Builders' Library: Instrumentasi sistem terdistribusi untuk visibilitas operasional
- Menggunakan Dasbor Amazon CloudWatch
- AWS Praktik Terbaik Observabilitas
- Kerangka Kerja Analisis Ketahanan
- Kerangka Kerja Analisis Ketahanan - Observabilitas
- Peninjauan Kesiapan Operasional - ORR

REL06-BP07 Memantau pelacakan permintaan menyeluruh melalui sistem Anda

Lacak permintaan yang sedang diproses melalui komponen layanan agar tim produk dapat lebih mudah menganalisis dan menemukan serta memperbaiki masalah dan meningkatkan kinerja.

Hasil yang diinginkan: Beban kerja dengan penelusuran komprehensif di semua komponen biasanya mudah di-debug, meningkatkan mean time to resolution (MTTR) kesalahan dan latensi dengan menyederhanakan penemuan akar masalah yang menjadi penyebabnya. Penelusuran yang menyeluruh akan mempersingkat waktu yang diperlukan untuk menemukan komponen-komponen yang terdampak dan mencari tahu akar masalah yang menyebabkan kesalahan atau latensi secara mendetail.

Anti-pola umum:

- Penelusuran digunakan untuk beberapa komponen, tidak semuanya. Misalnya, tanpa penelusuran untuk AWS Lambda, tim mungkin tidak akan memahami dengan jelas latensi yang disebabkan oleh cold start dalam beban kerja yang mengalami fluktuasi.
- Canary sintetis atau pemantauan pengguna nyata (RUM) tidak dikonfigurasi dengan penelusuran. Tanpa canary atau RUM, telemetri interaksi klien dihilangkan dari analisis jejak, hal ini berimbas pada tidak lengkapnya profil kinerja.
- Beban kerja hibrida mencakup alat-alat penelusuran cloud-native dan pihak ketiga, tetapi langkah-langkah belum dilakukan untuk memilih dan sepenuhnya mengintegrasikan solusi penelusuran tunggal. Berdasarkan solusi penelusuran yang dipilih, SDK penelusuran cloud-native harus digunakan untuk melakukan instrumentasi terhadap komponen-komponen yang bukan cloud-native, atau alat-alat pihak ketiga harus dikonfigurasi untuk menyerap telemetri pelacakan cloud-native.

Manfaat menerapkan praktik terbaik ini: Saat tim pengembangan menerima peringatan masalah, mereka dapat melihat gambaran utuh dari interaksi komponen sistem, termasuk korelasi komponen per komponen dengan pembuatan log, kinerja, dan kegagalan. Karena penelusuran memudahkan Anda untuk mengidentifikasi akar masalah secara visual, waktu penyelidikan akar masalah akan menjadi lebih singkat. Tim yang memahami interaksi komponen secara detail dapat mengambil keputusan yang lebih baik dan lebih cepat saat menyelesaikan masalah. Keputusan seperti kapan harus menginvokasi failover pemulihan bencana (DR) atau lokasi terbaik untuk menerapkan strategi penyembuhan mandiri dapat ditingkatkan dengan menganalisis jejak sistem, dan pada akhirnya meningkatkan kepuasan pelanggan terhadap layanan Anda.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Sedang

Panduan implementasi

Tim yang mengoperasikan aplikasi-aplikasi yang terdistribusi dapat menggunakan alat penelusuran untuk membuat sebuah pengidentifikasi korelasi, mengumpulkan jejak permintaan, dan membuat peta layanan dari komponen-komponen yang terhubung. Semua komponen aplikasi harus disertakan dalam jejak-jejak permintaan, termasuk klien layanan, gateway perangkat lunak perantara (middleware) dan bus peristiwa, komponen komputasi, dan penyimpanan, termasuk penyimpanan nilai kunci dan basis data. Sertakan canary sintetis dan pemantauan pengguna nyata dalam konfigurasi penelusuran menyeluruh (end-to-end) Anda untuk mengukur interaksi dan latensi klien jarak jauh sehingga Anda dapat secara akurat mengevaluasi kinerja sistem Anda berdasarkan perjanjian dan tujuan-tujuan tingkat layanan Anda.

Anda dapat menggunakan AWS X-Ray dan layanan-layanan instrumentasi Pemantauan Aplikasi Amazon CloudWatch untuk menyediakan tampilan permintaan yang lengkap saat mereka melakukan perjalanan melalui aplikasi Anda. X-Ray mengumpulkan telemetri aplikasi dan memungkinkan Anda untuk memvisualisasikan dan memfilter telemetri tersebut di seluruh muatan, fungsi, jejak, layanan, API, dan dapat diaktifkan untuk komponen sistem tanpa perlu melakukan coding atau coding minimum. Pemantauan aplikasi CloudWatch mencakup ServiceLens yang bisa digunakan untuk mengintegrasikan jejak Anda dengan metrik, log, dan alarm. Pemantauan aplikasi CloudWatch juga mencakup pemantauan sintetis untuk memantau titik akhir dan API Anda, serta pemantauan pengguna nyata untuk melengkapi instrumen klien aplikasi web Anda.

Langkah-langkah implementasi

- Gunakan AWS X-Ray di semua layanan native yang didukung seperti Amazon S3AWS Lambda, dan Amazon API Gateway. Semua layanan AWS ini mengaktifkan X-Ray dengan pengalih konfigurasi menggunakan infrastruktur sebagai kode, SDK AWS, atau Konsol Manajemen AWS.
- Aplikasi instrumen AWS Distro for Open Telemetry dan X-Ray atau agen pengumpulan pihak ketiga.
- Tinjau Panduan Pengembang AWS X-Ray untuk penerapan khusus bahasa pemrograman. Bagian dokumentasi ini menjelaskan cara melakukan instrumentasi terhadap permintaan HTTP, mengkueri SQL, dan proses lain yang spesifik untuk bahasa pemrograman aplikasi Anda.
- Gunakan penelusuran X-Ray untuk Canary Amazon CloudWatch Synthetics dan Amazon CloudWatch RUM untuk melakukan analisis terhadap jalur permintaan dari klien pengguna akhir Anda melalui infrastruktur AWS hilir Anda.

- Konfigurasikan metrik dan alarm CloudWatch berdasarkan telemetri kesehatan sumber daya dan canary sehingga tim dapat menerima peringatan masalah dengan cepat, kemudian dapat mempelajari jejak dan peta layanan dengan ServiceLens.
- Aktifkan integrasi X-Ray untuk alat-alat penelusuran pihak ketiga seperti Datadog, New Relic, atau Dynatrace jika Anda menggunakan alat-alat pihak ketiga untuk solusi penelusuran utama Anda.

Sumber daya

Praktik-praktik terbaik terkait:

- REL06-BP01 Memantau semua komponen untuk beban kerja (Pembuatan)
- REL11-BP01 Memantau semua komponen beban kerja untuk mendeteksi kegagalan

Dokumen terkait:

- Apa itu AWS X-Ray?
- Amazon CloudWatch: Pemantauan Aplikasi
- Melakukan Debugging dengan Amazon CloudWatch Synthetics dan AWS X-Ray
- Amazon Builders' Library: Instrumentasi sistem terdistribusi untuk visibilitas operasional
- Mengintegrasikan AWS X-Ray dengan layanan AWS lainnya
- AWS Distro for OpenTelemetry dan AWS X-Ray
- Amazon CloudWatch: Menggunakan pemantauan sintetis
- Amazon CloudWatch: Gunakan CloudWatch RUM
- Siapkan alarm canary Amazon CloudWatch synthetics dan Amazon CloudWatch
- Ketersediaan dan Lainnya: Memahami dan Meningkatkan Ketangguhan Sistem Terdistribusi di AWS

Contoh terkait:

- Lokakarya Satu Observabilitas

Video terkait:

- AWS re:Invent 2022 - Cara memantau aplikasi di beberapa akun

- Cara Memantau Aplikasi AWS Anda

Alat terkait:

- AWS X-Ray
- Amazon CloudWatch
- Amazon Route 53

Rancang beban kerja Anda agar dapat beradaptasi dengan perubahan dalam permintaan

Beban kerja yang dapat diskalakan memberikan elastisitas untuk menambahkan atau mengeluarkan sumber daya secara otomatis sehingga sangat sesuai dengan permintaan saat ini pada titik waktu tertentu.

Praktik terbaik

- REL07-BP01 Menggunakan otomatisasi ketika mendapatkan atau menskalakan sumber daya
- REL07-BP02 Mendapatkan sumber daya setelah deteksi gangguan pada beban kerja
- REL07-BP03 Menambah sumber daya berdasarkan deteksi bahwa beban kerja memerlukan lebih banyak sumber daya
- REL07-BP04 Beban uji beban kerja Anda

REL07-BP01 Menggunakan otomatisasi ketika mendapatkan atau menskalakan sumber daya

Landasan keandalan di cloud adalah definisi, penyediaan, dan pengelolaan infrastruktur dan sumber daya Anda secara terprogram. Otomatisasi membantu Anda menyederhanakan penyediaan sumber daya, memfasilitasi deployment yang konsisten dan aman, serta menskalakan sumber daya di seluruh infrastruktur Anda.

Hasil yang diinginkan: Anda mengelola infrastruktur sebagai kode (IaC). Anda menentukan dan memelihara kode infrastruktur Anda dalam sistem kontrol versi (VCS). Anda mendelegasikan penyediaan sumber daya AWS ke mekanisme otomatis dan memanfaatkan layanan yang dikelola seperti Penyeimbang Beban Aplikasi (ALB), Penyeimbang Beban Jaringan (NLB), dan grup Auto

Scaling. Anda menyediakan sumber daya Anda menggunakan pipeline integrasi berkelanjutan/pengiriman berkelanjutan (CI/CD) sehingga perubahan kode secara otomatis memulai pembaruan sumber daya, termasuk pembaruan pada konfigurasi Auto Scaling Anda.

Anti-pola umum:

- Anda melakukan deployment sumber daya secara manual menggunakan baris perintah atau di Konsol Manajemen AWS (juga dikenal sebagai click-ops).
- Anda melakukan tight coupling terhadap komponen atau sumber daya aplikasi Anda, dan akibatnya menciptakan arsitektur yang tidak fleksibel.
- Anda menerapkan kebijakan penskalaan yang tidak fleksibel dan tidak beradaptasi dengan perubahan persyaratan bisnis, pola lalu lintas, atau jenis sumber daya baru.
- Anda secara manual memperkirakan kapasitas untuk memenuhi permintaan yang diantisipasi.

Manfaat menjalankan praktik terbaik ini: Infrastruktur sebagai kode (IaC) memungkinkan infrastruktur didefinisikan secara terprogram. Hal ini membantu Anda mengelola perubahan infrastruktur melalui siklus hidup pengembangan perangkat lunak yang sama dengan perubahan aplikasi, sehingga meningkatkan konsistensi dan pengulangan, serta mengurangi risiko tugas manual yang rawan kesalahan. Anda dapat menyederhanakan lebih lanjut proses penyediaan dan pembaruan sumber daya melalui penerapan IaC dengan pipeline pengiriman otomatis. Anda dapat melakukan deployment pembaruan infrastruktur secara andal dan efisien tanpa perlu intervensi manual. Ketangkasan ini sangat penting ketika menskalakan sumber daya untuk memenuhi permintaan yang berfluktuasi.

Anda dapat mencapai penskalaan sumber daya otomatis yang dinamis bersama dengan IaC dan pipeline pengiriman. Dengan memantau metrik utama dan menerapkan kebijakan penskalaan yang telah ditentukan sebelumnya, Auto Scaling dapat secara otomatis menyediakan atau menghentikan penyediaan sumber daya sesuai kebutuhan, sehingga meningkatkan kinerja dan efisiensi biaya. Hal ini mengurangi potensi kesalahan manual atau keterlambatan dalam menanggapi perubahan pada persyaratan aplikasi atau beban kerja.

Kombinasi IaC, pipeline pengiriman otomatis, dan Auto Scaling membantu organisasi menyediakan, memperbarui, dan menskalakan lingkungan mereka dengan percaya diri. Otomatisasi ini sangat penting untuk menjaga infrastruktur cloud yang responsif, tangguh, dan dikelola secara efisien.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Tinggi

Panduan implementasi

Untuk mengatur otomatisasi dengan pipeline CI/CD dan infrastruktur sebagai kode (IaC) untuk arsitektur AWS Anda, pilih sistem kontrol versi seperti Git untuk menyimpan templat dan konfigurasi IaC Anda. Templat ini dapat ditulis menggunakan alat-alat seperti AWS CloudFormation. Untuk memulai, tentukan komponen infrastruktur Anda (seperti VPC AWS, Grup Amazon EC2 Auto Scaling, dan basis data Amazon RDS) dalam templat-templat ini.

Selanjutnya, integrasikan templat IaC ini dengan pipeline CI/CD untuk mengotomatiskan proses deployment. AWS CodePipelinemenyediakan solusi bawaan AWS yang lancar, atau Anda dapat menggunakan solusi CI/CD pihak ketiga lainnya. Buat pipeline yang aktif saat terjadi perubahan pada repositori kontrol versi Anda. Konfigurasikan pipeline untuk menyertakan tahapan yang melakukan linting dan validasi terhadap templat IaC Anda, melakukan deployment infrastruktur ke lingkungan staging, menjalankan pengujian otomatis, dan terakhir, melakukan deployment ke produksi. Sertakan langkah-langkah persetujuan jika diperlukan untuk mempertahankan kontrol atas perubahan. Pipeline otomatis ini tidak hanya mempercepat deployment, tetapi juga memfasilitasi konsistensi dan keandalan di seluruh lingkungan.

Konfigurasikan Auto Scaling sumber daya seperti instans Amazon EC2, tugas Amazon ECS, dan replika basis data di IaC Anda untuk menyediakan penambahan skala dan pengurangan skala secara otomatis sesuai kebutuhan. Pendekatan ini meningkatkan ketersediaan dan kinerja aplikasi serta mengoptimalkan biaya dengan menyesuaikan sumber daya secara dinamis berdasarkan permintaan. Untuk mengetahui daftar sumber daya yang didukung, lihat Amazon EC2 Auto Scaling dan AWS Auto Scaling.

Langkah-langkah implementasi

1. Buat dan gunakan repositori kode sumber untuk menyimpan kode yang mengontrol konfigurasi infrastruktur Anda. Lakukan commit perubahan pada repositori ini untuk mencerminkan perubahan yang sedang berlangsung yang ingin Anda buat.
2. Pilih solusi infrastruktur sebagai kode seperti AWS CloudFormation untuk menjaga infrastruktur Anda tetap mutakhir dan mendeteksi inkonsistensi (pergeseran) dari kondisi yang Anda inginkan.
3. Integrasikan platform IaC Anda dengan pipeline CI/CD Anda untuk mengotomatiskan deployment.
4. Tentukan dan kumpulkan metrik yang sesuai untuk penskalaan sumber daya secara otomatis.
5. Konfigurasikan penskalaan otomatis sumber daya menggunakan kebijakan penambahan skala dan pengurangan skala yang sesuai untuk komponen beban kerja Anda. Pertimbangkan untuk menggunakan penskalaan terjadwal untuk pola penggunaan yang dapat diprediksi.

6. Pantau deployment untuk mendeteksi kegagalan dan regresi. Terapkan mekanisme rollback dalam platform CI/CD Anda untuk mengembalikan perubahan, jika perlu.

Sumber daya

Dokumen terkait:

- AWS Auto Scaling: Cara Kerja Penskalaan
- AWS Marketplace: produk yang dapat digunakan dengan penskalaan otomatis
- Mengelola Kapasitas Throughput Secara Otomatis dengan Menggunakan Penskalaan Otomatis DynamoDB
- Menggunakan penyeimbang beban dengan grup Auto Scaling
- Apa itu AWS Akselerator Global?
- Apa itu Amazon EC2 Auto Scaling?
- Apa itu AWS Auto Scaling?
- Apa yang dimaksud dengan Amazon CloudFront?
- Apa itu Amazon Route 53?
- Apa itu Penyeimbangan Beban Elastis?
- Apa itu Penyeimbang Beban Jaringan?
- Apa itu Penyeimbang Beban Aplikasi?
- Mengintegrasikan Jenkins dengan AWS CodeBuild dan AWS CodeDeploy
- Membuat pipeline empat tahap dengan AWS CodePipeline

Video terkait:

- Back to Basics: Melakukan Deployment Kode Anda ke Amazon EC2
- AWS Supports You | Memulai Solusi Infrastruktur sebagai Kode Anda Menggunakan Templat
- Merampingkan Proses Rilis Perangkat Lunak Menggunakan AWS CodePipeline
- Memantau Sumber Daya AWS Menggunakan Dasbor Amazon CloudWatch
- Membuat Dasbor CloudWatch Lintas Akun & Lintas Wilayah | Amazon Web Services

REL07-BP02 Mendapatkan sumber daya setelah deteksi gangguan pada beban kerja

Skalakan sumber daya secara reaktif saat diperlukan jika ketersediaan terganggu, guna memulihkan ketersediaan beban kerja.

Anda terlebih dahulu harus mengonfigurasi pemeriksaan kondisi dan kriteria pada pemeriksaan ini agar memberikan penanda saat ada ketersediaan yang terganggu karena kurangnya sumber daya. Lalu, beri tahu personel yang bersangkutan untuk menskalakan sumber daya secara manual, atau mulai lakukan otomatisasi untuk menskalakannya secara otomatis.

Skala dapat disesuaikan secara manual untuk beban kerja Anda (misalnya, penggantian jumlah instans EC2 di grup Auto Scaling atau modifikasi throughput tabel DynamoDB yang dapat dilakukan melalui Konsol Manajemen AWS atau AWS CLI). Namun, otomatisasi harus digunakan bila memungkinkan (silakan lihat Gunakan otomatisasi saat memperoleh atau menskalakan sumber daya).

Hasil yang diinginkan: Aktivitas penskalaan (baik secara otomatis atau manual) dimulai untuk memulihkan ketersediaan setelah mendeteksi adanya kegagalan atau pengalaman pelanggan yang mengalami penurunan kualitas.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Sedang

Panduan implementasi

Terapkan observabilitas dan pemantauan di semua komponen yang ada dalam beban kerja Anda untuk memantau pengalaman pelanggan dan mendeteksi terjadinya kegagalan. Tentukan prosedur, baik manual ataupun otomatis, yang menskalakan sumber daya yang diperlukan. o Untuk informasi selengkapnya, silakan lihat REL11-BP01 Memantau semua komponen beban kerja untuk mendeteksi kegagalan.

Langkah-langkah implementasi

- Tentukan prosedur, baik manual ataupun otomatis, yang menskalakan sumber daya yang dibutuhkan.
  - Prosedur penskalaan tergantung pada bagaimana rancangan berbagai komponen yang ada dalam beban kerja Anda.
  - Prosedur penskalaan juga bisa berbeda-beda, tergantung pada teknologi dasar yang digunakan.

- Komponen-komponen yang menggunakan AWS Auto Scaling dapat menggunakan rencana penskalaan untuk mengonfigurasi serangkaian instruksi guna menskalakan sumber daya Anda. Jika Anda menggunakan AWS CloudFormation atau menambahkan tag ke sumber daya AWS, Anda dapat menyiapkan rencana penskalaan untuk berbagai sumber daya, untuk masing-masing aplikasi. Auto Scaling (penskalaan otomatis) memberikan rekomendasi untuk strategi penyekalaan yang disesuaikan dengan setiap sumber daya. Setelah membuat rencana penskalaan, Auto Scaling (penskalaan otomatis) menggabungkan metode penskalaan dinamik dan penskalaan prediktif secara bersama-sama untuk mendukung strategi penskalaan Anda. Untuk detail selengkapnya, silakan lihat Cara kerja rencana penskalaan.
- Amazon EC2 Auto Scaling membantu Anda memastikan bahwa Anda memiliki jumlah instans Amazon EC2 yang tepat tersedia untuk menangani beban untuk aplikasi Anda. Anda membuat koleksi instans EC2, yang disebut grup Auto Scaling (penskalaan otomatis). Anda dapat menentukan jumlah instans minimum dan maksimum di setiap grup Auto Scaling, dan Amazon EC2 Auto Scaling akan memastikan bahwa grup Anda tidak pernah berada di bawah atau di atas batas yang ditentukan ini. Untuk detail selengkapnya, silakan lihat Apa itu Amazon EC2 Auto Scaling?
- Penskalaan otomatis Amazon DynamoDB menggunakan layanan Penskalaan Otomatis Aplikasi untuk secara dinamis menyesuaikan kapasitas throughput tersedia untuk merespons pola lalu lintas aktual. Ini memungkinkan tabel atau indeks sekunder global meningkatkan kapasitas baca dan tulis yang disediakan untuk menangani peningkatan lalu lintas tiba-tiba, tanpa throttling. Untuk detail lebih lanjut, silakan lihat Mengelola kapasitas throughput secara otomatis dengan menggunakan penskalaan otomatis DynamoDB.

Sumber daya

Praktik-praktik terbaik terkait:

- REL07-BP01 Menggunakan otomatisasi ketika mendapatkan atau menskalakan sumber daya
- REL11-BP01 Memantau semua komponen beban kerja untuk mendeteksi kegagalan

Dokumen terkait:

- AWS Auto Scaling: Cara Kerja Penskalaan
- Mengelola Kapasitas Throughput Secara Otomatis dengan Menggunakan Penskalaan Otomatis DynamoDB
- Apa itu Amazon EC2 Auto Scaling?

REL07-BP03 Menambah sumber daya berdasarkan deteksi bahwa beban kerja memerlukan lebih banyak sumber daya

Salah satu fitur yang paling berharga dari komputasi cloud adalah kemampuan untuk menyediakan sumber daya secara dinamis.

Dalam lingkungan komputasi on-premise tradisional, Anda harus mengidentifikasi dan menyediakan kapasitas yang cukup sebelumnya untuk melayani permintaan puncak. Hal ini menjadi masalah karena mahal dan menimbulkan risiko ketersediaan jika Anda meremehkan kebutuhan kapasitas puncak beban kerja.

Di cloud, Anda tidak perlu melakukan ini. Sebagai gantinya, Anda dapat menyediakan komputasi, basis data, dan kapasitas sumber daya lainnya sesuai kebutuhan untuk memenuhi permintaan saat ini maupun perkiraan. Solusi otomatis seperti Amazon EC2 Auto Scaling dan Application Auto Scaling dapat membuat sumber daya siap digunakan berdasarkan metrik yang Anda tentukan. Hal ini dapat membuat proses penskalaan lebih mudah dan terprediksi, serta dapat membuat beban kerja Anda jauh lebih andal dengan memastikan Anda memiliki ketersediaan sumber daya yang memadai setiap saat.

Hasil yang diinginkan: Anda mengonfigurasi penskalaan otomatis komputasi dan sumber daya lainnya untuk memenuhi permintaan. Anda menyediakan kapasitas tambahan yang cukup dalam kebijakan penskalaan Anda untuk memungkinkan lonjakan lalu lintas dilayani sementara sumber daya tambahan siap digunakan.

Anti-pola umum:

- Anda menyediakan sumber daya yang dapat diskalakan dengan jumlah tetap.
- Anda memilih metrik penskalaan yang tidak berkorelasi dengan permintaan aktual.
- Anda gagal menyediakan kapasitas tambahan yang cukup dalam rencana penskalaan Anda untuk mengakomodasi lonjakan permintaan.
- Kebijakan penskalaan Anda menambahkan kapasitas terlalu lambat, sehingga menyebabkan kehabisan kapasitas dan penurunan layanan saat sumber daya tambahan siap digunakan.
- Anda gagal mengonfigurasi jumlah sumber daya minimum dan maksimum dengan benar, sehingga menyebabkan kegagalan penskalaan.

Manfaat menjalankan praktik terbaik ini: Memiliki sumber daya yang cukup untuk memenuhi permintaan saat ini sangat penting untuk memberikan ketersediaan tinggi beban kerja Anda dan mematuhi tujuan tingkat layanan (SLO) yang Anda tetapkan. Penskalaan otomatis memungkinkan

Anda menyediakan jumlah komputasi, database, dan sumber daya lain yang dibutuhkan beban kerja Anda secara tepat untuk melayani permintaan saat ini dan yang diperkirakan. Anda tidak perlu menentukan kebutuhan kapasitas puncak dan mengalokasikan sumber daya secara statis untuk melayaninya. Sebaliknya, ketika permintaan meningkat, Anda dapat mengalokasikan lebih banyak sumber daya untuk mengakomodasi peningkatan tersebut, dan setelah permintaan menurun, Anda dapat menonaktifkan sumber daya untuk mengurangi biaya.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Sedang

Panduan implementasi

Pertama, tentukan apakah komponen beban kerja cocok untuk penskalaan otomatis. Komponen-komponen ini disebut dapat diskalakan secara horizontal karena menyediakan sumber daya yang sama dan berperilaku identik. Contoh komponen yang dapat diskalakan secara horizontal termasuk instans EC2 yang dikonfigurasi sama, tugas Amazon Elastic Container Service (ECS), dan pod yang berjalan di Amazon Elastic Kubernetes Service (EKS). Sumber daya komputasi ini biasanya terletak di belakang penyeimbang beban dan disebut sebagai replika.

Sumber daya lain yang direplikasi dapat meliputi replika baca database, tabel Amazon DynamoDB, dan klaster Amazon ElastiCache (Redis OSS). Untuk daftar lengkap sumber daya yang didukung, lihat Layanan AWS yang dapat Anda gunakan dengan Application Auto Scaling.

Untuk arsitektur berbasis kontainer, Anda mungkin perlu menskalakan dengan dua cara berbeda. Pertama, Anda mungkin perlu menskalakan kontainer yang menyediakan layanan yang dapat diskalakan secara horizontal. Kedua, Anda mungkin perlu menskalakan sumber daya komputasi untuk memberi ruang bagi kontainer baru. Mekanisme penskalaan otomatis yang berbeda ada untuk setiap lapisan. Untuk menskalakan tugas ECS, Anda dapat menggunakan Application Auto Scaling. Untuk menskalakan pod Kubernetes, Anda dapat menggunakan Horizontal Pod Autoscaler (HPA) atau Kubernetes Event-driven Autoscaling (KEDA). Untuk menskalakan sumber daya komputasi, Anda dapat menggunakan Penyedia Kapasitas untuk ECS, atau untuk Kubernetes, Anda dapat menggunakan Karpenter atau Penskala Otomatis Klaster.

Selanjutnya, pilih bagaimana Anda akan melakukan penskalaan otomatis. Ada tiga opsi utama: penskalaan berbasis metrik, penskalaan terjadwal, dan penskalaan prediktif.

Penskalaan berbasis metrik

Penskalaan berbasis metrik menyediakan sumber daya berdasarkan nilai satu metrik penskalaan atau lebih. Metrik penskalaan adalah metrik yang sesuai dengan permintaan beban kerja Anda. Cara

yang baik untuk menentukan metrik penskalaan yang tepat adalah dengan melakukan pengujian beban di lingkungan non-produksi. Selama pengujian beban Anda, jangan ubah jumlah sumber daya yang dapat diskalakan, dan tingkatkan permintaan secara bertahap (misalnya, throughput, konkurensi, atau pengguna simulasi). Kemudian cari metrik yang meningkat (atau menurun) seiring dengan peningkatan permintaan, dan menurun (atau meningkat) seiring dengan penurunan permintaan. Metrik penskalaan yang umum mencakup pemanfaatan CPU, kedalaman antrean kerja (seperti antrean Amazon SQS), jumlah pengguna aktif, dan throughput jaringan.

> **Note**
>
> AWS mengamati bahwa pada sebagian besar aplikasi, pemanfaatan memori meningkat saat aplikasi dimulai dan mencapai nilai yang stabil setelahnya. Ketika permintaan menurun, pemanfaatan memori biasanya tetap meningkat, bukan menurun secara paralel. Karena pemanfaatan memori tidak berkorelasi dengan permintaan di kedua arah–yaitu, naik dan turun mengikuti permintaan–pertimbangkan dengan cermat sebelum Anda memilih metrik ini untuk penskalaan otomatis.

Penskalaan berbasis metrik adalah operasi laten. Diperlukan waktu beberapa menit agar metrik pemanfaatan menyebar ke mekanisme penskalaan otomatis, dan mekanisme ini biasanya menunggu sinyal yang jelas akan peningkatan permintaan sebelum bereaksi. Kemudian, saat penskala otomatis menciptakan sumber daya baru, dibutuhkan waktu tambahan bagi sumber daya tersebut untuk sepenuhnya siap beroperasi. Karena itu, penting untuk tidak menetapkan target metrik penskalaan Anda terlalu dekat dengan pemanfaatan penuh (misalnya, 90% penggunaan CPU). Melakukan hal tersebut dapat menghabiskan kapasitas sumber daya yang ada sebelum kapasitas tambahan siap digunakan. Target penggunaan sumber daya tipikal dapat berkisar antara 50-70% untuk ketersediaan optimal, tergantung pada pola permintaan dan waktu yang diperlukan untuk menyediakan sumber daya tambahan.

Penskalaan terjadwal

Penskalaan terjadwal menyediakan atau menghapus sumber daya berdasarkan kalender atau waktu dalam satu hari. Hal ini sering digunakan untuk beban kerja yang memiliki permintaan yang dapat diprediksi, seperti pemanfaatan puncak selama jam kerja pada hari kerja atau aktivitas penjualan. Baik Amazon EC2 Auto Scaling dan Application Auto Scaling mendukung penskalaan terjadwal. Penskala cron KEDA mendukung penskalaan terjadwal pod Kubernetes.

Penskalaan prediktif

Penskalaan prediktif menggunakan machine learning untuk secara otomatis menskalakan sumber daya berdasarkan permintaan yang diantisipasi. Penskalaan prediktif menganalisis nilai historis dari metrik pemanfaatan yang Anda berikan dan terus memprediksi nilai masa depannya. Nilai yang diprediksi kemudian digunakan untuk meningkatkan atau menurunkan skala sumber daya. Amazon EC2 Auto Scaling dapat melakukan penskalaan prediktif.

Langkah-langkah implementasi

1. Tentukan apakah komponen beban kerja cocok untuk penskalaan otomatis.
2. Tentukan jenis mekanisme penskalaan yang paling sesuai untuk beban kerja: penskalaan berbasis metrik, penskalaan terjadwal, atau penskalaan prediktif.
3. Pilih mekanisme penskalaan otomatis yang sesuai untuk komponen. Untuk instans Amazon EC2, gunakan Amazon EC2 Auto Scaling. Untuk layanan AWS lainnya, gunakan Application Auto Scaling. Untuk pod Kubernetes (seperti yang berjalan di klaster Amazon EKS), pertimbangkan Horizontal Pod Autoscaler (HPA) atau Kubernetes Event-driven Autoscaling (KEDA). Untuk simpul Kubernetes atau EKS, pertimbangkan Karpenter dan Cluster Auto Scaler (CAS).
4. Untuk penskalaan metrik atau terjadwal, lakukan pengujian beban untuk menentukan metrik penskalaan dan nilai target yang sesuai untuk beban kerja Anda. Untuk penskalaan terjadwal, tentukan jumlah sumber daya yang dibutuhkan pada tanggal dan waktu yang Anda pilih. Tentukan jumlah maksimum sumber daya yang dibutuhkan untuk melayani lalu lintas puncak yang diharapkan.
5. Konfigurasikan penskala otomatis berdasarkan informasi yang dikumpulkan di atas. Lihat dokumentasi layanan penskalaan otomatis untuk detailnya. Verifikasi bahwa batas penskalaan maksimum dan minimum dikonfigurasi dengan benar.
6. Verifikasi bahwa konfigurasi penskalaan berfungsi sesuai harapan. Lakukan pengujian beban di lingkungan non-produksi dan amati bagaimana sistem bereaksi, dan sesuaikan sesuai kebutuhan. Saat mengaktifkan penskalaan otomatis dalam produksi, konfigurasikan alarm yang sesuai untuk memberi tahu Anda tentang perilaku yang tidak terduga.

Sumber daya

Dokumen terkait:

- Apa itu Amazon EC2 Auto Scaling?
- Panduan Preskriptif AWS: Memuat aplikasi pengujian
- AWS Marketplace: produk yang dapat digunakan dengan penskalaan otomatis

- Mengelola Kapasitas Throughput Secara Otomatis dengan Menggunakan Penskalaan Otomatis DynamoDB
- Penskalaan Prediktif untuk EC2, Didukung oleh Machine Learning
- Penskalaan terjadwal untuk Amazon EC2 Auto Scaling
- Telling Stories About Little's Law

REL07-BP04 Beban uji beban kerja Anda

Adopsi metodologi pengujian beban untuk mengukur apakah aktivitas penskalaan memenuhi persyaratan beban kerja.

Pengujian beban yang berkelanjutan penting untuk dilakukan. Tes beban harus menemukan titik puncaknya dan menguji kinerja beban kerja Anda. AWS memudahkan untuk mengatur lingkungan pengujian sementara yang memodelkan skala beban kerja produksi Anda. Di cloud, Anda dapat membuat sebuah lingkungan pengujian berskala produksi sesuai permintaan, menyelesaikan pengujian, dan kemudian menonaktifkan sumber dayanya. Karena Anda hanya membayar lingkungan pengujian saat sedang berjalan, Anda dapat menyimulasikan lingkungan langsung Anda dengan biaya yang lebih murah daripada pengujian on-premise.

Pengujian beban di lingkungan produksi juga harus dipertimbangkan sebagai bagian dari game day di mana sistem produksi diberikan tekanan, selama jam-jam penggunaan pelanggan yang lebih rendah, dengan semua personel yang siap untuk menerjemahkan hasilnya dan menangani masalah-masalah yang muncul.

Anti-pola umum:

- Melakukan pengujian beban di lingkungan deployment yang tidak memiliki konfigurasi yang sama dengan lingkungan produksi Anda.
- Melakukan pengujian beban hanya pada beban kerja Anda secara terpisah-pisah, bukan pada keseluruhan beban kerja.
- Melakukan pengujian beban dengan subset permintaan, bukan set permintaan riil yang representatif.
- Melakukan pengujian beban ke satu faktor keselamatan kecil di atas beban yang diharapkan.

Manfaat menerapkan praktik terbaik ini: Anda mengetahui komponen apa saja di dalam arsitektur Anda yang gagal saat menerima beban, dan mampu mengidentifikasi metrik apa saja yang perlu

diamati sebagai indikator bahwa Anda mendekati beban tersebut tepat waktu untuk mengatasi masalah dan mencegah dampak kegagalan tersebut.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Sedang

Panduan implementasi

- Lakukan pengujian beban untuk mengidentifikasi aspek mana dalam beban kerja Anda yang menunjukkan bahwa Anda harus menambah atau menghapus kapasitas. Pengujian beban harus memiliki lalu lintas representatif yang serupa dengan yang Anda terima di lingkungan produksi. Tingkatkan beban sambil mengamati metrik yang telah Anda instrumentasikan untuk menentukan metrik mana yang menunjukkan kapan Anda harus menambah atau menghapus sumber daya.
  - Pengujian Beban Terdistribusi pada AWS: mensimulasikan ribuan pengguna yang terhubung
    - Identifikasi gabungan permintaan. Anda mungkin memiliki gabungan permintaan yang beragam, sehingga Anda harus melihat berbagai kerangka waktu saat mengidentifikasi gabungan lalu lintas tersebut.
    - Implementasikan pendorong beban. Anda dapat menggunakan perangkat lunak kode kustom, sumber terbuka, atau komersial untuk mengimplementasikan pendorong beban.
    - Lakukan uji beban di awal dengan menggunakan kapasitas kecil. Anda melihat beberapa efek langsung dengan mendorong beban ke kapasitas yang lebih kecil, kemungkinan seukuran satu instans atau kontainer.
    - Uji beban dengan kapasitas yang lebih besar. Efek-efek tersebut akan berbeda di beban yang terdistribusi, sehingga Anda harus melakukan pengujian di lingkungan yang semirip mungkin dengan lingkungan produksi.

Sumber daya

Dokumen terkait:

- Pengujian Beban Terdistribusi pada AWS: mensimulasikan ribuan pengguna yang terhubung
- Aplikasi pengujian beban

Video terkait:

- AWS Summit ANZ 2023: Mempercepat dengan percaya diri melalui Pengujian Beban AWS Terdistribusi

Implementasikan perubahan

Perubahan-perubahan terkontrol diperlukan untuk melakukan deployment fungsionalitas baru, dan untuk memastikan bahwa beban kerja dan lingkungan operasi menjalankan perangkat lunak yang dikenal dan dapat di-patch dengan baik. Jika perubahan-perubahan ini tidak terkontrol, maka akan sulit untuk memprediksi efek dari perubahan-perubahan tersebut, atau untuk mengatasi masalah yang timbul sebagai akibatnya.

Pola deployment tambahan untuk meminimalkan risiko

Bendera fitur (juga dikenal sebagai fitur toggle) adalah opsi konfigurasi pada sebuah aplikasi. Anda dapat men-deploy perangkat lunak dengan fitur yang dimatikan, sehingga pelanggan Anda tidak melihat fitur tersebut. Anda kemudian dapat mengaktifkan fitur tersebut, seperti yang akan Anda lakukan untuk deployment canary, atau Anda dapat mengatur kecepatan perubahan menjadi 100% untuk melihat efeknya. Jika deployment bermasalah, Anda cukup mematikan fitur tanpa perlu melakukan roll back.

Deployment zona terisolasi kesalahan: Salah satu aturan terpenting yang telah ditetapkan AWS untuk deployment-nya sendiri adalah menghindari tersentuhnya beberapa Zona Ketersediaan dalam suatu Wilayah secara bersamaan. Hal ini penting untuk memastikan bahwa Zona Ketersediaan tersebut sudah bersifat independen untuk tujuan perhitungan ketersediaan kami. Sebaiknya Anda menggunakan pertimbangan serupa dalam deployment Anda.

Peninjauan Kesiapan Operasional (ORR)

AWS merasakan bahwa akan sangat bermanfaat untuk melakukan peninjauan kesiapan operasional yang mengevaluasi kelengkapan pengujian, kemampuan untuk memantau, dan yang penting, kemampuan untuk mengaudit kinerja aplikasi ke SLA dan memberikan data jika terjadi gangguan atau anomali operasional lainnya. ORR formal dilakukan sebelum deployment produksi awal. AWS akan mengulangi ORR secara berkala (sekali per tahun, atau sebelum periode kinerja kritis) untuk memastikan bahwa tidak terjadi penyimpangan dari ekspektasi operasional. Untuk informasi lebih lanjut tentang kesiapan operasional ini, silakan lihat Pilar Keunggulan Operasional dari Kerangka Kerja AWS Well-Architected.

Praktik terbaik

- REL08-BP01 Menggunakan runbook untuk aktivitas standar seperti deployment
- REL08-BP02 Integrasikan pengujian fungsional sebagai bagian dari deployment Anda
- REL08-BP03 Mengintegrasikan pengujian ketahanan sebagai bagian dari deployment Anda

- REL08-BP04 Melakukan deployment dengan menggunakan infrastruktur yang tidak bisa diubah
- REL08-BP05 Melakukan deployment perubahan dengan otomatisasi

REL08-BP01 Menggunakan runbook untuk aktivitas standar seperti deployment

Runbook adalah prosedur terdokumentasi untuk mencapai hasil-hasil tertentu. Gunakan runbook untuk melakukan aktivitas-aktivitas standar, baik yang dilakukan secara manual maupun otomatis. Contohnya adalah men-deploy beban kerja, mem-patch beban kerja, atau membuat modifikasi DNS.

Misalnya, terapkan proses untuk memastikan keamanan rollback selama deployment. Memastikan bahwa Anda dapat membatalkan deployment tanpa menimbulkan gangguan terhadap pelanggan adalah sesuatu yang penting dalam menciptakan keandalan layanan.

Untuk prosedur runbook, mulailah dengan proses manual efektif yang valid, implementasikan dalam kode, dan lakukan invokasi agar berjalan secara otomatis saat diperlukan.

Bahkan untuk beban kerja canggih yang sangat otomatis, runbook masih berguna untuk menjalankan game day atau memenuhi persyaratan pelaporan dan audit yang ketat.

Perlu diperhatikan bahwa playbook digunakan untuk merespons insiden-insiden tertentu, sedangkan runbook digunakan untuk meraih hasil-hasil tertentu. Sering kali, runbook ditujukan untuk aktivitas rutin, sedangkan playbook digunakan untuk merespons peristiwa-peristiwa non-rutin.

Anti-pola umum:

- Melakukan perubahan tidak terencana pada konfigurasi di lingkungan produksi.
- Melewatkan langkah-langkah yang diuraikan dalam rencana Anda untuk melakukan deployment yang lebih cepat, sehingga mengakibatkan deployment mengalami kegagalan.
- Membuat perubahan tanpa melakukan uji pembatalan perubahan.

Manfaat menjalankan praktik terbaik ini: Perencanaan perubahan yang efektif meningkatkan kemampuan Anda untuk berhasil menjalankan perubahan karena Anda mengetahui semua sistem yang terpengaruh. Validasi perubahan di lingkungan pengujian meningkatkan kepercayaan diri Anda.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Tinggi

Panduan implementasi

- Sediakan respons yang cepat dan konsisten terhadap peristiwa yang dipahami dengan baik dengan cara membuat dokumentasi prosedur-prosedur penanganan peristiwa di dalam runbook.
- Gunakan prinsip infrastruktur sebagai kode untuk menentukan infrastruktur Anda. Dengan menggunakan AWS CloudFormation (atau pihak ketiga tepercaya) untuk menentukan infrastruktur Anda, Anda dapat menggunakan perangkat lunak kontrol versi untuk membuat versi baru dan melacak perubahan.
  - Gunakan AWS CloudFormation (atau penyedia pihak ketiga tepercaya) untuk menentukan infrastruktur Anda.
    - Apa itu AWS CloudFormation?
  - Buat templat yang bersifat tunggal dan terpisah-pisah, dengan menggunakan prinsip desain perangkat lunak yang baik.
    - Tentukan izin, templat, dan pihak-pihak yang bertanggung jawab untuk implementasi.
      - Mengendalikan akses dengan AWS Identity and Access Management
    - Gunakan sistem manajemen kode sumber yang di-hosting berdasarkan teknologi populer seperti Git untuk menyimpan kode sumber dan konfigurasi infrastruktur sebagai kode (IaC) Anda.

Sumber daya

Dokumen terkait:

- Partner APN: partner yang dapat membantu Anda membuat solusi deployment yang diotomatisasi
- AWS Marketplace: produk yang dapat digunakan untuk mengotomatisasi deployment Anda
- Apa itu AWS CloudFormation?

Contoh terkait:

- Melakukan otomatisasi operasi dengan Playbook dan Runbook

REL08-BP02 Integrasikan pengujian fungsional sebagai bagian dari deployment Anda

Gunakan teknik seperti pengujian unit dan pengujian integrasi yang memvalidasi fungsionalitas.

Pengujian unit adalah proses pengujian unit fungsional terkecil dari kode untuk memvalidasi perilakunya. Pengujian integrasi berupaya untuk memvalidasi bahwa setiap fitur aplikasi bekerja sesuai dengan persyaratan perangkat lunak. Sementara pengujian unit berfokus pada pengujian bagian dari aplikasi secara terpisah, pengujian integrasi mempertimbangkan efek samping (misalnya, efek dari data yang diubah melalui operasi mutasi). Dalam kedua kasus tersebut, pengujian harus diintegrasikan ke dalam pipeline deployment, dan jika kriteria keberhasilan tidak terpenuhi, pipeline dihentikan atau di-rollback. Pengujian ini dijalankan dalam lingkungan pra-produksi, yang di-staging sebelum produksi dalam pipeline.

Anda akan meraih hasil terbaik saat pengujian ini dijalankan secara otomatis sebagai bagian dari tindakan deployment dan build. Misalnya, dengan AWS CodePipeline, developer melakukan perubahan pada repositori sumber tempat CodePipeline mendeteksi perubahan secara otomatis. Aplikasi dibangun, dan pengujian unit dijalankan. Setelah pengujian selesai, kode yang dibangun di-deploy ke server staging untuk pengujian. Dari server staging, CodePipeline menjalankan lebih banyak pengujian, seperti integrasi atau pengujian beban. Setelah berhasil menyelesaikan pengujian tersebut, CodePipeline melakukan deployment kode yang telah diuji dan disetujui ke instans produksi.

Hasil yang diinginkan: Anda menggunakan otomatisasi untuk melakukan pengujian unit dan pengujian integrasi untuk memvalidasi bahwa kode Anda berperilaku sesuai harapan. Pengujian ini diintegrasikan ke dalam proses deployment, dan kegagalan pengujian membatalkan deployment.

Anti-pola umum:

- Anda mengabaikan atau tidak menghiraukan kegagalan dan rencana pengujian selama proses deployment untuk mempercepat jadwal deployment.
- Anda melakukan pengujian secara manual di luar pipeline deployment.
- Anda melewatkan langkah-langkah pengujian dalam otomatisasi melalui alur kerja darurat manual.
- Anda menjalankan pengujian otomatis di lingkungan yang tidak mirip dengan lingkungan produksi.
- Anda membangun rangkaian pengujian yang tidak cukup fleksibel dan sulit untuk dipelihara, diperbarui, atau diskalakan seiring aplikasi berkembang.

Manfaat menjalankan praktik terbaik ini: Pengujian otomatis selama proses deployment akan menemukan masalah lebih awal, sehingga mengurangi risiko rilis ke produksi dengan bug atau perilaku tak terduga. Pengujian unit memvalidasi bahwa kode berperilaku sesuai keinginan dan mematuhi kontrak API. Pengujian integrasi memvalidasi bahwa sistem beroperasi sesuai dengan

persyaratan yang ditentukan. Jenis-jenis pengujian ini memverifikasi fungsi komponen yang diinginkan, seperti antarmuka pengguna, API, basis data, dan kode sumber secara konsisten.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Tinggi

Panduan implementasi

Adopsi pendekatan pengembangan berbasis pengujian (TDD) untuk menulis perangkat lunak agar Anda dapat mengembangkan kasus pengujian untuk menentukan dan memvalidasi kode Anda. Untuk memulai, buat kasus pengujian untuk setiap fungsi. Jika pengujian gagal, Anda menulis kode baru yang cukup untuk memenuhi persyaratan pengujian. Pendekatan ini membantu Anda memvalidasi hasil yang diharapkan dari setiap fungsi. Jalankan pengujian unit dan validasi bahwa pengujian tersebut dijalankan tanpa kesalahan sebelum Anda menyimpan perubahan kode ke repositori kode sumber.

Implementasikan pengujian unit dan pengujian integrasi sebagai bagian dari tahap pembuatan, pengujian, dan deployment pipeline CI/CD. Otomatiskan pengujian, dan mulai pengujian secara otomatis setiap kali versi baru aplikasi siap di-deploy. Jika kriteria keberhasilan tidak terpenuhi, maka alur akan dihentikan atau di-rollback.

Jika aplikasi tersebut adalah aplikasi web atau seluler, lakukan pengujian integrasi otomatis pada beberapa browser desktop atau perangkat sebenarnya. Pendekatan ini sangat berguna untuk memvalidasi kompatibilitas dan fungsionalitas aplikasi seluler di berbagai perangkat.

Langkah-langkah implementasi

1. Tulis pengujian unit sebelum Anda menulis kode fungsional (pengembangan berbasis pengujian, atau TDD). Buat pedoman kode sehingga menulis dan menjalankan pengujian unit menjadi persyaratan pengodean non-fungsional.
2. Buat serangkaian pengujian integrasi otomatis yang mencakup fungsionalitas yang dapat diuji yang telah diidentifikasi. Serangkaian pengujian tersebut harus menyimulasikan interaksi pengguna dan memvalidasi hasil yang diharapkan.
3. Buat lingkungan pengujian yang diperlukan untuk menjalankan pengujian integrasi. Hal ini mungkin termasuk lingkungan staging atau pra-produksi yang sangat mirip dengan lingkungan produksi.
4. Siapkan tahap sumber, pembuatan, pengujian, dan deployment menggunakan konsol AWS CodePipeline atau AWS Command Line Interface (CLI).
5. Lakukan deployment aplikasi setelah kode telah dibangun dan diuji. AWS CodeDeploy dapat melakukan deployment aplikasi ke lingkungan staging (pengujian) dan produksi Anda. Lingkungan

ini dapat mencakup instans Amazon EC2, fungsi AWS Lambda, atau server on-premise. Mekanisme deployment yang sama harus digunakan untuk melakukan deployment aplikasi ke semua lingkungan.

6. Pantau progres pipeline Anda dan status setiap tahap. Gunakan pemeriksaan kualitas untuk memblokir pipeline berdasarkan status pengujian Anda. Anda juga dapat menerima notifikasi untuk setiap kegagalan tahap pipeline atau penyelesaian pipeline.
7. Terus pantau hasil pengujian, dan cari pola, regresi, atau area yang membutuhkan perhatian lebih. Gunakan informasi ini untuk memperbaiki rangkaian pengujian, mengidentifikasi area aplikasi yang membutuhkan pengujian yang lebih ketat, dan mengoptimalkan proses deployment.

Sumber daya

Praktik-praktik terbaik terkait:

- REL07-BP04 Menguji beban untuk beban kerja Anda
- REL08-BP03 Mengintegrasikan pengujian ketahanan sebagai bagian dari deployment Anda
- REL12-BP04 Menguji ketahanan menggunakan chaos engineering

Dokumen terkait:

- Panduan Preskriptif AWS: Otomatisasi pengujian
- Pengiriman Berkelanjutan dan Integrasi Berkelanjutan
- Indikator untuk pengujian fungsional
- Memantau pipeline
- Gunakan AWS CodePipeline dengan AWS CodeBuild untuk menguji kode dan menjalankan build
- AWS Device Farm

REL08-BP03 Mengintegrasikan pengujian ketahanan sebagai bagian dari deployment Anda

Integrasikan pengujian ketahanan dengan memasukkan kegagalan secara sadar ke dalam sistem Anda guna mengukur kemampuannya apabila terjadi skenario yang mengganggu. Pengujian ketahanan berbeda dari pengujian unit dan fungsi yang biasanya terintegrasi dalam siklus deployment, karena pengujian ini berfokus pada identifikasi terhadap kegagalan-kegagalan yang

tidak terduga di dalam sistem Anda. Meskipun memulai dengan integrasi pengujian ketahanan dalam tahap praproduksi aman dilakukan, tetapkan tujuan untuk mengimplementasikan pengujian ini dalam produksi sebagai bagian dari game day Anda.

Hasil yang diinginkan: Pengujian ketahanan akan membantu Anda membangun kepercayaan pada kemampuan sistem untuk bertahan dari degradasi dalam produksi. Eksperimen mengidentifikasi titik lemah yang dapat menyebabkan kegagalan, yang akan membantu Anda meningkatkan kualitas sistem Anda untuk mengurangi terjadinya kegagalan dan penurunan kualitas secara otomatis dan efisien.

Anti-pola umum:

- Kurangnya observabilitas dan pemantauan dalam proses deployment
- Ketergantungan pada manusia untuk mengatasi kegagalan sistem
- Mekanisme analisis kualitas yang buruk
- Fokus pada masalah yang diketahui dalam suatu sistem dan kurangnya eksperimen untuk mengidentifikasi masalah yang belum diketahui
- Identifikasi kegagalan yang tidak mempunyai penyelesaian
- Tidak ada dokumentasi temuan dan runbook

Manfaat menerapkan praktik terbaik: Pengujian ketahanan yang terintegrasi dalam deployment Anda akan membantu Anda dalam mengidentifikasi masalah yang tidak diketahui yang terjadi dalam sistem yang tidak diketahui, yang dapat menyebabkan waktu henti dalam produksi. Melakukan identifikasi terhadap masalah-masalah yang tidak diketahui di dalam sistem akan membantu Anda mendokumentasikan temuan, mengintegrasikan pengujian ke dalam proses CI/CD Anda, dan membangun runbook, yang pada akhirnya akan menyederhanakan mitigasi melalui mekanisme yang efisien dan dapat diulang.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Sedang

Panduan implementasi

Bentuk pengujian ketahanan yang paling umum yang dapat diintegrasikan dalam deployment sistem Anda adalah pemulihan bencana dan rekayasa kekacauan (chaos engineering).

- Sertakan pembaruan pada rencana pemulihan bencana dan prosedur operasi standar (SOP) Anda dengan melakukan deployment yang signifikan.

- Integrasikan pengujian keandalan ke dalam pipeline deployment otomatis Anda. Layanan-layanan seperti AWS Resilience Hub dapat diintegrasikan ke dalam pipeline CI/CD Anda untuk membuat penilaian ketahanan berkelanjutan yang secara otomatis dievaluasi sebagai bagian dari setiap deployment.
- Tentukan aplikasi Anda di AWS Resilience Hub. Penilaian ketahanan akan menghasilkan cuplikan kode yang dapat membantu Anda dalam menciptakan prosedur pemulihan dalam bentuk dokumen AWS Systems Manager untuk aplikasi Anda dan menyediakan daftar monitor dan alarm Amazon CloudWatch yang direkomendasikan.
- Setelah rencana DR dan SOP Anda sudah diperbarui, selesaikan pengujian pemulihan bencana untuk memverifikasi efektivitasnya. Pengujian pemulihan bencana akan membantu Anda untuk menentukan apakah Anda dapat memulihkan sistem setelah terjadinya peristiwa tertentu dan kembali ke beroperasi secara normal. Anda dapat membuat simulasi dari berbagai strategi pemulihan bencana dan mengidentifikasi apakah perencanaan Anda sudah memadai untuk memenuhi persyaratan-persyaratan waktu aktif Anda. Strategi pemulihan bencana yang biasanya dibuat mencakup pencadangan dan pemulihan, pilot light, cold standby, warm standby, hot standby, dan active-active, dan semuanya memerlukan biaya dan memiliki kompleksitas yang berbeda-beda. Sebelum pengujian pemulihan bencana, kami menyarankan Anda untuk menentukan sasaran waktu pemulihan (RTO) dan sasaran titik pemulihan (RPO) Anda untuk menyederhanakan pilihan strategi yang akan disimulasikan. AWS menawarkan alat pemulihan bencana seperti AWS Elastic Disaster Recovery untuk membantu Anda memulai perencanaan dan pengujian Anda.
- Eksperimen chaos engineering memasukkan gangguan (disruptions) ke dalam sistem, seperti pemadaman jaringan dan kegagalan layanan. Dengan melakukan simulasi dengan kegagalan terkontrol, Anda akan dapat menemukan kerentanan sistem Anda sambil mengendalikan dampak-dampak yang ditimbulkan oleh kegagalan yang dimasukkan. Sama seperti strategi lainnya, jalankan simulasi kegagalan terkontrol di lingkungan non-produksi dengan menggunakan layanan-layanan seperti AWS Fault Injection Service untuk mendapatkan kepercayaan diri sebelum melakukan deployment di lingkungan produksi.

Sumber daya

Dokumen terkait:

- Bereksperimen dengan kegagalan menggunakan pengujian ketahanan untuk membangun kesiapan pemulihan
- Terus-menerus menilai ketahanan aplikasi dengan AWS Resilience Hub dan AWS CodePipeline

- Arsitektur Pemulihan Bencana (DR) di AWS, bagian 1: Strategi untuk Pemulihan di Cloud
- Verifikasi ketahanan beban kerja Anda dengan menggunakan Chaos Engineering
- Prinsip-prinsip Chaos Engineering
- Lokakarya Chaos Engineering

Video terkait:

- AWS re:Invent 2020: Menguji Ketahanan menggunakan Perekayasaan Chaos
- Meningkatkan Ketahanan Aplikasi dengan Layanan Injeksi Kesalahan AWS
- Mempersiapkan & Melindungi Aplikasi Anda Dari Gangguan Dengan AWS Resilience Hub

REL08-BP04 Melakukan deployment dengan menggunakan infrastruktur yang tidak bisa diubah

Infrastruktur tetap adalah model yang menuntut bahwa tidak ada pembaruan, patch keamanan, atau perubahan konfigurasi yang terjadi di tempat pada beban kerja produksi. Saat perubahan diperlukan, arsitektur dibangun ke infrastruktur baru dan di-deploy ke dalam lingkungan produksi.

Ikuti strategi penerapan infrastruktur tetap untuk meningkatkan keandalan, konsistensi, dan keterulangan (reproducibility) dalam deployment beban kerja Anda.

Hasil yang diinginkan: Dengan infrastruktur yang tidak dapat diubah, tidak ada modifikasi di tempat yang diizinkan untuk menjalankan sumber daya infrastruktur dalam sebuah beban kerja. Sebaliknya, ketika ada sebuah perubahan yang diperlukan, kumpulan sumber daya infrastruktur baru yang sudah diperbarui, yang berisi semua perubahan yang diperlukan, di-deploy secara paralel dengan sumber daya Anda yang ada. Deployment ini divalidasi secara otomatis, dan jika berhasil, lalu lintas akan dialihkan secara bertahap ke kumpulan sumber daya baru.

Strategi deployment ini berlaku di antaranya untuk pembaruan perangkat lunak, patch keamanan, perubahan infrastruktur, pembaruan konfigurasi, dan pembaruan aplikasi.

Anti-pola umum:

- Menerapkan perubahan di tempat untuk menjalankan sumber daya infrastruktur.

Manfaat menjalankan praktik terbaik ini:

- Peningkatan konsistensi di seluruh lingkungan: Karena tidak ada perbedaan dalam sumber daya infrastruktur di seluruh lingkungan, maka konsistensi ditingkatkan dan pengujian disederhanakan.
- Mengurangi penyimpangan konfigurasi: Dengan mengganti sumber daya infrastruktur dengan konfigurasi yang diketahui dan dikontrol versi, infrastruktur tersebut diatur ke status yang diketahui, diuji, dan tepercaya, menghindari penyimpangan konfigurasi.
- Deployment atom yang andal: Deployment berhasil diselesaikan atau tidak ada yang berubah, meningkatkan konsistensi dan keandalan dalam proses deployment.
- Deployment disederhanakan: Deployment disederhanakan karena tidak memerlukan pembaruan dukungan. Pembaruan hanyalah berupa deployment baru.
- Deployment yang lebih aman dengan proses rollback dan pemulihan yang cepat: Deployment lebih aman karena versi sebelumnya yang digunakan tidak diubah. Anda dapat melakukan rollback jika ada kesalahan yang terdeteksi.
- Postur keamanan yang ditingkatkan: Dengan tidak mengizinkan perubahan infrastruktur, mekanisme akses jarak jauh (seperti SSH) dapat Anda nonaktifkan. Hal ini akan mengurangi vektor serangan, sehingga meningkatkan postur keamanan organisasi.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Sedang

Panduan implementasi

Otomatisasi

Saat menentukan strategi deployment infrastruktur yang tidak dapat diubah, Anda sebaiknya menggunakan otomatisasi sebanyak mungkin untuk meningkatkan reproduksibilitas dan meminimalkan potensi kesalahan manusia. Untuk detail selengkapnya, silakan lihat REL08-BP05 Melakukan deployment perubahan dengan otomatisasi dan Melakukan otomatisasi deployment secara aman dan otonom.

Dengan infrastruktur sebagai kode (IaC), langkah-langkah penyediaan infrastruktur, orkestrasi, dan deployment ditentukan dengan cara terprogram, deskriptif, dan deklaratif dan disimpan dalam sistem kontrol sumber. Memanfaatkan infrastruktur sebagai kode akan makin memudahkan otomatisasi deployment infrastruktur dan membantu Anda mewujudkan ketetapan (immutability) infrastruktur.

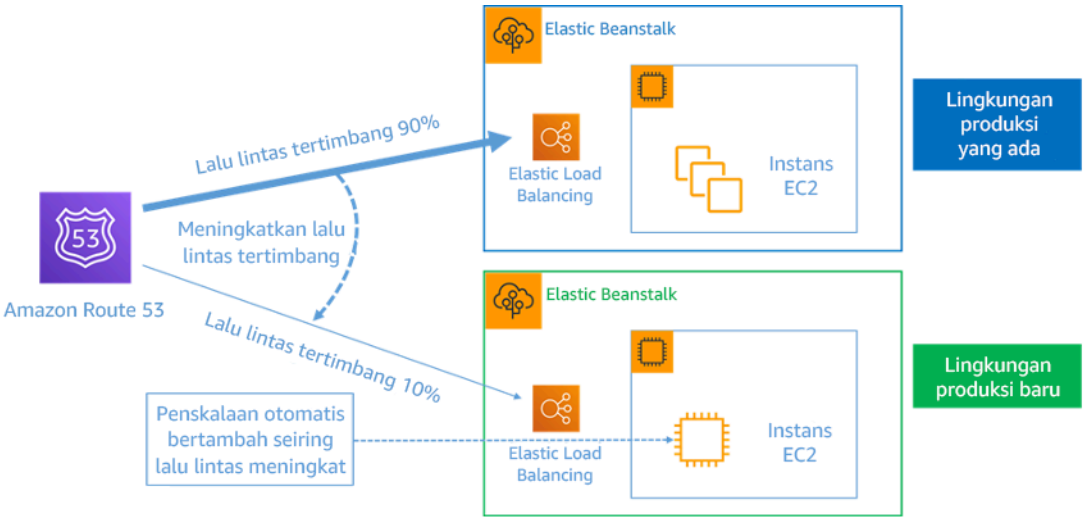
Pola deployment

Ketika ada perubahan dalam beban kerja yang diperlukan, strategi deployment tetap mengharuskan deployment sumber daya infrastruktur yang baru, termasuk semua perubahan yang diperlukan.

Kumpulan sumber daya baru ini harus mengikuti pola rollout yang meminimalkan dampak terhadap pengguna. Ada dua strategi utama untuk deployment ini:

Deployment canary: Praktik ini mengarahkan sejumlah kecil pelanggan kepada versi baru, yang biasanya dijalankan di sebuah instans layanan tunggal (canary). Lalu, Anda meneliti secara mendalam setiap perubahan perilaku atau kesalahan yang dihasilkan. Anda dapat menghapus lalu lintas dari canary jika menemui masalah-masalah kritis dan mengembalikan pengguna ke versi sebelumnya. Jika deployment berhasil, Anda dapat melanjutkan melakukan deployment pada kecepatan yang diinginkan, sambil memantau perubahan kesalahan, hingga Anda di-deploy sepenuhnya. AWS CodeDeploy dapat dikonfigurasi dengan konfigurasi deployment yang akan mengaktifkan deployment canary.

Deployment blue/green: Deployment yang bersifat serupa dengan deployment canary, kecuali armada penuh aplikasi di-deploy secara paralel. Anda mengubah deployment Anda di dua tumpukan (blue and green). Sekali lagi, Anda mengirimkan lalu lintas ke versi yang baru, dan kembali ke versi lama jika Anda melihat ada masalah yang terjadi dengan deployment. Biasanya semua lalu lintas dialihkan sekaligus, tetapi Anda juga dapat menggunakan sebagian lalu lintas ke setiap versi untuk meningkatkan adopsi versi baru menggunakan kemampuan perutean DNS tertimbang dari Amazon Route 53. AWS CodeDeploy dan AWS Elastic Beanstalk dapat dikonfigurasikan dengan konfigurasi deployment yang memungkinkan deployment blue/green.



Gambar 8: Deployment blue/green dengan AWS Elastic Beanstalk dan Amazon Route 53

Deteksi penyimpangan

Penyimpangan didefinisikan sebagai setiap perubahan yang menyebabkan sumber daya infrastruktur memiliki status atau konfigurasi yang berbeda dengan apa yang diharapkan. Setiap jenis perubahan konfigurasi yang tidak terkelola bertentangan dengan gagasan infrastruktur tetap dan tidak

bisa diubah (immutable), dan harus dideteksi dan diperbaiki agar infrastruktur tetap berhasil diimplementasikan.

Langkah-langkah implementasi

- Larang modifikasi di tempat pada sumber daya infrastruktur yang sedang berjalan.
  - Anda dapat menggunakan AWS Identity and Access Management (IAM) untuk menentukan siapa atau apa yang dapat mengakses layanan dan sumber daya yang ada di AWS, mengelola izin terperinci secara terpusat, dan menganalisis akses untuk menyempurnakan izin di seluruh AWS.
- Lakukan otomatisasi terhadap deployment sumber daya infrastruktur untuk meningkatkan reproduksibilitas dan meminimalkan terjadinya potensi kesalahan manusia.
  - Seperti yang dijelaskan dalam Pengantar DevOps di laporan resmi AWS, otomatisasi adalah sebuah landasan dengan layanan AWS dan didukung secara internal di semua layanan, fitur, dan penawaran.
  - Melakukan prebaking terhadap Amazon Machine Image (AMI) Anda dapat mempercepat waktu peluncurannya. EC2 Image Builder adalah sebuah layanan AWS terkelola penuh yang dapat membantu Anda mengotomatiskan pembuatan, pemeliharaan, validasi, berbagi, dan deployment AMI kustom Linux atau Windows yang aman dan terbaru.
  - Beberapa layanan yang mendukung otomatisasi adalah:
    - AWS Elastic Beanstalk adalah sebuah layanan untuk melakukan deployment dan menskalakan aplikasi web dan layanan yang dikembangkan dengan Java, NET, PHP, Node.js, Python, Ruby, Go, dan Docker pada server-server yang sudah dikenal seperti Apache, NGINX, Passenger, dan IIS.
    - AWS Proton membantu tim platform untuk menghubungkan dan mengoordinasikan semua alat yang berbeda yang dibutuhkan tim pengembangan Anda untuk melakukan penyediaan infrastruktur, deployment kode, pemantauan, dan pembaruan. AWS Proton mengaktifkan infrastruktur otomatis sebagai penyediaan kode dan deployment aplikasi nirserver dan berbasis kontainer.
  - Memanfaatkan infrastruktur sebagai kode akan memudahkan Anda dalam melakukan otomatisasi deployment infrastruktur, dan membantu mewujudkan ketetapan infrastruktur. AWS menyediakan layanan yang memungkinkan pembuatan, deployment, dan pemeliharaan infrastruktur dengan cara yang terprogram, deskriptif, dan deklaratif.
    - AWS CloudFormation membantu para developer membuat sumber daya AWS secara teratur dan dapat diprediksi. Sumber daya ditulis dalam file teks menggunakan format JSON atau

YAML. Templat memerlukan sebuah sintaks dan struktur tertentu sesuai dengan jenis sumber daya yang sedang dibuat dan dikelola. Anda menulis sumber daya Anda dalam JSON atau YAML dengan editor kode apa pun, melakukan check-in sumber daya ke dalam sebuah sistem kontrol versi, lalu CloudFormation membangun layanan yang ditentukan dengan cara yang aman dan dapat diulang.

  - AWS Serverless Application Model (AWS SAM) merupakan sebuah kerangka kerja sumber terbuka yang dapat Anda gunakan untuk membangun aplikasi nirserver di AWS. AWS SAM terintegrasi dengan layanan-layanan AWS lain, dan merupakan ekstensi dari CloudFormation.
  - AWS Cloud Development Kit (AWS CDK) adalah sebuah kerangka kerja pengembangan perangkat lunak sumber terbuka untuk membuat model dan menyediakan sumber daya aplikasi cloud Anda menggunakan bahasa pemrograman yang sudah dikenal. Anda dapat menggunakan AWS CDK untuk membuat model infrastruktur aplikasi dengan menggunakan TypeScript, Python, Java, dan .NET. AWS CDK menggunakan CloudFormation di latar belakang untuk menyediakan sumber daya dengan cara yang aman dan dapat diulang.
  - AWS Cloud Control API memperkenalkan satu set umum API Buat, Baca, Perbarui, Hapus, dan Daftar (CRUDL) untuk membantu para developer mengelola infrastruktur cloud mereka dengan cara yang mudah dan konsisten. Cloud Control API, adalah API umum yang memungkinkan para developer mengelola siklus hidup AWS dan layanan pihak ketiga secara seragam.
- Implementasikan pola deployment yang meminimalkan dampak-dampaknya terhadap pengguna.
  - Deployment canary:
    - Siapkan deployment rilis canary API Gateway
    - Buat pipeline dengan deployment canary untuk Amazon ECS dengan menggunakan AWS App Mesh
  - Deployment blue/green: Deployment Blue/Green, pada laporan resmi AWS dijelaskan contoh teknik untuk menerapkan strategi deployment blue/green.
- Deteksi konfigurasi atau penyimpangan status. Untuk informasi selengkapnya, silakan lihat Mendeteksi perubahan konfigurasi yang tidak terkelola pada tumpukan dan sumber daya.

Sumber daya

Praktik-praktik terbaik terkait:

- REL08-BP05 Melakukan deployment perubahan dengan otomatisasi

Dokumen terkait:

- Melakukan otomatisasi deployment secara aman dan otonom
- Memanfaatkan AWS CloudFormation untuk menciptakan sebuah infrastruktur yang tidak dapat diubah di Nubank
- Infrastruktur sebagai kode
- Menerapkan alarm untuk secara otomatis mendeteksi penyimpangan yang terjadi di tumpukan AWS CloudFormation

Video terkait:

- AWS re:Invent 2020: Keandalan, konsistensi, dan kepercayaan diri melalui kekekalan (immutability)

REL08-BP05 Melakukan deployment perubahan dengan otomatisasi

Deployment dan patching diotomatisasi untuk menghilangkan dampak-dampak negatif.

Membuat perubahan pada sistem produksi adalah salah satu area risiko terbesar bagi banyak organisasi. Kami menganggap deployment sebagai masalah kelas pertama untuk diatasi bersama dengan masalah-masalah bisnis yang ditangani oleh perangkat lunak. Saat ini, ini artinya penggunaan otomatisasi kapan saja memungkinkan dalam operasi, termasuk untuk menguji dan melakukan deployment perubahan, menambah atau menghapus kapasitas, dan memigrasikan data.

Hasil yang diinginkan: Anda membangun keamanan deployment otomatis ke dalam proses rilis dengan pengujian pra-produksi yang ekstensif, rollback otomatis, dan deployment produksi yang sangat baik. Otomatisasi ini meminimalkan potensi dampak pada produksi yang disebabkan oleh deployment yang gagal, dan developer tidak perlu lagi mengawasi tahapan deployment hingga produksi secara aktif.

Anti-pola umum:

- Anda melakukan perubahan secara manual.
- Anda melewatkan langkah-langkah dalam otomatisasi Anda melalui alur kerja darurat manual.
- Anda tidak mengikuti rencana dan proses yang telah ditetapkan demi mempercepat kronologi (timeline).
- Anda melakukan deployment susulan cepat tanpa menyediakan waktu menanam.

Manfaat menjalankan praktik terbaik ini: Ketika Anda menggunakan otomatisasi untuk melakukan deployment atas semua perubahan, Anda menghapus kemungkinan adanya kesalahan manusia dan memberikan kemampuan untuk melakukan pengujian sebelum Anda mengubahnya ke tahap produksi. Melakukan proses ini sebelum deployment di lingkungan produksi (push) dapat memverifikasi bahwa rencana Anda sudah lengkap. Selain itu, rollback otomatis ke dalam proses rilis Anda dapat mengidentifikasi masalah-masalah produksi dan mengembalikan beban kerja Anda ke keadaan operasional yang diketahui berfungsi sebelumnya.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Sedang

Panduan implementasi

Lakukan otomatisasi terhadap pipeline deployment Anda. Pipeline deployment memungkinkan Anda untuk menginvokasi pengujian dan deteksi anomali secara otomatis, serta memberi Anda pilihan untuk menghentikan pipeline pada langkah tertentu sebelum deployment produksi atau membatalkan perubahan secara otomatis. Bagian integral dari hal ini adalah adopsi budaya integrasi berkelanjutan dan pengiriman/deployment berkelanjutan (CI/CD), di mana pelaksanaan commit atau perubahan kode melewati berbagai gerbang tahapan otomatis dari tahap pembuatan dan pengujian hingga tahap deployment pada lingkungan produksi.

Meskipun kebijaksanaan konvensional menyarankan Anda untuk melibatkan personel untuk prosedur operasional paling sulit, kami justru menyarankan Anda mengotomatiskan prosedur paling sulit karena alasan tersebut.

Langkah-langkah implementasi

Anda dapat mengotomatiskan deployment untuk menghapus operasi-operasi manual dengan mengikuti langkah-langkah berikut:

- Siapkan repositori kode untuk menyimpan kode Anda dengan aman: Gunakan sistem manajemen kode sumber yang di-hosting berdasarkan teknologi populer seperti Git untuk menyimpan kode sumber dan konfigurasi infrastruktur sebagai kode (IaC) Anda.
- Konfigurasikan layanan integrasi berkelanjutan untuk mengompilasi kode sumber, menjalankan pengujian, dan membuat artefak deployment: Untuk menyiapkan proyek pembangunan (build) untuk tujuan ini, lihat Memulai AWS CodeBuild menggunakan konsol.
- Siapkan layanan deployment yang mengotomatiskan deployment aplikasi dan menangani kompleksitas pembaruan aplikasi tanpa bergantung pada deployment manual yang rawan kesalahan: AWS CodeDeploy mengotomatiskan deployment perangkat lunak ke berbagai layanan

komputasi, seperti Amazon EC2, AWS Fargate, AWS Lambda, dan server on-premise Anda. Untuk mengonfigurasi langkah-langkah ini, silakan lihat Memulai CodeDeploy.

- Siapkan layanan pengiriman berkelanjutan yang mengotomatiskan pipeline rilis Anda untuk pembaruan aplikasi dan infrastruktur yang lebih cepat dan lebih andal: Pertimbangkan menggunakan AWS CodePipeline yang akan membantu Anda mengotomatiskan pipeline rilis. Untuk detail selengkapnya, lihat tutorial CodePipeline.

Sumber daya

Praktik-praktik terbaik terkait:

- OPS05-BP04 Menggunakan sistem manajemen build dan deployment
- OPS05-BP10 Mengotomatiskan integrasi dan deployment sepenuhnya
- OPS06-BP02 Menguji deployment
- OPS06-BP04 Mengotomatiskan pengujian dan rollback

Dokumen terkait:

- Pengiriman Berkelanjutan dari Tumpukan AWS CloudFormation Bersarang Menggunakan AWS CodePipeline
- Partner APN: partner yang dapat membantu Anda membuat solusi deployment yang diotomatisasi
- AWS Marketplace: produk yang dapat digunakan untuk mengotomatisasi deployment Anda
- Otomatiskan pesan obrolan dengan webhooks.
- Amazon Builders' Library: Memastikan keamanan rollback selama deployment
- Amazon Builders' Library: Melaju lebih cepat dengan pengiriman berkelanjutan
- Apa itu AWS CodePipeline?
- Apa itu CodeDeploy?
- AWS Systems Manager Patch Manager
- Apa itu Amazon SES?
- Apa itu Amazon Simple Notification Service?

Video terkait:

- AWS Summit 2019: CI/CD di AWS

Manajemen kegagalan

> Kegagalan tidak bisa dihindari dan semua hal akan mengalami kegagalan seiring berjalannya waktu: mulai dari router hingga diska keras, dari sistem operasi hingga unit memori yang membuat paket TCP mengalami kerusakan, dari kesalahan sementara hingga kegagalan permanen. Ini tidak bisa dihindari, meskipun Anda menggunakan perangkat keras berkualitas tinggi, apalagi komponen dengan biaya termurah - Werner Vogels, CTO - Amazon.com

Kegagalan komponen perangkat keras tingkat rendah menjadi hal yang harus dihadapi setiap hari di sebuah pusat data on-premise. Namun, saat di cloud, Anda harus terlindungi dari sebagian besar jenis kegagalan ini. Misalnya, volume Amazon EBS ditempatkan di Zona Ketersediaan tertentu dan direplikasi secara otomatis di dalamnya guna melindungi Anda dari kegagalan komponen tunggal. Semua volume EBS dirancang untuk ketersediaan 99,999%. Objek Amazon S3 disimpan di minimal tiga Zona Ketersediaan, menyediakan ketahanan objek sebesar 99,999999999% selama satu tahun. Terlepas dari penyedia cloud Anda, potensi kegagalan pasti ada dan bisa berdampak pada beban kerja Anda. Oleh karena itu, Anda harus mengambil langkah-langkah untuk mengimplementasikan ketangguhan jika Anda membutuhkan beban kerja yang bisa diandalkan.

Prasyarat untuk menerapkan praktik terbaik yang didiskusikan di sini adalah Anda harus memastikan bahwa orang-orang yang merancang desain, mengimplementasikan, dan mengoperasikan beban kerja Anda paham tentang tujuan bisnis dan tujuan keandalan untuk mencapai tujuan bisnis ini. Mereka harus paham dan dilatih untuk persyaratan keandalan ini.

Bagian ini menjelaskan praktik terbaik untuk mengelola kegagalan guna mencegah dampaknya pada beban kerja Anda.

Topik

Cadangkan data

Cadangkan data, aplikasi, dan konfigurasi untuk memenuhi persyaratan-persyaratan untuk sasaran waktu pemulihan (RTO) dan sasaran titik pemulihan (RPO).

Praktik terbaik

- REL09-BP01 Mengidentifikasi dan mencadangkan data yang perlu dicadangkan, atau melakukan reproduksi ulang data dari sumber
- REL09-BP02 Mengamankan dan mengenkripsikan cadangan
- REL09-BP03 Melakukan pencadangan data secara otomatis
- REL09-BP04 Melakukan pemulihan data secara berkala untuk memverifikasi integritas dan proses pencadangan

REL09-BP01 Mengidentifikasi dan mencadangkan data yang perlu dicadangkan, atau melakukan reproduksi ulang data dari sumber

Pahami dan gunakan kemampuan-kemampuan pencadangan sumber daya dan layanan data yang digunakan oleh beban kerja. Sebagian besar layanan menyediakan kemampuan untuk mencadangkan data beban kerja.

Hasil yang diinginkan: Sumber data telah diidentifikasi dan diklasifikasikan berdasarkan tingkat kekritisan. Kemudian, bangun strategi untuk pemulihan data berdasarkan RPO. Strategi ini melibatkan pencadangan sumber-sumber data, atau memiliki kemampuan untuk memproduksi ulang data dari sumber yang lain. Untuk kasus kehilangan data, strategi yang diimplementasikan akan memungkinkan pemulihan atau produksi ulang data dalam RPO dan RTO yang ditetapkan.

Fase kematangan cloud: Dasar

Anti-pola umum:

- Tidak mengetahui semua sumber data untuk beban kerja serta tingkat kekritisannya.
- Tidak melakukan pencadangan sumber data kritis.
- Melakukan pencadangan hanya beberapa sumber data tanpa menggunakan tingkat kekritisan sebagai kriteria.
- Tidak ada RPO yang ditetapkan, atau frekuensi pencadangan tidak memenuhi RPO.

- Tidak mengevaluasi apakah cadangan diperlukan atau apakah data dapat diproduksi ulang dari sumber yang lain.

Manfaat menerapkan praktik terbaik ini: Mengidentifikasi tempat-tempat yang memerlukan pencadangan dan mengimplementasikan mekanisme untuk membuat cadangan, atau mampu memproduksi ulang data dari sumber eksternal, semuanya dapat meningkatkan kemampuan untuk memulihkan dan mengembalikan data selama pemadaman.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Tinggi

Panduan implementasi

Semua penyimpanan data AWS menawarkan kemampuan pencadangan. Layanan-layanan seperti Amazon RDS dan Amazon DynamoDB memberikan dukungan tambahan pada pencadangan otomatis yang memungkinkan pemulihan titik waktu (PITR), yang akan memungkinkan Anda untuk memulihkan cadangan ke waktu kapan pun hingga lima menit atau kurang sebelum waktu saat ini. Banyak layanan AWS yang menawarkan kemampuan untuk menyalin cadangan ke Wilayah AWS yang lain. AWS Backup adalah sebuah alat yang akan memberi Anda kemampuan untuk melakukan sentralisasi dan otomatisasi terhadap perlindungan data di seluruh layanan AWS. AWS Elastic Disaster Recovery akan memungkinkan Anda untuk menyalin beban kerja server penuh dan mempertahankan perlindungan data berkelanjutan dari on-premise, lintas Zona Ketersediaan atau lintas Wilayah, dengan Sasaran Titik Pemulihan (RPO) yang diukur dalam hitungan detik.

Amazon S3 dapat digunakan sebagai tujuan pencadangan untuk sumber data yang dikelola mandiri dan yang dikelola oleh AWS. Layanan-layanan AWS seperti Amazon EBS, Amazon RDS, dan Amazon DynamoDB memiliki kemampuan bawaan untuk membuat cadangan. Perangkat lunak pencadangan pihak ketiga juga dapat digunakan.

Data on-premise dapat dicadangkan ke AWS Cloud dengan menggunakan AWS Storage Gateway atau AWS DataSync. Bucket Amazon S3 dapat digunakan untuk menyimpan data ini di AWS. Amazon S3 menawarkan beberapa tingkatan penyimpanan seperti Amazon Glacier atau Amazon Glacier Deep Archive untuk mengurangi biaya penyimpanan data.

Anda mungkin dapat memenuhi kebutuhan pemulihan data Anda dengan memproduksi ulang data dari sumber yang lain. Misalnya, simpul replika Amazon ElastiCache atau replika baca Amazon RDS dapat digunakan untuk memproduksi ulang data jika data primer hilang. Dalam kasus di mana sumber-sumber data seperti ini dapat digunakan untuk memenuhi Sasaran Titik Pemulihan (RPO) dan Sasaran Waktu Pemulihan (RTO), Anda mungkin tidak memerlukan cadangan. Contoh lainnya,

jika Anda menggunakan Amazon EMR, pencadangan penyimpanan data HDFS Anda mungkin tidak diperlukan, selama Anda dapat memproduksi ulang data ke Amazon EMR dari Amazon S3.

Ketika memilih strategi pencadangan, pertimbangkan waktu yang diperlukan untuk melakukan pemulihan data. Waktu yang diperlukan untuk melakukan pemulihan data tergantung pada tipe cadangan (untuk kasus strategi pencadangan), atau kompleksitas mekanisme produksi ulang data. Waktu ini termasuk dalam RTO untuk beban kerja.

Langkah-langkah implementasi

1. Mengidentifikasi semua sumber daya untuk beban kerja. Data dapat disimpan pada sejumlah sumber daya seperti basis data, volume, sistem file, sistem pencatatan log, dan penyimpanan objek. Lihat bagian Sumber Daya untuk menemukan Dokumen terkait mengenai berbagai layanan AWS tempat data disimpan, dan kemampuan cadangan yang disediakan oleh layanan-layanan ini.
2. Klasifikasikan sumber data berdasarkan tingkat kekritisan. Set data yang berbeda akan memiliki tingkat kekritisan yang berbeda untuk suatu beban kerja, sehingga memiliki persyaratan ketahanan yang berbeda pula. Misalnya, beberapa data mungkin kritis dan memerlukan RPO hampir nol, sedangkan data lain mungkin tidak terlalu kritis dan dapat mentoleransi RPO yang lebih tinggi dan beberapa kehilangan data. Demikian juga, set data yang berbeda mungkin memiliki persyaratan RTO yang berbeda.
3. Gunakan AWS atau layanan pihak ketiga untuk membuat cadangan data. AWS Backup adalah sebuah layanan terkelola yang memungkinkan pembuatan cadangan dari berbagai sumber data di AWS. AWS Elastic Disaster Recovery menangani replikasi data otomatis di bawah satu detik (sub-second) ke Wilayah AWS. Sebagian besar layanan AWS juga memiliki kemampuan native untuk membuat cadangan. AWS Marketplace juga memiliki banyak solusi untuk menyediakan kemampuan-kemampuan ini. Lihat Sumber Daya yang disebutkan di bawah ini untuk mendapatkan informasi tentang cara membuat cadangan data dari berbagai layanan AWS.
4. Untuk data yang tidak dicadangkan, bangun mekanisme produksi ulang data. Anda mungkin memilih untuk tidak mencadangkan data yang dapat diproduksi ulang dari sumber yang lain karena berbagai alasan. Mungkin terdapat situasi di mana produksi ulang data dari sumber yang lain saat diperlukan lebih murah daripada membuat cadangan, karena mungkin ada biaya-biaya yang timbul terkait penyimpanan cadangan. Contoh lainnya adalah ketika pemulihan dari cadangan memerlukan waktu lebih lama daripada produksi ulang data dari sumber-sumber lain, sehingga mengakibatkan pelanggaran RTO. Pada situasi-situasi demikian, pertimbangkan semua kompromi dan bangun sebuah proses yang ditetapkan dengan baik terkait bagaimana data dapat diproduksi ulang dari sumber-sumber ini saat pemulihan data diperlukan. Misalnya, jika Anda telah memuat data dari Amazon S3 ke gudang data (seperti Amazon Redshift), atau klaster MapReduce

(seperti Amazon EMR) untuk melakukan analisis pada data tersebut, ini mungkin adalah contoh data yang dapat diproduksi ulang dari sumber lain. Selama hasil dari semua analisis ini disimpan di suatu tempat atau dapat diproduksi ulang, Anda tidak akan mengalami kehilangan data akibat kegagalan pada gudang data atau klaster MapReduce. Contoh lain data yang dapat diproduksi ulang dari sumber lain adalah cache (seperti Amazon ElastiCache) atau replika baca RDS.

5. Buat jadwal pencadangan data. Membuat cadangan sumber data adalah proses berkala dan frekuensinya seharusnya tergantung pada RPO.

Tingkat upaya untuk Rencana Implementasi: Sedang

Sumber daya

Praktik-Praktik Terbaik Terkait:

REL13-BP01 Menetapkan sasaran pemulihan untuk waktu henti dan kehilangan data

REL13-BP02 Menggunakan strategi pemulihan untuk memenuhi sasaran pemulihan

Dokumen terkait:

- Apa itu AWS Backup?
- Apa itu AWS DataSync?
- Apa itu Volume Gateway?
- Partner APN: partner yang dapat membantu terkait pencadangan
- AWS Marketplace: produk yang dapat digunakan untuk pencadangan
- Snapshot Amazon EBS
- Mencadangkan Amazon EFS
- Mencadangkan Amazon FSx for Windows File Server
- Pencadangan dan pemulihan untuk ElastiCache for Redis
- Membuat Snapshot Klaster DB di Neptune
- Membuat Snapshot DB
- Membuat Aturan EventBridge yang Memicu Berdasarkan Jadwal
- Replikasi Lintas Wilayah dengan Amazon S3
- AWS Backup EFS ke EFS

- Mengekspor Data Log ke Amazon S3
- Manajemen siklus aktif objek
- Pencadangan Sesuai Permintaan dan Pemulihan untuk DynamoDB
- Pemulihan titik waktu untuk DynamoDB
- Menggunakan Snapshot Indeks Amazon OpenSearch Service
- Apa itu AWS Elastic Disaster Recovery?

Video terkait:

- AWS re:Invent 2021 - Pencadangan, pemulihan bencana, dan perlindungan ransomware dengan AWS
- Demo AWS Backup: Pencadangan Lintas Akun dan Lintas Wilayah
- AWS re:Invent 2019: Memahami lebih dalam mengenai AWS Backup, ft. Rackspace (STG341)

REL09-BP02 Mengamankan dan mengenkripsikan cadangan

Kontrol dan deteksi akses ke cadangan menggunakan autentikasi dan otorisasi. Gunakan enkripsi untuk mencegah dan mendeteksi jika integritas data cadangan terancam.

Implementasikan kontrol keamanan untuk mencegah akses tidak sah ke data cadangan. Enkripsi cadangan untuk melindungi kerahasiaan dan integritas data Anda.

Anti-pola umum:

- Buatlah akses yang sama ke cadangan dan otomatisasi pemulihan seperti yang dilakukan pada data.
- Tidak mengenkripsi cadangan.
- Tidak menerapkan imutabilitas untuk perlindungan terhadap penghapusan atau manipulasi.
- Menggunakan domain keamanan yang sama untuk sistem produksi dan pencadangan.
- Tidak memvalidasi integritas cadangan melalui pengujian reguler.

Manfaat menjalankan praktik terbaik ini:

- Mengamankan cadangan Anda akan mencegah manipulasi terhadap data, dan enkripsi data mencegah akses ke data tersebut jika tidak sengaja terekspos.

- Meningkatkan perlindungan terhadap ransomware dan ancaman siber lainnya yang menarget infrastruktur pencadangan.
- Mengurangi waktu pemulihan setelah insiden siber melalui proses pemulihan yang divalidasi.
- Meningkatkan kemampuan kontinuitas bisnis selama insiden keamanan.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Tinggi

Panduan implementasi

Kontrol dan deteksi akses ke cadangan dengan menggunakan autentikasi dan otorisasi seperti AWS Identity and Access Management (IAM). Gunakan enkripsi untuk mencegah dan mendeteksi jika integritas data cadangan terancam.

Amazon S3 mendukung beberapa metode enkripsi data diam. Dengan menggunakan enkripsi di sisi server, Amazon S3 dapat menerima objek sebagai data yang tidak terenkripsi dan mengenkripsinya saat disimpan. Dengan menggunakan enkripsi di sisi klien, aplikasi beban kerja bertanggung jawab untuk mengenkripsi data sebelum mengirimkannya ke Amazon S3. Kedua metode tersebut akan memungkinkan Anda menggunakan AWS Key Management Service (AWS KMS) guna menciptakan dan menyimpan kunci data. Anda dapat menyediakan kunci Anda sendiri dan bertanggung jawab atas kunci tersebut. Dengan menggunakan AWS KMS, Anda dapat menetapkan kebijakan menggunakan IAM terkait siapa yang dapat dan tidak dapat mengakses kunci data dan data terdekripsi.

Untuk Amazon RDS, cadangan juga akan dienkripsi jika Anda memilih untuk mengenkripsi basis data Anda. Pencadangan DynamoDB selalu dienkripsi. Ketika menggunakan AWS Elastic Disaster Recovery, semua data bergerak dan data diam dienkripsi. Dengan Pemulihan Bencana Elastis, data diam dapat dienkripsi dengan menggunakan Kunci Enkripsi Volume enkripsi Amazon EBS atau kunci kustom yang dikelola pelanggan.

Pertimbangan ketahanan siber

Guna meningkatkan keamanan cadangan terhadap ancaman siber, pertimbangkan untuk menerapkan kontrol tambahan berikut selain enkripsi:

- Implementasikan imutabilitas menggunakan Kunci Penyimpanan AWS Backup atau Kunci Objek Amazon S3 untuk mencegah data cadangan diubah atau dihapus selama periode retensi sehingga melindungi terhadap ransomware dan penghapusan berniat jahat.

- Tetapkan isolasi logis antara lingkungan produksi dan pencadangan menggunakan penyimpanan yang terisolasi secara logis dari AWS Backup untuk sistem krusial, sehingga menciptakan pemisahan yang membantu mencegah disusupinya kedua lingkungan tersebut secara bersamaan.
- Validasikan integritas cadangan secara teratur menggunakan uji pemulihan AWS Backup untuk memverifikasi bahwa cadangan tidak rusak dan dapat berhasil dipulihkan setelah insiden siber.
- Implementasikan persetujuan multipihak untuk operasi pemulihan krusial menggunakan persetujuan multipihak AWS Backup guna mencegah upaya pemulihan yang tidak sah atau berniat jahat dengan meminta otorisasi dari beberapa pemberi persetujuan yang ditunjuk.

Langkah-langkah implementasi

1. Gunakan enkripsi untuk setiap penyimpanan data. Jika sumber data terenkripsi, maka cadangannya juga akan terenkripsi.
   - Gunakan enkripsi di Amazon RDS. . Anda dapat mengonfigurasi enkripsi diam dengan menggunakan AWS Key Management Service saat membuat instans RDS.
   - Gunakan enkripsi pada volume Amazon EBS.. Anda dapat mengonfigurasi enkripsi default atau menentukan sebuah kunci unik saat pembuatan volume.
   - Gunakan enkripsi Amazon DynamoDB yang diperlukan. DynamoDB mengenkripsi semua data diam. Anda dapat menggunakan kunci AWS KMS yang dimiliki AWS atau kunci KMS yang dikelola AWS, yang menentukan sebuah kunci yang disimpan di akun Anda.
   - Lakukan enkripsi pada data Anda yang disimpan di Amazon EFS. Konfigurasikan enkripsi saat Anda membuat sistem file.
   - Konfigurasikan enkripsi di Wilayah sumber dan tujuan. Anda dapat mengonfigurasikan enkripsi saat diam di Amazon S3 dengan menggunakan kunci yang disimpan di KMS, tetapi kunci tersebut bersifat spesifik Wilayah. Anda dapat menentukan kunci tujuan saat Anda mengonfigurasikan replikasi.
   - Pilih apakah akan Anda akan menggunakan enkripsi Amazon EBS default atau eknripsi kustom untuk Pemulihan Bencana Elastis. Opsi ini akan mengenkripsi data diam yang direplikasi di disk Subnet Area Staging dan disk yang direplikasi.
2. Implementasikan izin hak akses paling rendah untuk mengakses cadangan Anda. Ikuti praktik-praktik terbaik untuk membatasi akses ke cadangan, snapshot, dan replika sesuai dengan praktik terbaik untuk keamanan.
3. Konfigurasikan imutabilitas untuk pencadangan penting. Untuk data krusial, terapkan Kunci Penyimpanan AWS Backup atau Kunci Objek S3 untuk mencegah penghapusan atau perubahan

selama periode retensi yang ditentukan. Untuk detail implementasi, lihat Kunci Penyimpanan AWS Backup.

4. Buat pemisahan logis untuk lingkungan pencadangan. Implementasikan penyimpanan yang terisolasi secara logis dari AWS Backup untuk sistem krusial yang membutuhkan perlindungan yang ditingkatkan terhadap ancaman siber. Untuk panduan implementasi, lihat Membangun ketahanan siber dengan penyimpanan yang terisolasi secara logis dari AWS Backup.
5. Implementasikan proses validasi cadangan. Konfigurasikan uji pemulihan AWS Backup untuk memverifikasi secara teratur bahwa cadangan tidak rusak dan dapat berhasil dipulihkan setelah insiden siber. Untuk informasi selengkapnya, lihat Validasikan kesiapan pemulihan dengan uji pemulihan AWS Backup.
6. Konfigurasikan persetujuan multipihak untuk operasi pemulihan sensitif. Untuk sistem krusial, terapkan persetujuan multipihak AWS Backup untuk meminta otorisasi dari beberapa pemberi persetujuan yang ditunjuk sebelum pemulihan dapat dilanjutkan. Untuk detail implementasi, lihat Tingkatkan ketahanan pemulihan dengan dukungan AWS Backup untuk persetujuan multipihak.

Sumber daya

Dokumen terkait:

- AWS Marketplace: produk yang dapat digunakan untuk pencadangan
- Enkripsi Amazon EBS
- Amazon S3: Melindungi Data Menggunakan Enkripsi
- Konfigurasi Tambahan CRR: Mereplika Objek yang Dibuat dengan Enkripsi di Sisi Server (SSE) Menggunakan Kunci Enkripsi yang disimpan di AWS KMS
- Enkripsi DynamoDB Saat Diam
- Mengenkripsi Sumber Daya Amazon RDS
- Mengenkripsi Data dan Metadata di Amazon EFS
- Enkripsi untuk Cadangan di AWS
- Mengelola Tabel yang Dienkripsi
- Pilar Keamanan - Kerangka Kerja AWS Well-Architected
- Apa itu AWS Elastic Disaster Recovery?
- FSISEC11: Bagaimana cara melindungi terhadap ransomware?
- Manajemen Risiko Ransomware di AWS Menggunakan Kerangka Kerja Keamanan Siber NIST
- Membangun ketahanan siber dengan penyimpanan yang terisolasi secara logis dari AWS Backup

- Validasikan kesiapan pemulihan dengan uji pemulihan AWS Backup
- Tingkatkan ketahanan pemulihan dengan dukungan AWS Backup untuk persetujuan multipihak

Contoh terkait:

- Mengimplementasikan Replikasi Lintas-Wilayah (CRR) Dua Arah untuk Amazon S3

REL09-BP03 Melakukan pencadangan data secara otomatis

Konfigurasikan pencadangan untuk dilakukan secara otomatis berdasarkan jadwal berkala yang diberikan oleh Sasaran Titik Pemulihan (RPO), atau berdasarkan perubahan dalam set data. Set data kritis dengan persyaratan data hilang yang rendah perlu dicadangkan otomatis secara rutin, sedangkan data yang tidak terlalu kritis yang apabila hilang masih dapat dimaklumi dapat dicadangkan tidak terlalu sering.

Hasil yang Diinginkan: Sebuah proses otomatis yang membuat cadangan sumber data dengan jadwal yang ditetapkan.

Anti-pola umum:

- Melakukan pencadangan secara manual.
- Menggunakan sumber daya yang memiliki kemampuan pencadangan, tetapi tidak termasuk pencadangan dalam otomatisasi Anda.

Manfaat menerapkan praktik terbaik ini: Melakukan otomatisasi pencadangan yang memverifikasi bahwa pencadangan dilakukan secara teratur berdasarkan RPO Anda dan memberi tahu Anda jika pencadangan tidak dilakukan.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Sedang

Panduan implementasi

AWS Backup dapat digunakan untuk membuat cadangan data secara otomatis dari berbagai sumber data AWS. Instans Amazon RDS dapat dicadangkan hampir secara berkelanjutan setiap lima menit dan objek Amazon S3 dapat dicadangkan hampir secara berkelanjutan setiap lima belas menit, dan memungkinkan pemulihan titik waktu (PITR) ke titik waktu tertentu di dalam riwayat pencadangan. Untuk sumber data AWS lainnya, seperti volume Amazon EBS, tabel Amazon DynamoDB, atau sistem file Amazon FSx, AWS Backup dapat menjalankan pencadangan secara otomatis setiap

satu jam. Layanan ini juga menawarkan kemampuan cadangan native. Layanan-layanan AWS yang menawarkan pencadangan otomatis dengan pemulihan titik waktu termasuk Amazon DynamoDB, Amazon RDS, dan Amazon Keyspaces (untuk Apache Cassandra) – ini dapat dikembalikan ke titik waktu tertentu dalam riwayat pencadangan. Sebagian besar layanan penyimpanan data AWS lainnya menawarkan kemampuan untuk menjadwalkan pencadangan secara berkala, dengan frekuensi setiap satu jam.

Amazon RDS dan Amazon DynamoDB menawarkan pencadangan berkelanjutan dengan pemulihan titik waktu. Penentuan versi Amazon S3, setelah dihidupkan, akan berjalan otomatis. Amazon Data Lifecycle Manager dapat digunakan untuk melakukan otomatisasi terhadap pembuatan, penyalinan dan penghapusan snapshot Amazon EBS. Layanan ini juga dapat mengotomatiskan pembuatan, penyalinan, penghentian, dan pembatalan registrasi Amazon Machine Image (AMI) yang dicadangkan Amazon EBS dan snapshot Amazon EBS yang melandasinya.

AWS Elastic Disaster Recovery memberikan replikasi tingkat blok yang berkelanjutan dari lingkungan sumber (on-premise atau AWS) ke wilayah pemulihan target. Snapshot titik waktu Amazon EBS dibuat dan dikelola secara otomatis oleh layanan tersebut.

Untuk tampilan otomatisasi dan riwayat pencadangan tersentralisasi, AWS Backup menyediakan solusi pencadangan berbasis kebijakan yang terkelola penuh. Layanan ini memusatkan dan mengotomatiskan pencadangan data di beberapa layanan AWS di cloud serta on-premise dengan menggunakan AWS Storage Gateway.

Selain penentuan versi, Amazon S3 dilengkapi dengan fitur replikasi. Seluruh bucket S3 dapat direplikasi secara otomatis ke bucket lain yang ada di Wilayah AWS yang sama atau berbeda.

Langkah-langkah implementasi

1. Identifikasi sumber data yang sedang dicadangkan secara manual. Untuk detail selengkapnya, lihat REL09-BP01 Mengidentifikasi dan mencadangkan data yang perlu dicadangkan, atau melakukan reproduksi ulang data dari sumber.
2. Tentukan RPO untuk beban kerja. Untuk detail selengkapnya, lihat REL13-BP01 Menetapkan sasaran pemulihan untuk waktu henti dan kehilangan data.
3. Gunakan solusi pencadangan otomatis atau layanan terkelola. AWS Backup adalah sebuah layanan terkelola sepenuhnya yang memudahkan untuk memusatkan dan mengotomatiskan perlindungan data di seluruh layanan AWS, di cloud, dan on-premise. Dengan menggunakan rencana cadangan di AWS Backup, buatlah aturan yang menetapkan sumber daya yang akan dicadangkan, dan frekuensi pembuatan cadangan ini. Frekuensi ini harus mengacu pada RPO

yang ditetapkan pada Langkah 2. Untuk panduan langsung tentang cara membuat cadangan otomatis dengan menggunakan AWS Backup, silakan lihat Menguji Cadangan dan Penyimpanan Kembali Data. Kemampuan pencadangan native ditawarkan oleh sebagian besar layanan AWS yang menyimpan data. Misalnya, RDS dapat dimanfaatkan untuk pencadangan secara otomatis dengan pemulihan titik waktu (PITR).

4. Untuk sumber data yang tidak didukung oleh solusi pencadangan otomatis atau layanan terkelola seperti sumber data on-premise atau antrean pesan, pertimbangkan untuk menggunakan solusi pihak ketiga tepercaya untuk membuat cadangan secara otomatis. Pilihan lainnya, Anda dapat membuat otomatisasi untuk melakukannya dengan menggunakan AWS CLI atau SDK. Anda dapat menggunakan Fungsi AWS Lambda atau AWS Step Functions untuk menetapkan logika yang terlibat dalam pembuatan cadangan data, dan gunakan Amazon EventBridge untuk menginvokasinya dengan frekuensi yang didasarkan pada RPO Anda.

Tingkat upaya untuk Rencana Implementasi: Rendah

Sumber daya

Dokumen terkait:

- Partner APN: partner yang dapat membantu terkait pencadangan
- AWS Marketplace: produk yang dapat digunakan untuk pencadangan
- Membuat Aturan EventBridge yang Memicu Berdasarkan Jadwal
- Apa itu AWS Backup?
- Apa itu AWS Step Functions?
- Apa itu AWS Elastic Disaster Recovery?

Video terkait:

- AWS re:Invent 2019: Memahami lebih dalam mengenai AWS Backup, ft. Rackspace (STG341)

REL09-BP04 Melakukan pemulihan data secara berkala untuk memverifikasi integritas dan proses pencadangan

Validasikan bahwa implementasi proses pencadangan Anda memenuhi Sasaran Waktu Pemulihan (RTO) dan Sasaran Titik Pemulihan (RPO) dengan melakukan uji pemulihan.

Hasil yang Diinginkan: Data dari cadangan dipulihkan secara berkala dengan menggunakan mekanisme yang ditentukan dengan baik untuk memverifikasi bahwa pemulihan tersebut dapat dilakukan dalam sasaran waktu pemulihan (RTO) yang ditetapkan untuk beban kerja. Pastikan bahwa pemulihan dari pencadangan menghasilkan sumber daya yang berisi data asli tanpa ada data yang rusak atau tidak dapat diakses, serta dengan kehilangan data dalam sasaran titik pemulihan (RPO).

Anti-pola umum:

- Memulihkan cadangan, tetapi tidak mengambil data atau membuat kueri data apa pun untuk memastikan bahwa data hasil pemulihan dapat digunakan.
- Dengan anggapan bahwa cadangan sudah ada.
- Dengan anggapan bahwa cadangan sistem dapat dioperasikan sepenuhnya dan data dapat dipulihkan dari sistem.
- Dengan anggapan bahwa waktu untuk memulihkan data dari cadangan termasuk dalam RTO untuk beban kerja.
- Dengan anggapan bahwa data dalam cadangan termasuk dalam RPO untuk beban kerja
- Melakukan pemulihan apabila diperlukan, tanpa menggunakan runbook, atau di luar prosedur otomatis yang ditetapkan.

Manfaat menerapkan praktik terbaik ini: Pengujian pemulihan cadangan akan memverifikasi bahwa data dapat dipulihkan saat dibutuhkan tanpa perlu khawatir data akan hilang atau rusak, bahwa pemulihan dapat dilakukan dalam RTO untuk beban kerja, dan kehilangan data apa pun masih termasuk dalam RPO untuk beban kerja.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Sedang

Panduan implementasi

Pengujian kemampuan pencadangan dan pemulihan akan meningkatkan keyakinan Anda pada kemampuan untuk menjalankan tindakan ini selama terjadi pemadaman (outage). Pulihkan cadangan ke lokasi baru secara berkala dan lakukan pengujian untuk memastikan integritas data. Beberapa pengujian umum yang harus dilakukan adalah memeriksa apakah semua data tersedia, tidak mengalami kerusakan, dapat diakses, dan setiap kehilangan data masih termasuk dalam RPO untuk beban kerja. Pengujian tersebut dapat juga membantu memastikan apakah mekanisme pemulihan cukup cepat untuk mengakomodasi RTO beban kerja.

Dengan menggunakan AWS, Anda dapat mempertahankan lingkungan pengujian dan memulihkan cadangan Anda untuk menilai kemampuan RTO dan RPO, serta menjalankan pengujian pada konten dan integritas data.

Selain itu, Amazon RDS dan Amazon DynamoDB dapat memungkinkan pemulihan titik waktu (PITR). Dengan menggunakan pencadangan yang berkelanjutan, Anda dapat memulihkan set data ke statusnya sesuai dengan waktu dan tanggal yang ditentukan.

Jika semua data tersedia, tidak mengalami kerusakan, dapat diakses, dan kehilangan data apa pun masih termasuk dalam RPO untuk beban kerja. Pengujian tersebut dapat juga membantu memastikan apakah mekanisme pemulihan cukup cepat untuk mengakomodasi RTO beban kerja.

AWS Elastic Disaster Recovery menawarkan snapshot pemulihan titik waktu volume Amazon EBS secara berkelanjutan. Saat server sumber direplikasi, status point-in-time dicatat dari waktu ke waktu berdasarkan kebijakan yang dikonfigurasi. Pemulihan Bencana Elastis dapat membantu Anda untuk memverifikasi integritas snapshot ini dengan meluncurkan instans untuk tujuan pengujian dan latihan tanpa mengarahkan lalu lintas.

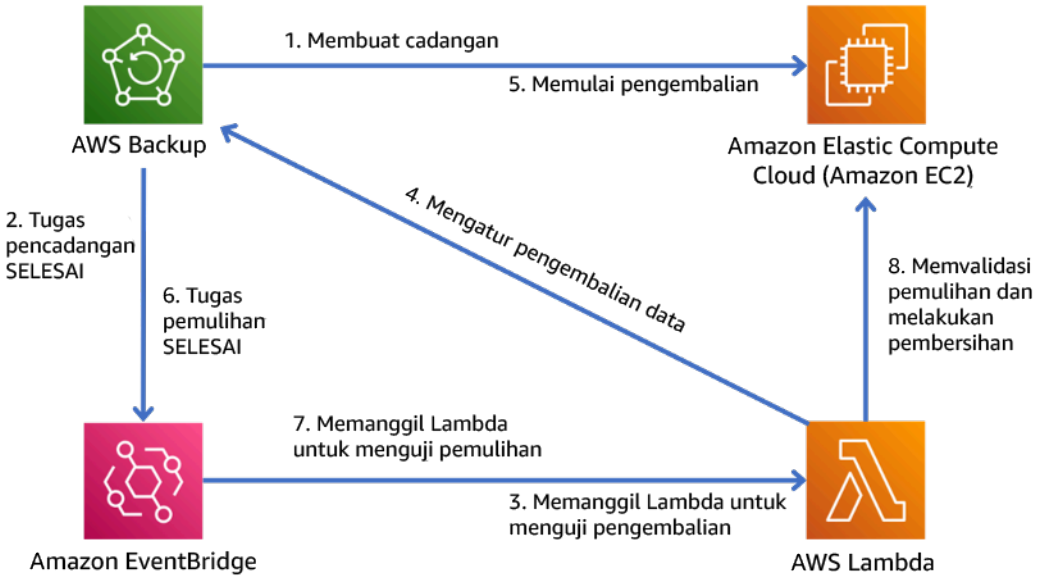
Langkah-langkah implementasi

1. Identifikasi sumber data yang dicadangkan saat ini dan lokasi penyimpanan cadangan tersebut. Untuk panduan implementasi, lihat REL09-BP01 Mengidentifikasi dan mencadangkan data yang perlu dicadangkan, atau melakukan reproduksi ulang data dari sumber.
2. Menetapkan kriteria untuk validasi data untuk masing-masing sumber data. Jenis data yang berbeda akan memiliki properti data yang berbeda, yang dapat memerlukan mekanisme validasi yang berbeda. Pertimbangkan bagaimana data ini dapat divalidasi sebelum Anda yakin untuk menggunakannya dalam produksi. Beberapa cara umum untuk memvalidasi data adalah dengan menggunakan data dan properti pencadangan seperti jenis data, format, checksum, ukuran, atau gabungan darinya dengan logika validasi kustom. Misalnya, hal ini dapat dilakukan dengan perbandingan nilai checksum antara sumber daya yang dipulihkan dan sumber data pada waktu cadangan dibuat.
3. Menetapkan RTO dan RPO untuk memulihkan data berdasarkan tingkat kekritisan data. Untuk panduan implementasi, lihat REL13-BP01 Menetapkan sasaran pemulihan untuk waktu henti dan kehilangan data.
4. Menilai kemampuan pemulihan Anda. Tinjau strategi pencadangan dan pemulihan untuk memahami apakah hal tersebut memenuhi RTO dan RPO, serta sesuaikan strategi sesuai keperluan. Dengan menggunakan AWS Resilience Hub, Anda dapat menjalankan penilaian

terhadap beban kerja Anda. Penilaian tersebut mengevaluasi konfigurasi aplikasi terhadap kebijakan dan pelaporan ketahanan jika target RTO dan RPO dapat dipenuhi.

5. Lakukan penyimpanan kembali pengujian dengan menggunakan proses yang ditetapkan saat ini yang digunakan dalam produksi untuk pemulihan data. Proses ini bergantung pada cara sumber data asli dicadangkan, format dan lokasi penyimpanan cadangan tersebut, atau apakah data diproduksi ulang dari sumber-sumber lainnya. Misalnya, jika Anda menggunakan sebuah layanan terkelola seperti AWS Backup, hal ini mungkin sesederhana memulihkan cadangan ke sumber daya baru. Jika Anda menggunakan AWS Elastic Disaster Recovery, maka Anda dapat meluncurkan latihan pemulihan.
6. Validasi pemulihan data dari sumber daya yang dipulihkan berdasarkan kriteria validasi data yang sebelumnya Anda buat. Apakah data yang direstorasi dan dipulihkan memiliki sebagian besar catatan atau item terbaru pada waktu pencadangan? Apakah data ini masih termasuk dalam RPO untuk beban kerja?
7. Ukur waktu yang diperlukan untuk menyimpan kembali dan melakukan pemulihan dan kemudian bandingkan dengan RTO Anda yang sudah ditetapkan. Apakah data ini masih termasuk dalam RTO untuk beban kerja? Misalnya, bandingkan stempel waktu dari kapan proses restorasi dimulai dan kapan validasi pemulihan selesai untuk menghitung waktu yang diperlukan proses ini. Semua panggilan API AWS diberi stempel waktu dan informasi ini tersedia di AWS CloudTrail. Meskipun informasi ini dapat menyediakan detail waktu kapan proses pemulihan dimulai, namun stempel waktu akhir yang menunjukkan kapan validasi diselesaikan harus dicatat berdasarkan logika validasi Anda. Jika Anda menggunakan proses otomatis, maka layanan-layanan seperti Amazon DynamoDB dapat digunakan untuk menyimpan informasi ini. Selain itu, banyak layanan AWS yang menyediakan riwayat peristiwa yang berisi informasi dengan yang dilengkapi stempel waktu tentang kapan tindakan-tindakan diambil. Dalam AWS Backup, tindakan pembuatan cadangan dan penyimpanan kembali disebut sebagai tugas, dan tugas tersebut berisi informasi stempel waktu sebagai bagian dari metadata yang dapat digunakan untuk mengukur waktu yang diperlukan untuk pemulihan.
8. Berikan notifikasi untuk pemangku kepentingan jika validasi data gagal, atau jika waktu yang diperlukan untuk pemulihan melebihi RTO yang ditetapkan untuk beban kerja. Saat menerapkan otomatisasi untuk melakukan langkah ini, seperti di lab ini, layanan-layanan seperti Amazon Simple Notification Service (Amazon SNS) dapat digunakan untuk mengirim pemberitahuan push seperti email atau SMS kepada para pemangku kepentingan. Pesan-pesan ini juga dapat dipublikasikan ke aplikasi perpesanan seperti Amazon Chime, Slack, atau Microsoft Teams atau digunakan untuk membuat tugas sebagai OpsItems dengan menggunakan AWS Systems Manager OpsCenter.

9. Otomatiskan proses ini untuk menjalankannya secara berkala. Sebagai contoh, layanan-layanan seperti AWS Lambda atau State Machine di AWS Step Functions dapat digunakan untuk melakukan otomatisasi proses pemulihan, dan Amazon EventBridge dapat digunakan untuk menginvokasi alur kerja otomatisasi ini secara berkala seperti yang ditampilkan dalam diagram arsitektur di bawah ini. Pelajari cara Mengotomatiskan validasi pemulihan data dengan AWS Backup. Selain itu, lab Well-Architected ini memberikan pengalaman langsung tentang satu cara untuk melakukan otomatisasi untuk beberapa langkah yang diuraikan di sini.



Gambar 9. Proses pencadangan dan pemulihan otomatis

Tingkat upaya untuk Rencana Implementasi: Sedang hingga tinggi tergantung pada kompleksitas kriteria validasinya.

Sumber daya

Dokumen terkait:

- Mengotomatiskan validasi pemulihan data dengan AWS Backup
- Partner APN: partner yang dapat membantu terkait pencadangan
- AWS Marketplace: produk yang dapat digunakan untuk pencadangan
- Membuat Aturan EventBridge yang Memicu Berdasarkan Jadwal
- Pencadangan sesuai permintaan dan pemulihan untuk DynamoDB
- Apa itu AWS Backup?

- Apa itu AWS Step Functions?
- Apa itu AWS Elastic Disaster Recovery
- AWS Elastic Disaster Recovery

Gunakan isolasi kesalahan untuk melindungi beban kerja Anda

Isolasi kesalahan membatasi dampak kegagalan komponen atau sistem ke batas yang ditentukan. Dengan isolasi yang baik, komponen-komponen yang ada di luar batas ini tidak terpengaruh oleh kegagalan. Menjalankan beban kerja Anda di beberapa batas isolasi kesalahan dapat membuatnya lebih tahan terhadap kegagalan.

Praktik terbaik

- REL10-BP01 Melakukan deployment beban kerja ke beberapa lokasi
- REL10-BP02 Mengotomatiskan pemulihan untuk komponen yang dibatasi dalam satu lokasi
- REL10-BP03 Menggunakan arsitektur bulkhead untuk membatasi cakupan dampak

REL10-BP01 Melakukan deployment beban kerja ke beberapa lokasi

Distribusikan sumber daya dan data beban kerja ke beberapa Zona Ketersediaan atau, jika diperlukan, ke beberapa Wilayah AWS.

Prinsip dasar untuk desain layanan di AWS adalah untuk menghindari titik kegagalan tunggal, termasuk infrastruktur fisik yang mendasarinya. AWS menyediakan sumber daya dan layanan komputasi cloud secara global di berbagai lokasi geografis yang disebut Wilayah. Setiap Wilayah berdiri sendiri secara fisik dan logis, dan terdiri dari tiga Zona Ketersediaan (AZ) atau lebih. Zona Ketersediaan terletak berdekatan secara geografis, tetapi terpisah dan terisolasi secara fisik. Ketika Anda mendistribusikan beban kerja Anda di beberapa Zona Ketersediaan dan Wilayah, Anda mengurangi risiko ancaman seperti kebakaran, banjir, bencana terkait cuaca, gempa bumi, dan kesalahan manusia.

Buat strategi lokasi untuk memastikan ketersediaan tinggi yang sesuai dengan beban kerja Anda.

Hasil yang diinginkan: Beban kerja produksi didistribusikan di beberapa Zona Ketersediaan (AZ) atau Wilayah untuk mencapai toleransi kesalahan dan ketersediaan tinggi.

Anti-pola umum:

- Anda hanya menempatkan beban kerja produksi di satu Zona Ketersediaan.
- Anda mengimplementasikan arsitektur multi-Wilayah ketika arsitektur multi-AZ sudah dapat memenuhi persyaratan bisnis.
- Deployment atau data Anda tidak lagi sinkron, sehingga terjadi penyimpangan konfigurasi atau data yang tidak direplikasi secara memadai.
- Anda tidak memperhitungkan dependensi antar-komponen aplikasi jika ketahanan dan persyaratan multi-lokasi berbeda di antara komponen-komponen tersebut.

Manfaat menjalankan praktik terbaik ini:

- Beban kerja Anda lebih tahan terhadap insiden, seperti pemadaman listrik atau kegagalan kontrol lingkungan, bencana alam, kegagalan layanan hulu, atau masalah jaringan yang berdampak pada AZ atau seluruh Wilayah.
- Anda dapat mengakses inventaris instans Amazon EC2 yang lebih luas dan mengurangi kemungkinan InsufficientCapacityExceptions (ICE) saat meluncurkan jenis instans EC2 tertentu.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Tinggi

Panduan implementasi

Lakukan deployment dan operasikan semua beban kerja produksi di setidaknya dua Zona Ketersediaan (AZ) di sebuah Wilayah.

Menggunakan beberapa Zona Ketersediaan

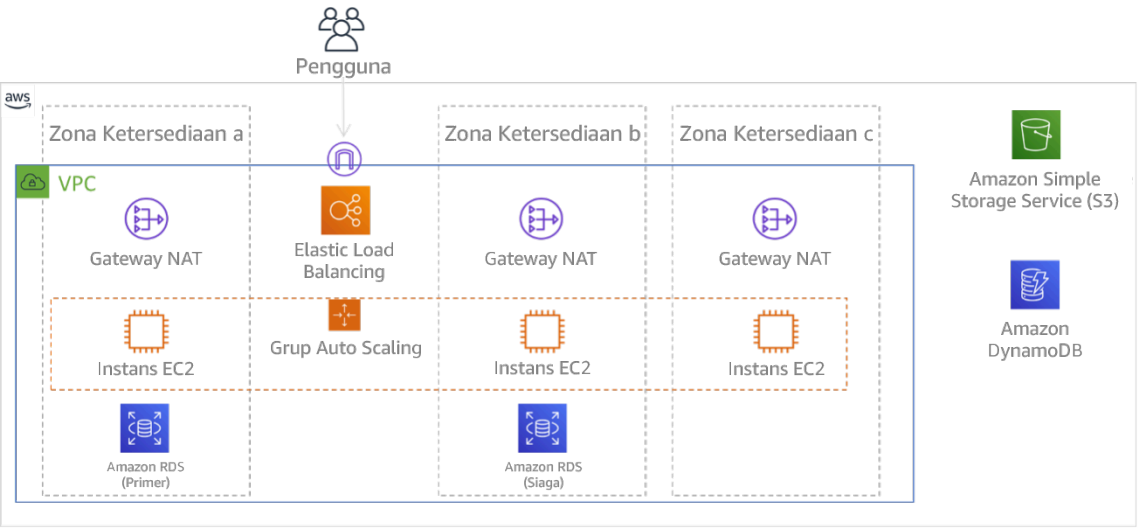
Zona Ketersediaan adalah lokasi hosting sumber daya yang secara fisik terpisah satu sama lain untuk menghindari kegagalan yang berkorelasi akibat risiko seperti kebakaran, banjir, dan tornado. Setiap Zona Ketersediaan memiliki infrastruktur fisik independen, termasuk sambungan listrik lokal, sumber listrik cadangan, sistem mekanis, dan koneksi jaringan. Desain infrastruktur yang independen ini membatasi dampak kegagalan di salah satu komponen ini hanya pada Zona Ketersediaan yang terkena dampak. Misalnya, jika insiden di seluruh AZ membuat instans EC2 tidak tersedia di Zona Ketersediaan yang terpengaruh, instans Anda di Zona Ketersediaan lainnya tetap tersedia.

Meskipun terpisah secara fisik, Zona Ketersediaan di Wilayah AWS yang sama cukup dekat untuk menyediakan jaringan throughput tinggi dengan latensi rendah (milidetik satu digit). Anda dapat mereplikasi data secara sinkron antara Zona Ketersediaan untuk sebagian besar beban kerja tanpa memengaruhi pengalaman pengguna secara signifikan. Artinya Anda dapat menggunakan Zona Ketersediaan di sebuah Wilayah dalam konfigurasi aktif/aktif atau aktif/siaga.

Semua komputasi yang terkait dengan beban kerja Anda harus didistribusikan di antara beberapa Zona Ketersediaan. Hal ini termasuk instans Amazon EC2, tugas AWS Fargate, dan fungsi AWS Lambda yang terhubung dengan VPC. Layanan komputasi AWS, termasuk EC2 Auto Scaling, Amazon Elastic Container Service (ECS), dan Amazon Elastic Kubernetes Service (EKS), menyediakan cara bagi Anda untuk meluncurkan dan mengelola komputasi di beberapa Zona Ketersediaan. Konfigurasikan hal-hal tersebut untuk secara otomatis mengganti komputasi sesuai kebutuhan di Zona Ketersediaan yang berbeda untuk menjaga ketersediaan. Untuk mengarahkan lalu lintas ke Zona Ketersediaan yang tersedia, tempatkan penyeimbang beban di depan komputasi Anda, seperti Penyeimbang Beban Aplikasi atau Penyeimbang Beban Jaringan. Penyeimbang beban AWS dapat mengalihkan lalu lintas ke instans yang tersedia jika terjadi gangguan pada suatu Zona Ketersediaan.

Anda juga harus mereplikasi data untuk beban kerja Anda dan membuatnya tersedia di beberapa Zona Ketersediaan. Beberapa layanan data yang dikelola AWS, seperti Amazon S3, Amazon Elastic File Service (EFS), Amazon Aurora, Amazon DynamoDB, Amazon Simple Queue Service (SQS), dan Amazon Kinesis Data Streams mereplikasi data di beberapa Zona Ketersediaan secara default dan tangguh terhadap gangguan Zona Ketersediaan. Dengan layanan data lain yang dikelola AWS, seperti Amazon Relational Database Service (RDS), Amazon Redshift, dan Amazon ElastiCache, Anda harus mengaktifkan replikasi multi-AZ. Setelah diaktifkan, layanan-layanan ini secara otomatis mendeteksi gangguan pada Zona Ketersediaan, mengalihkan permintaan ke Zona Ketersediaan yang tersedia, dan mereplikasi ulang data sesuai kebutuhan setelah pemulihan tanpa campur tangan pelanggan. Pelajari panduan pengguna untuk setiap layanan data yang dikelola AWS yang Anda gunakan untuk memahami kemampuan, perilaku, dan operasi multi-AZ pada layanan tersebut.

Jika Anda menggunakan penyimpanan yang dikelola sendiri, seperti volume Amazon Elastic Block Store (EBS) atau penyimpanan instans Amazon EC2, Anda harus mengelola sendiri replikasi multi-AZ.



Gambar 9: Arsitektur multitingkat di-deploy di tiga Zona Ketersediaan. Perhatikan bahwa Amazon S3 dan Amazon DynamoDB selalu Multi-AZ secara otomatis. ELB juga di-deploy ke tiga zona.

Menggunakan beberapa Wilayah AWS

Jika Anda memiliki beban kerja yang memerlukan ketahanan ekstrem (seperti infrastruktur sangat penting, aplikasi terkait kesehatan, atau layanan dengan persyaratan ketersediaan yang ketat dari pelanggan atau berdasarkan peraturan), Anda mungkin memerlukan ketersediaan tambahan di luar apa yang dapat disediakan oleh sebuah Wilayah AWS. Dalam hal ini, Anda harus melakukan deployment dan mengoperasikan beban kerja Anda di setidaknya dua Wilayah AWS (dengan asumsi bahwa persyaratan residensi data Anda memungkinkan hal ini).

Wilayah AWS terletak di wilayah geografis yang berbeda-beda di seluruh dunia dan di berbagai benua. Wilayah AWS memiliki pemisahan dan isolasi fisik yang lebih besar daripada Zona Ketersediaan. Layanan AWS, dengan beberapa pengecualian, memanfaatkan desain ini untuk beroperasi sepenuhnya secara independen di antara Wilayah yang berbeda-beda (juga dikenal sebagai layanan Regional). Kegagalan suatu layanan Wilayah AWS dirancang agar tidak memengaruhi layanan di Wilayah yang berbeda.

Ketika Anda mengoperasikan beban kerja Anda di beberapa Wilayah, Anda harus mempertimbangkan persyaratan tambahan. Karena sumber daya di Wilayah yang berbeda-beda saling terpisah dan independen satu sama lain, Anda harus menduplikasi komponen beban kerja Anda di setiap Wilayah. Hal ini termasuk infrastruktur dasar, seperti VPC, selain komputasi dan layanan data.

CATATAN: Saat Anda mempertimbangkan desain multi-Regional, pastikan bahwa beban kerja Anda mampu berjalan di satu Wilayah. Jika Anda membuat dependensi antar-Wilayah yang membuat

komponen di satu Wilayah bergantung pada layanan atau komponen di Wilayah lain, Anda dapat meningkatkan risiko kegagalan dan melemahkan postur keandalan Anda secara signifikan.

Untuk memudahkan deployment multi-Regional dan menjaga konsistensi, AWS CloudFormation StackSets dapat mereplikasi seluruh infrastruktur AWS Anda di beberapa Wilayah. AWS CloudFormation juga dapat mendeteksi penyimpangan konfigurasi dan memberi tahu Anda ketika sumber daya AWS Anda di sebuah Wilayah tidak sinkron. Banyak layanan AWS menawarkan replikasi multi-wilayah untuk aset beban kerja penting. Misalnya, EC2 Image Builder dapat menerbitkan image mesin EC2 (AMI) Anda setelah setiap kali proses build selesai ke setiap Wilayah yang Anda gunakan. Amazon Elastic Container Registry (ECR) dapat mereplikasi image kontainer Anda ke Wilayah yang Anda pilih.

Anda juga harus mereplikasi data Anda di setiap Wilayah yang Anda pilih. Banyak layanan data yang dikelola AWS menyediakan kemampuan replikasi lintas Wilayah, termasuk Amazon S3, Amazon DynamoDB, Amazon RDS, Amazon Aurora, Amazon Redshift, Amazon ElastiCache, dan Amazon EFS. Tabel global Amazon DynamoDB menerima penulisan di Wilayah mana pun yang didukung dan akan mereplikasi data di antara semua Wilayah lain yang Anda konfigurasikan. Dengan layanan lain, Anda harus menetapkan Wilayah utama untuk penulisan karena Wilayah lain berisi replika hanya-baca. Untuk setiap layanan data terkelola AWS yang digunakan beban kerja Anda, lihat panduan pengguna dan panduan developernya untuk memahami kemampuan dan batasan multi-Wilayahnya. Perhatikan secara khusus ke mana penulisan harus diarahkan, kemampuan dan batasan transaksi, bagaimana replikasi dilakukan, dan bagaimana memantau sinkronisasi antar-Wilayah.

AWS juga menyediakan kemampuan untuk merutekan lalu lintas permintaan ke deployment Regional Anda dengan fleksibilitas tinggi. Misalnya, Anda dapat mengonfigurasi catatan DNS Anda menggunakan Amazon Route 53 untuk mengarahkan lalu lintas ke Wilayah terdekat yang tersedia bagi pengguna. Selain itu, Anda dapat mengonfigurasi catatan DNS Anda dalam konfigurasi aktif/siaga, yang menetapkan satu Wilayah sebagai primer dan beralih ke replika Regional hanya jika Wilayah primer mengalami masalah. Anda dapat mengonfigurasi pemeriksaan kondisi Route 53 untuk mendeteksi titik akhir yang bermasalah dan melakukan failover otomatis. Selain itu Anda juga dapat menggunakan Amazon Application Recovery Controller (ARC) untuk menyediakan kontrol perutean dengan ketersediaan tinggi untuk merutekan ulang lalu lintas secara manual sesuai kebutuhan.

Meskipun Anda memilih untuk tidak beroperasi di beberapa Wilayah untuk ketersediaan tinggi, pertimbangkanlah untuk menggunakan beberapa Wilayah sebagai bagian dari strategi pemulihan bencana (DR) Anda. Jika memungkinkan, replikasi data dan komponen infrastruktur beban kerja Anda dalam konfigurasi warm standby atau pilot light di Wilayah sekunder. Dalam desain ini, Anda

mereplikasi infrastruktur dasar dari Wilayah primer seperti VPC, grup Auto Scaling, orkestrator kontainer, dan komponen lainnya, tetapi Anda mengonfigurasi komponen berukuran variabel di Wilayah siaga (seperti jumlah instans EC2 dan replika basis data) menjadi ukuran minimal yang dapat beroperasi. Anda juga mengatur replikasi data berkelanjutan dari Wilayah primer ke Wilayah siaga. Jika suatu insiden terjadi, Anda kemudian dapat menskalakan, atau menambah, sumber daya di Wilayah siaga, lalu mengubahnya menjadi Wilayah primer.

Langkah-langkah implementasi

1. Bekerjasamalah dengan pemangku kepentingan bisnis dan ahli residensi data untuk menentukan Wilayah AWS mana yang dapat digunakan untuk meng-host sumber daya dan data Anda.
2. Libatkan para pemangku kepentingan bisnis dan teknis untuk mengevaluasi beban kerja Anda, dan tentukan apakah kebutuhan ketahanannya dapat dipenuhi dengan pendekatan multi-AZ (Wilayah AWS tunggal) atau jika kebutuhannya adalah pendekatan multi-Wilayah (jika penggunaan beberapa Wilayah diizinkan). Penggunaan beberapa Wilayah dapat mencapai ketersediaan yang lebih besar, tetapi juga dapat meningkatkan kompleksitas dan biaya. Pertimbangkan faktor-faktor berikut dalam evaluasi Anda:
    a. Tujuan bisnis dan persyaratan pelanggan: Berapa lama waktu henti yang diizinkan jika insiden yang berdampak pada beban kerja terjadi di sebuah Zona Ketersediaan atau Wilayah? Evaluasi tujuan titik pemulihan Anda seperti yang dibahas dalam REL13-BP01 Menetapkan sasaran pemulihan untuk waktu henti dan kehilangan data.
    b. Persyaratan pemulihan bencana (DR): Potensi bencana seperti apa yang ingin Anda antisipasi? Pertimbangkan kemungkinan kehilangan data atau ketidaktersediaan dalam jangka panjang pada berbagai cakupan dampak, mulai dari satu Zona Ketersediaan hingga keseluruhan Wilayah. Jika Anda mereplikasi data dan sumber daya di beberapa Zona Ketersediaan, dan satu Zona Ketersediaan mengalami gangguan berkepanjangan, Anda dapat memulihkan layanan di Zona Ketersediaan lainnya. Jika Anda mereplikasi data dan sumber daya di beberapa Wilayah, Anda dapat memulihkan layanan di Wilayah lain.
3. Lakukan deployment sumber daya komputasi Anda ke beberapa Zona Ketersediaan.
    a. Di VPC Anda, buat beberapa subnet di Zona Ketersediaan yang berbeda-beda. Konfigurasikan setiap subnet agar cukup besar untuk mengakomodasi sumber daya yang dibutuhkan untuk melayani beban kerja, bahkan saat terjadi gangguan. Untuk detail selengkapnya, lihat REL02-BP03 Pastikan alokasi subnet IP memperhitungkan ekspansi dan ketersediaan.
    b. Jika Anda menggunakan instans Amazon EC2, gunakan EC2 Auto Scaling untuk mengelola instans Anda. Tentukan subnet yang Anda pilih pada langkah sebelumnya saat Anda membuat grup Auto Scaling.

   c. Jika Anda menggunakan komputasi AWS Fargate untuk Amazon ECS atau Amazon EKS, pilih subnet yang Anda pilih pada langkah pertama saat Anda membuat Layanan ECS, luncurkan tugas ECS, atau buat profil Fargate untuk EKS.

   d. Jika Anda menggunakan fungsi AWS Lambda yang perlu dijalankan di VPC Anda, pilih subnet yang Anda pilih pada langkah pertama saat Anda membuat fungsi Lambda. Untuk fungsi apa pun yang tidak memiliki konfigurasi VPC, AWS Lambda mengelola ketersediaan untuk Anda secara otomatis.

   e. Tempatkan pengarah lalu lintas seperti penyeimbang beban di depan sumber daya komputasi Anda. Jika penyeimbangan beban lintas zona diaktifkan, Penyeimbang Beban Aplikasi AWS dan Penyeimbang Beban Jaringan akan mendeteksi saat target seperti instans dan kontainer EC2 tidak dapat dijangkau karena gangguan Zona Ketersediaan lalu mengalihkan lalu lintas ke target di Zona Ketersediaan yang sehat. Jika Anda menonaktifkan penyeimbangan beban lintas zona, gunakan Amazon Application Recovery Controller (ARC) untuk menyediakan kemampuan peralihan zona. Jika Anda menggunakan penyeimbang beban pihak ketiga atau telah menerapkan penyeimbang beban Anda sendiri, konfigurasikan dengan beberapa frontend di Zona Ketersediaan yang berbeda-beda.

4. Replikasi data beban kerja Anda di beberapa Zona Ketersediaan.

   a. Jika Anda menggunakan layanan data yang dikelola AWS seperti Amazon RDS, Amazon ElastiCache, atau Amazon FSx, pelajari panduan penggunanya untuk memahami kemampuan replikasi data dan ketahanannya. Aktifkan replikasi dan failover lintas-AZ jika perlu.

   b. Jika Anda menggunakan layanan penyimpanan yang dikelola AWS seperti Amazon S3, Amazon EFS, dan Amazon FSx, jangan menggunakan konfigurasi satu AZ atau Satu Zona untuk data yang memerlukan daya tahan tinggi. Gunakan konfigurasi multi-AZ untuk layanan-layanan ini. Periksa panduan pengguna layanan masing-masing untuk menentukan apakah replikasi multi-AZ diaktifkan secara default atau apakah Anda harus mengaktifkannya.

   c. Jika Anda menjalankan basis data, antrean, atau layanan penyimpanan lainnya yang Anda kelola sendiri, atur replikasi multi-AZ sesuai dengan petunjuk atau praktik terbaik aplikasi. Pahami prosedur failover untuk aplikasi Anda.

5. Konfigurasikan layanan DNS Anda untuk mendeteksi gangguan pada AZ dan mengalihkan lalu lintas ke Zona Ketersediaan yang sehat. Amazon Route 53, apabila penggunaannya dikombinasikan dengan Penyeimbang Beban Elastis, dapat melakukan hal ini secara otomatis. Route 53 juga dapat dikonfigurasikan dengan catatan failover yang menggunakan pemeriksaan kondisi untuk merespons kueri dengan alamat IP yang sehat saja. Untuk data DNS apa pun yang digunakan untuk failover, tentukan nilai time to live (TTL) yang singkat (misalnya, 60 detik atau

kurang) untuk membantu mencegah caching data menghambat pemulihan (data alias Route 53 menyediakan TTL yang sesuai untuk Anda).

Langkah-langkah tambahan saat menggunakan beberapa Wilayah AWS

1. Buat replikasi semua sistem operasi (OS) dan kode aplikasi yang digunakan oleh beban kerja Anda di seluruh Wilayah yang Anda pilih. Buat replikasi Amazon Machine Image (AMI) yang digunakan oleh instans EC2 Anda jika perlu menggunakan solusi seperti Amazon EC2 Image Builder. Buat replikasi image kontainer yang disimpan dalam registri menggunakan solusi seperti replikasi lintas Wilayah Amazon ECR. Aktifkan replikasi Regional untuk bucket Amazon S3 apa pun yang digunakan untuk menyimpan sumber daya aplikasi.
2. Lakukan deployment sumber daya komputasi dan metadata konfigurasi Anda (seperti parameter yang disimpan di Penyimpanan Parameter AWS Systems Manager) ke beberapa Wilayah. Gunakan prosedur yang sama yang dijelaskan dalam langkah sebelumnya, tetapi replikasikan konfigurasi untuk setiap Wilayah yang Anda gunakan untuk beban kerja Anda. Gunakan solusi infrastruktur sebagai kode seperti AWS CloudFormation untuk mereproduksi konfigurasi antar-Wilayah secara seragam. Jika Anda menggunakan Wilayah sekunder dalam konfigurasi pilot light untuk pemulihan bencana, Anda dapat mengurangi jumlah sumber daya komputasi Anda ke nilai minimum untuk menghemat biaya, dengan peningkatan waktu pemulihan yang sesuai.
3. Buat replikasi data Anda dari Wilayah primer Anda ke Wilayah sekunder Anda.
    a. Tabel global Amazon DynamoDB menyediakan replika global data Anda yang dapat ditulis dari dan ke Wilayah mana pun yang didukung. Dengan layanan data lain yang dikelola AWS, seperti Amazon RDS, Amazon Aurora, dan Amazon ElastiCache, Anda menetapkan Wilayah primer (baca/tulis) dan Wilayah replika (hanya-baca). Lihat panduan pengguna dan developer layanan masing-masing untuk detail tentang replikasi Regional.
    b. Jika Anda menjalankan basis data yang Anda kelola sendiri, atur replikasi multi-Wilayah sesuai dengan petunjuk atau praktik terbaik aplikasi tersebut. Pahami prosedur failover untuk aplikasi Anda.
    c. Jika beban kerja Anda menggunakan AWS EventBridge, Anda mungkin perlu meneruskan peristiwa yang dipilih dari Wilayah primer Anda ke Wilayah sekunder Anda. Untuk melakukannya, tentukan bus peristiwa di Wilayah sekunder Anda sebagai target untuk peristiwa yang cocok di Wilayah primer Anda.
4. Pertimbangkan apakah Anda ingin menggunakan kunci enkripsi yang identik di seluruh Wilayah dan sejauh apa penggunaannya. Pendekatan umum yang menyeimbangkan keamanan dan kemudahan penggunaan adalah melalui penggunaan kunci dengan cakupan Wilayah untuk data

dan autentikasi lokal Wilayah, dan penggunaan kunci dengan cakupan global untuk enkripsi data yang direplikasi di antara Wilayah yang berbeda. AWS Key Management Service (KMS) mendukung kunci multi-wilayah untuk mendistribusikan dan melindungi kunci yang dibagikan secara aman di berbagai Wilayah.

5. Pertimbangkan AWS Global Accelerator untuk meningkatkan ketersediaan aplikasi Anda dengan mengarahkan lalu lintas ke Wilayah yang berisi titik akhir yang sehat.

Sumber daya

Praktik-praktik terbaik terkait:

- REL02-BP03 Pastikan alokasi subnet IP memperhitungkan ekspansi dan ketersediaan
- REL11-BP05 Menggunakan stabilitas statis untuk mencegah perilaku bimodal
- REL13-BP01 Menetapkan sasaran pemulihan untuk waktu henti dan kehilangan data

Dokumen terkait:

- Infrastruktur Global AWS
- Laporan resmi: Batas Isolasi Kesalahan AWS
- Ketahanan di Amazon EC2 Auto Scaling
- Amazon EC2 Auto Scaling: Contoh: Mendistribusikan instans di beberapa Zona Ketersediaan
- Cara kerja EC2 Image Builder
- Bagaimana Amazon ECS menempatkan tugas pada instans kontainer (termasuk Fargate)
- Ketahanan di AWS Lambda
- Ikhtisar Amazon S3: Mereplikasi objek
- Replikasi image privat di Amazon ECR
- Tabel Global: Replikasi Multi-Wilayah dengan DynamoDB
- Amazon ElastiCache for Redis OSS: Replikasi di beberapa Wilayah AWS menggunakan penyimpanan data global
- Ketahanan dalam Amazon RDS
- Menggunakan basis data global Amazon Aurora
- Panduan Developer AWS Global Accelerator

- Kunci Multi-Wilayah di AWS KMS
- Amazon Route 53: Mengonfigurasi failover DNS
- Panduan Developer Amazon Application Recovery Controller (ARC)
- Mengirim dan menerima peristiwa Amazon EventBridge di antara Wilayah AWS
- Membuat Aplikasi Multi-Wilayah dengan seri blog Layanan AWS
- Arsitektur Pemulihan Bencana (DR) di AWS, Bagian I: Strategi untuk Pemulihan di Cloud
- Arsitektur Pemulihan Bencana (DR) di AWS, Bagian III: Pilot Light dan Warm Standby

Video terkait:

- AWS re:Invent 2018: Pola Arsitektur untuk Aplikasi Aktif-Aktif Multi-Wilayah
- AWS re:Invent 2019: Inovasi dan operasi infrastruktur jaringan global AWS

REL10-BP02 Mengotomatiskan pemulihan untuk komponen yang dibatasi dalam satu lokasi

Jika komponen beban kerja hanya dapat dijalankan di satu Zona Ketersediaan atau di pusat data on-premise, implementasikan kemampuan untuk membangun kembali beban kerja sepenuhnya dalam lingkup tujuan pemulihan yang telah ditetapkan.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Sedang

Panduan implementasi

Jika praktik terbaik untuk melakukan deployment beban kerja ke beberapa lokasi tidak mungkin dilakukan karena adanya kendala teknologi, Anda harus mengimplementasikan jalur alternatif untuk mewujudkan ketahanan. Anda harus melakukan otomatisasi terhadap kemampuan untuk membuat ulang infrastruktur yang dibutuhkan, melakukan deployment ulang aplikasi, dan membuat ulang data yang diperlukan untuk kasus ini.

Misalnya, Amazon EMR meluncurkan semua simpul untuk klaster tertentu yang tersedia dalam Zona Ketersediaan yang sama karena menjalankan klaster di zona yang sama dapat meningkatkan kinerja aliran tugas berkat tingkat akses data yang lebih tinggi. Jika komponen ini tidak dibutuhkan untuk ketahanan beban kerja, maka Anda harus mencari cara lain untuk melakukan deployment ulang klaster dan datanya. Selain itu, untuk Amazon EMR, Anda harus menyediakan redundansi selain dengan menggunakan Multi-AZ. Anda dapat menyediakan beberapa simpul. Menggunakan EMR File

System (EMRFS), data yang ada dalam EMR dapat disimpan dalam Amazon S3, sehingga nantinya dapat direplikasi di seluruh Zona Ketersediaan atau Wilayah AWS.

Dengan cara yang serupa, untuk Amazon Redshift, secara default menyediakan klaster dalam Zona Ketersediaan yang dipilih secara acak dalam Wilayah AWS pilihan Anda. Semua simpul klaster disediakan dalam zona yang sama.

Untuk beban kerja stateful berbasis server yang di-deploy ke sebuah pusat data on-premise, Anda dapat menggunakan AWS Elastic Disaster Recovery untuk melindungi beban kerja Anda yang ada di AWS. Jika Anda sudah di-hosting di AWS, Anda dapat menggunakan Elastic Disaster Recovery untuk memberikan proteksi terhadap beban kerja Anda ke Zona Ketersediaan atau Wilayah alternatif. Elastic Disaster Recovery menggunakan replikasi tingkat blok berkelanjutan ke area staging yang ringan untuk menyediakan pemulihan aplikasi berbasis cloud dan on-premise yang cepat dan andal.

Langkah-langkah implementasi

1. Implementasikan pemulihan mandiri. Deploy instans dan kontainer Anda dengan menggunakan penskalaan otomatis jika memungkinkan. Jika tidak dapat menggunakan penskalaan otomatis, Anda harus menggunakan pemulihan otomatis untuk instans EC2 atau implementasikan otomatisasi pemulihan mandiri berdasarkan peristiwa siklus hidup kontainer Amazon EC2 atau ECS.
   - Gunakan grup Amazon EC2 Auto Scaling untuk instans atau beban kerja kontainer yang tidak memiliki persyaratan untuk alamat IP instans tunggal, alamat IP pribadi, alamat IP Elastis, dan metadata instans.
     - Data pengguna templat peluncuran dapat Anda gunakan untuk mengimplementasikan otomatisasi yang dapat memulihkan sebagian besar beban kerja secara mandiri.
   - Gunakan pemulihan otomatis instans Amazon EC2 untuk beban kerja yang memerlukan instans tunggal alamat ID, alamat IP pribadi, alamat IP Elastis, dan instans metadata.
     - Pemulihan Otomatis akan mengirimkan peringatan status pemulihan kepada sebuah topik SNS saat ada kegagalan instans yang terdeteksi.
   - Gunakan peristiwa siklus hidup instans Amazon EC2 atau peristiwa Amazon ECS untuk mengotomatiskan pemulihan mandiri jika penskalaan otomatis atau pemulihan EC2 tidak dapat digunakan.
     - Gunakan peristiwa untuk menginvokasi otomatisasi yang akan memulihkan komponen Anda berdasarkan proses logika yang diperlukan.
   - Berikan proteksi terhadap beban kerja stateful yang terbatas pada satu lokasi dengan menggunakan AWS Elastic Disaster Recovery.

Sumber daya

Dokumen terkait:

- Peristiwa Amazon ECS
- Pengait siklus hidup Amazon EC2 Auto Scaling
- Pulihkan instans Anda.
- Penskalaan otomatis layanan
- Apa itu Amazon EC2 Auto Scaling?
- AWS Elastic Disaster Recovery

REL10-BP03 Menggunakan arsitektur bulkhead untuk membatasi cakupan dampak

Implementasikan arsitektur bulkhead (juga disebut sebagai arsitektur berbasis sel) untuk membatasi efek kegagalan dalam beban kerja hingga jumlah komponen yang terbatas.

Hasil yang diinginkan: Arsitektur berbasis sel menggunakan beberapa instans beban kerja yang terisolasi, di mana setiap instans dikenal sebagai sel. Setiap sel bersifat mandiri, tidak berbagi status dengan sel yang lain, dan menangani subset permintaan beban kerja secara keseluruhan. Hal ini dapat mengurangi potensi dampak yang mungkin ditimbulkan oleh sebuah kegagalan, seperti pembaruan perangkat lunak yang buruk, sehingga hanya berdampak pada satu sel individu dan permintaan yang diprosesnya. Jika beban kerja menggunakan 10 sel untuk melayani 100 permintaan, ketika ada satu kegagalan yang terjadi, 90% dari seluruh permintaan tidak akan terpengaruh oleh kegagalan tersebut.

Anti-pola umum:

- Membiarkan sel bertumbuh tanpa batas.
- Menerapkan pembaruan kode atau deployment ke semua sel pada waktu yang sama.
- Berbagi status atau komponen antara sel (dengan pengecualian lapisan router).
- Menambahkan bisnis yang kompleks atau mengarahkan rute logika ke lapisan router.
- Tidak meminimalkan interaksi lintas sel.

Manfaat menerapkan praktik terbaik ini: Dengan arsitektur berbasis sel, banyak jenis kegagalan umum akan dibatasi di dalam sel itu sendiri, sehingga memberikan isolasi kesalahan tambahan.

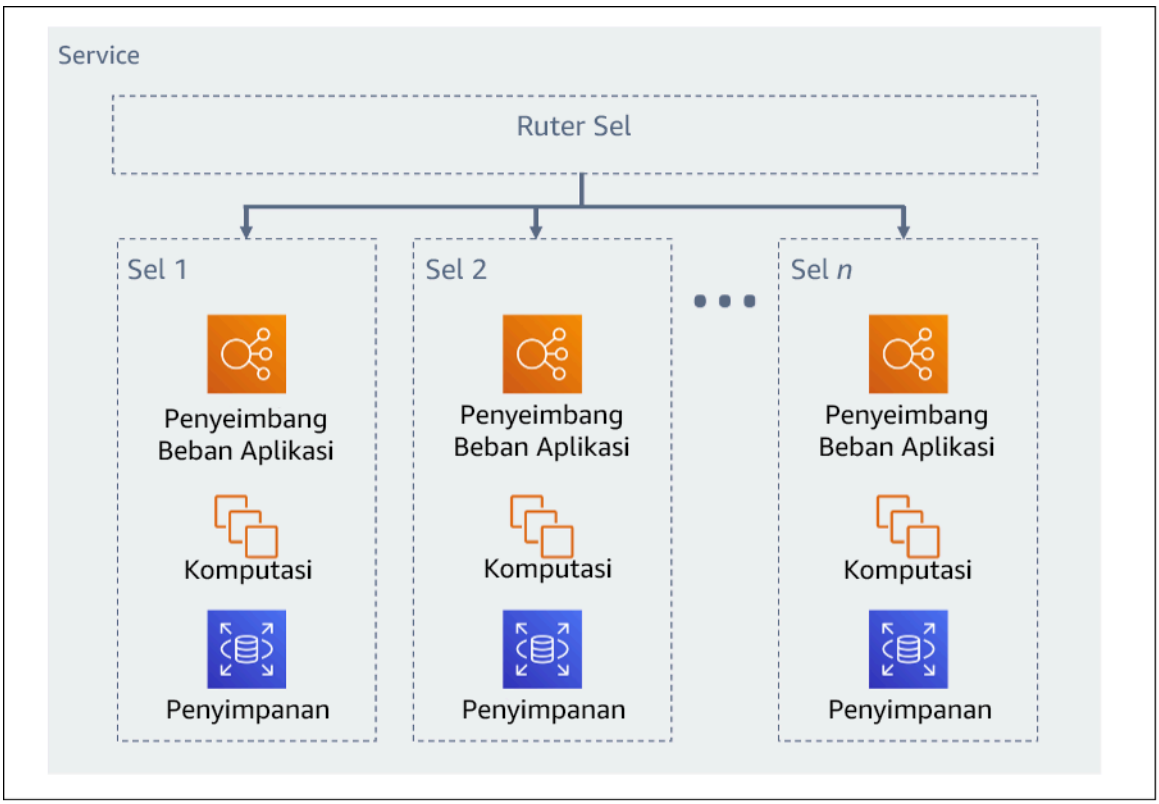
Batasan-batasan kesalahan ini dapat memberikan ketahanan terhadap berbagai tipe kegagalan yang sulit ditampung, seperti deployment kode yang tidak berhasil atau permintaan yang rusak atau menginvokasi mode kegagalan tertentu (juga dikenal sebagai permintaan pil racun).

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Tinggi

Panduan implementasi

Di sebuah kapal, bulkhead memastikan kebocoran lambung kapal hanya dibatasi dalam satu bagian lambung kapal saja. Di sistem yang kompleks, pola ini sering kali direplikasi untuk memungkinkan Anda melakukan isolasi kesalahan. Batasan-batasan isolasi kesalahan ini akan membatasi efek yang ditimbulkan kegagalan di dalam sebuah beban kerja sehingga hanya berdampak pada sejumlah komponen yang terbatas saja. Komponen-komponen yang ada di luar batasan-batasan ini tidak terpengaruh oleh kegagalan tersebut. Menggunakan beberapa batasan isolasi kesalahan, Anda dapat membatasi dampak pada beban kerja Anda. Di AWS, pelanggan dapat menggunakan beberapa Zona Ketersediaan dan Wilayah untuk menyediakan isolasi kesalahan, tetapi konsep isolasi kesalahan ini dapat diperluas ke arsitektur beban kerja juga.

Beban kerja secara keseluruhan adalah sel-sel yang dipartisi oleh sebuah kunci partisi. Kunci ini harus selaras dengan grain layanan, atau cara alami beban kerja layanan dapat dibagi lebih lanjut dengan interaksi lintas sel yang minim. Contoh kunci partisi ini antara lain ID pelanggan, ID sumber daya, atau parameter lainnya yang dapat diakses dengan mudah dalam sebagian besar panggilan API. Lapisan perutean sel mendistribusikan permintaan ke masing-masing sel berdasarkan kunci partisi tersebut dan memberikan satu titik akhir ke klien.



Gambar 11: Arsitektur berbasis sel

Langkah-langkah implementasi

Ketika mendesain sebuah arsitektur berbasis sel, ada beberapa pertimbangan desain yang harus Anda pikirkan:

1. Kunci partisi: Pertimbangan khusus harus diambil saat memilih kunci partisi.
    - Kunci ini harus sesuai dengan grain layanan, atau cara alami beban kerja dari sebuah layanan dapat dibagi lebih lanjut dengan interaksi lintas sel yang minimum. Contohnya adalah `customer ID` atau `resource ID`.
    - Kunci partisi harus tersedia dalam semua permintaan, baik secara langsung atau dengan cara yang dapat disimpulkan dengan pasti dan mudah berdasarkan parameter-parameter lain.
2. Pemetaan sel persisten: Layanan-layanan hulu seharusnya hanya berinteraksi dengan satu sel untuk siklus hidup sumber daya mereka.
    - Bergantung pada beban kerjanya, strategi migrasi sel mungkin perlu digunakan untuk memigrasikan data dari satu sel ke yang lain. Kemungkinan skenario ketika migrasi sel mungkin diperlukan adalah jika sumber daya atau pengguna tertentu yang ada di dalam beban kerja Anda menjadi terlalu besar dan memerlukan sebuah sel khusus.

- Sel tidak boleh berbagi status atau komponen antara sel.
- Oleh karena itu, interaksi lintas sel harus dihindari atau dijaga agar tetap minimum, karena interaksi tersebut akan menimbulkan dependensi antara sel, dan karena itu akan mengurangi peningkatan isolasi kesalahan.

3. Lapisan router: Lapisan router adalah sebuah komponen bersama antar sel, dan karena itu tidak dapat mengikuti strategi kompartementalisasi yang sama seperti sel.
   - Sebaiknya lapisan router mendistribusikan permintaan ke masing-masing sel dengan menggunakan algoritma pemetaan partisi dengan cara yang efisien secara komputasional, seperti menggabungkan fungsi hash kriptografi dan aritmetika modul untuk memetakan kunci partisi ke sel.
   - Untuk menghindari dampak yang timbul terhadap banyak sel, lapisan perutean harus dibuat sesederhana mungkin dan dapat diskalakan sehorizontal mungkin, sehingga Anda harus menghindari logika bisnis kompleks di dalam lapisan ini. Hal ini memiliki manfaat tambahan yakni mempermudah pemahaman terhadap harapan perilakunya di setiap waktu, sehingga memberikan kemampuan untuk diuji secara menyeluruh. Seperti yang dijelaskan oleh Colm MacCárthaigh dalam Keandalan, kerja konstan, dan secangkir kopi yang enak, desain sederhana dan pola kerja yang konstan akan menghasilkan sistem yang andal dan mengurangi anti-kerapuhan.
4. Ukuran sel: Sel harus memiliki ukuran maksimum dan tidak boleh dibiarkan tumbuh melampaui ukuran itu.
   - Ukuran maksimum harus diidentifikasi dengan melakukan pengujian menyeluruh, sampai titik rusak tercapai dan margin pengoperasian yang aman ditetapkan. Untuk detail selengkapnya tentang cara mengimplementasikan praktik pengujian, lihat REL07-BP04 Beban uji beban kerja Anda
   - Beban kerja secara keseluruhan harus bertumbuh dengan menambahkan sel-sel tambahan, sehingga beban kerja dapat diskalakan seiring dengan peningkatan permintaan.
5. Strategi Multi-AZ atau Multi-Wilayah: Beberapa lapisan ketahanan harus dimanfaatkan untuk memberikan perlindungan dari domain kegagalan yang berbeda.
   - Untuk ketahanan, Anda harus menggunakan sebuah pendekatan yang membangun berlapis-lapis pertahanan. Satu lapisan melindungi Anda dari gangguan yang lebih kecil dan lebih umum dengan membangun arsitektur yang memiliki ketersediaan tinggi dengan menggunakan beberapa AZ. Lapisan pertahanan lainnya ditujukan untuk memberikan proteksi dari peristiwa-peristiwa langka seperti bencana alam yang meluas dan gangguan tingkat Wilayah. Lapisan kedua ini melibatkan perancangan aplikasi agar menjangkau beberapa Wilayah AWS.

Mengimplementasikan strategi multi-Wilayah untuk beban kerja dapat membantu Anda melindunginya dari bencana alam yang menjangkau dan memengaruhi wilayah geografis yang luas di suatu negara, atau kegagalan-kegagalan teknis yang mencakup seluruh Wilayah. Perhatikan bahwa mengimplementasikan arsitektur multi-Wilayah dapat menjadi suatu hal yang sangat kompleks, dan biasanya tidak diperlukan untuk sebagian besar beban kerja. Untuk detail selengkapnya, lihat REL10-BP01 Melakukan deployment beban kerja ke beberapa lokasi.

6. Kode deployment: Strategi deployment kode yang bertahap seharusnya lebih disarankan untuk digunakan daripada menerapkan deployment perubahan kode ke semua sel secara bersamaan.
   - Hal ini akan membantu Anda dalam meminimalkan potensi kegagalan pada beberapa sel karena deployment yang buruk atau kesalahan manusia. Untuk detail selengkapnya, silakan lihat Melakukan otomatisasi deployment secara aman dan otonom.

Sumber daya

Praktik-praktik terbaik terkait:

- REL07-BP04 Beban uji beban kerja Anda
- REL10-BP01 Melakukan deployment beban kerja ke beberapa lokasi

Dokumen terkait:

- Keandalan, kerja konstan, dan secangkir kopi yang enak
- AWS dan Kompartementalisasi
- Isolasi beban kerja dengan menggunakan shuffle-sharding
- Melakukan otomatisasi deployment secara aman dan otonom

Video terkait:

- AWS re:Invent 2018: Loop Tertutup dan Pikiran Terbuka: Cara Mengendalikan Sistem, Besar dan Kecil
- AWS re:Invent 2018: Cara AWS Meminimalkan Radius Dampak Kegagalan (ARC338)
- Shuffle-sharding: AWS re:Invent 2019: Memperkenalkan Amazon Builders' Library (DOP328)
- AWS Summit ANZ 2021 - Semuanya gagal, sepanjang waktu: Merancang untuk ketahanan

Rancang beban kerja Anda agar bertahan dalam kegagalan komponen

Beban kerja yang membutuhkan ketersediaan tinggi dan waktu rata-rata untuk pemulihan (MTTR) rendah harus didesain dan dikonfigurasi agar tangguh.

Praktik terbaik

- REL11-BP01 Memantau semua komponen beban kerja untuk mendeteksi kegagalan
- REL11-BP02 Melakukan failover ke sumber daya yang sehat
- REL11-BP03 Melakukan otomatisasi pemulihan di semua lapisan
- REL11-BP04 Mengandalkan bidang data dan bukan bidang kontrol selama pemulihan
- REL11-BP05 Menggunakan stabilitas statis untuk mencegah perilaku bimodal
- REL11-BP06 Mengirimkan notifikasi ketika peristiwa memengaruhi ketersediaan
- REL11-BP07 Merancang produk Anda agar memenuhi target-target ketersediaan dan perjanjian tingkat layanan (SLA) waktu aktif

REL11-BP01 Memantau semua komponen beban kerja untuk mendeteksi kegagalan

Lakukan pemantauan terus-menerus terhadap kondisi beban kerja agar Anda dan sistem otomatis Anda langsung mengetahui adanya penurunan kualitas atau kegagalan ketika muncul. Pantau indikator kinerja utama (KPI) berdasarkan nilai bisnis.

Semua mekanisme pemulihan dan penyembuhan harus dimulai dengan kemampuan untuk mendeteksi adanya masalah secara cepat. Kegagalan teknis harus dideteksi terlebih dahulu sehingga dapat diselesaikan. Namun demikian, ketersediaan didasarkan pada kemampuan beban kerja Anda untuk menghadirkan nilai bisnis, sehingga indikator performa utama (KPI) yang mengukurnya harus menjadi bagian dari strategi deteksi dan perbaikan Anda.

Hasil yang diinginkan: Komponen penting dari suatu beban kerja dipantau secara independen untuk mendeteksi dan memperingatkan adanya kegagalan pada saat dan di bagian mana kegagalan tersebut terjadi.

Anti-pola umum:

- Tidak ada alarm yang dikonfigurasi, sehingga pemadaman (outage) terjadi tanpa notifikasi.
- Alarm tersedia, tetapi pada ambang batas yang tidak menyediakan waktu yang cukup untuk bereaksi.
- Metrik tidak dikumpulkan cukup sering untuk memenuhi sasaran waktu pemulihan (RTO).
- Hanya antarmuka beban kerja yang terlihat oleh pelanggan yang secara aktif dipantau.
- Hanya mengumpulkan metrik-metrik teknis, dan mengabaikan metrik-metrik fungsi bisnis.
- Tidak ada metrik yang mengukur pengalaman pengguna beban kerja.
- Terlalu banyak pemantau yang dibuat.

Manfaat menerapkan praktik terbaik ini: Pemantauan yang sesuai di semua lapisan akan memungkinkan Anda untuk menghemat waktu pemulihan karena berkurangnya waktu deteksi.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Tinggi

Panduan implementasi

Identifikasi semua beban kerja yang akan ditinjau untuk pemantauan. Setelah Anda mengidentifikasi semua komponen beban kerja yang perlu dipantau, selanjutnya Anda harus menentukan rentang waktu pemantauan. Rentang waktu pemantauan akan berdampak langsung pada seberapa cepat pemulihan dapat dimulai berdasarkan waktu yang diperlukan untuk mendeteksi sebuah kegagalan. Rata-rata waktu deteksi (MTTD) adalah lamanya waktu antara terjadinya kegagalan dan ketika operasi perbaikan dimulai. Daftar layanan harus luas dan lengkap.

Pemantauan harus mencakup semua lapisan tumpukan aplikasi termasuk aplikasi, platform, infrastruktur, dan jaringan.

Strategi pemantauan Anda harus mempertimbangkan dampak kegagalan abu-abu. Untuk detail lebih lanjut tentang kegagalan abu-abu (samar), lihat Kegagalan abu-abu di laporan resmi Pola Ketahanan Multi-AZ Tingkat Lanjut.

Langkah-langkah implementasi

- Rentang waktu pemantauan Anda bergantung pada seberapa cepat waktu pemulihan yang Anda perlukan. Waktu pemulihan Anda ditentukan oleh waktu yang diperlukan untuk pulih, sehingga Anda harus menentukan frekuensi pengumpulan dengan cara menghitung waktu ini dan sasaran waktu pemulihan (RTO) Anda.
- Konfigurasikan pemantauan mendetail untuk komponen dan layanan terkelola.

  - Tentukan apakah diperlukan pemantauan mendetail untuk instans EC2 dan Auto Scaling (penskalaan otomatis). Pemantauan mendetail menyediakan metrik rentang waktu satu menit, sedangkan pemantauan default menyediakan metrik rentang waktu lima menit.
  - Tentukan apakah diperlukan pemantauan yang ditingkatkan untuk RDS. Pemantauan yang ditingkatkan menggunakan sebuah agen di instans RDS untuk memperoleh informasi bermanfaat tentang berbagai alur atau proses.
  - Tentukan persyaratan-persyaratan pemantauan komponen nirserver penting untuk Lambda, API Gateway, Amazon EKS, Amazon ECS, dan semua jenis penyeimbang beban..
  - Tentukan persyaratan-persyaratan pemantauan komponen penyimpanan untuk Amazon S3, Amazon FSx, Amazon EFS, dan Amazon EBS.
- Buat metrik kustom untuk mengukur indikator performa utama (KPI) bisnis. Beban kerja mengimplementasikan fungsi-fungsi bisnis utama, yang harus digunakan sebagai KPI yang membantu mengidentifikasi kapan sebuah masalah tidak langsung terjadi.
- Pantau pengalaman pengguna untuk mendeteksi terjadinya kegagalan menggunakan canary pengguna. Pengujian transaksi sintetis (juga disebut pengujian canary, tetapi bedakan dengan deployment canary) yang dapat menjalankan dan menyimulasikan perilaku pelanggan adalah salah satu proses pengujian yang paling penting. Jalankan pengujian ini secara konstan terhadap titik akhir beban kerja Anda dari beragam lokasi jarak jauh.
- Buat metrik kustom yang melacak pengalaman pengguna. Jika Anda dapat menginstrumentasikan pengalaman pelanggan, Anda dapat menentukan kapan pengalaman pelanggan mengalami degradasi.
- Atur alarm untuk mendeteksi saat ada bagian dari beban kerja Anda yang tidak berfungsi dengan baik, dan untuk menunjukkan kapan harus menerapkan penskalaan secara otomatis pada sumber daya. Alarm dapat ditampilkan secara visual di dasbor, mengirimkan peringatan melalui Amazon SNS atau email, dan menggunakan Auto Scaling (penskalaan otomatis) untuk menaikkan atau menurunkan skala sumber daya beban kerja.
- Buat dasbor untuk memvisualisasikan metrik Anda. Dasbor dapat digunakan untuk melihat tren, penyimpangan, dan indikator masalah lain yang mungkin muncul, atau menyediakan penanda untuk masalah yang ingin Anda selidiki.
- Buat pemantauan penelusuran terdistribusi untuk layanan-layanan Anda. Dengan pemantauan terdistribusi, Anda dapat memahami cara kerja aplikasi Anda dan layanan-layanan yang mendasarinya dalam melakukan identifikasi dan memecahkan akar penyebab masalah dan kesalahan performa.

- Buat dasbor dan pengumpulan data sistem pemantauan (menggunakan CloudWatch atau X-Ray) di Wilayah dan akun terpisah.
- Terus dapatkan informasi tentang penurunan layanan terkait AWS Health. Buat notifikasi peristiwa AWS Health sesuai keperluan yang dikirim ke saluran email dan obrolan melalui Notifikasi Pengguna AWS serta integrasikan secara programatis dengan alat pemantauan dan peringatan Anda melalui Amazon EventBridge.

Sumber daya

Praktik-praktik terbaik terkait:

- Definisi Ketersediaan
- REL11-BP06 Mengirimkan Notifikasi ketika peristiwa memengaruhi ketersediaan

Dokumen terkait:

- Amazon CloudWatch Synthetics memungkinkan Anda untuk membuat canary pengguna
- Aktifkan atau Nonaktifkan Pemantauan Mendetail untuk Instans Anda
- Pemantauan yang Ditingkatkan
- Memantau Grup Auto Scaling dan Instans Anda Menggunakan Amazon CloudWatch
- Memublikasikan Metrik Kustom
- Menggunakan Alarm Amazon CloudWatch
- Menggunakan Dasbor CloudWatch
- Menggunakan Dasbor CloudWatch Lintas Akun dan Lintas Wilayah
- Menggunakan Penelusuran X-Ray Lintas Akun dan Lintas Wilayah
- Memahami ketersediaan

Video terkait:

- Memitigasi kegagalan abu-abu

Contoh terkait:

- One Observability Workshop: Explore X-Ray

Alat terkait:

- CloudWatch
- X-Ray CloudWatch

REL11-BP02 Melakukan failover ke sumber daya yang sehat

Jika terjadi kegagalan sumber daya, sumber daya yang sehat harus terus melayani permintaan. Untuk kerusakan lokasi (seperti Zona Ketersediaan atau Wilayah AWS) pastikan Anda sudah memiliki sistem untuk melakukan failover ke sumber daya yang sehat di lokasi-lokasi yang tidak terkena gangguan.

Saat merancang sebuah layanan, distribusikan beban di seluruh sumber daya, Zona Ketersediaan, atau Wilayah. Oleh karena itu, kegagalan atau gangguan yang terjadi pada satu sumber daya individu dapat dimitigasi dengan mengalihkan lalu lintas ke sumber daya sehat yang masih ada. Pertimbangkan bagaimana layanan ditemukan dan dirutekan jika ada suatu kegagalan yang terjadi.

Rancang layanan Anda dengan mempertimbangkan pemulihan kesalahan. Di AWS, kami merancang layanan untuk meminimalkan waktu pemulihan dari kegagalan dan dampak yang ditimbulkannya terhadap data. Layanan-layanan kami utamanya menggunakan penyimpanan data yang mengenali permintaan hanya setelah disimpan dalam waktu lama di beberapa replika di dalam suatu Wilayah. Layanan dan sumber daya ini dibangun untuk menggunakan isolasi berbasis sel dan menggunakan isolasi kesalahan yang disediakan oleh Zona Ketersediaan. Kami banyak menggunakan otomatisasi di dalam prosedur-prosedur operasional kami. Kami juga mengoptimalkan fungsionalitas “ganti dan mulai ulang” kami untuk melakukan pemulihan secara cepat dari gangguan.

Pola dan desain yang memungkinkan failover berbeda-beda untuk setiap layanan platform AWS. Banyak layanan terkelola native AWS adalah layanan yang secara native multi-Zona Ketersediaan (seperti Lambda atau API Gateway). Layanan-layanan AWS lain (seperti EC2 dan EKS) memerlukan desain praktik terbaik khusus untuk mendukung failover sumber daya atau penyimpanan data di seluruh AZ.

Pemantauan harus disiapkan untuk memeriksa apakah sumber daya failover sehat, melacak kemajuan sumber daya yang melakukan failover, dan memantau pemulihan proses bisnis.

Hasil yang diinginkan: Sistem mampu secara otomatis atau manual menggunakan sumber daya baru untuk pulih dari degradasi.

Anti-pola umum:

- Perencanaan kegagalan bukan bagian dari fase perencanaan dan desain.
- RTO dan RPO tidak ditetapkan.
- Pemantauan yang tidak memadai untuk mendeteksi sumber daya yang gagal.
- Isolasi domain kegagalan yang layak.
- Kegagalan Multi-Wilayah tidak dipertimbangkan.
- Deteksi kegagalan terlalu sensitif atau agresif saat memutuskan untuk melakukan failover.
- Tidak menguji atau memvalidasi desain failover.
- Melakukan otomatisasi pemulihan otomatis, tetapi tidak memberikan notifikasi bahwa pemulihan diperlukan.
- Kurangnya periode peredaman untuk menghindari kegagalan berulang yang terlalu cepat.

Manfaat menerapkan praktik terbaik ini: Anda dapat membangun sistem-sistem yang lebih tangguh yang mempertahankan keandalan saat mengalami kegagalan dengan melakukan degradasi secara mulus dan pulih dengan cepat.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Tinggi

Panduan implementasi

Layanan-layanan AWS, seperti Elastic Load Balancing dan Amazon EC2 Auto Scaling, dapat membantu Anda mendistribusikan beban di seluruh sumber daya dan Zona Ketersediaan. Oleh karena itu, kegagalan yang terjadi pada suatu sumber daya individu (seperti instans EC2) atau gangguan pada suatu Zona Ketersediaan dapat dimitigasi dengan mengalihkan lalu lintas ke sumber daya sehat yang masih ada.

Untuk beban kerja multi-Wilayah, desainnya lebih rumit. Misalnya, replika baca lintas Wilayah memungkinkan Anda untuk melakukan deployment data ke beberapa Wilayah AWS. Namun, failover masih diperlukan untuk mempromosikan replika baca ke primer kemudian mengarahkan lalu lintas Anda ke titik akhir baru. Amazon Route 53, Pengontrol Pemulihan Aplikasi Amazon (ARC), Amazon CloudFront, dan AWS Global Accelerator dapat membantu merutekan lalu lintas ke berbagai Wilayah AWS.

Layanan-layanan AWS, seperti Amazon S3, Lambda, API Gateway, Amazon SQS, Amazon SNS, Amazon SES, Amazon Pinpoint, Amazon ECR, AWS Certificate Manager, EventBridge, atau Amazon DynamoDB, secara otomatis di-deploy ke beberapa Zona Ketersediaan oleh AWS. Jika terjadi

kegagalan, layanan-layanan AWS ini secara otomatis merutekan lalu lintas ke lokasi yang sehat. Data disimpan secara redundan di beberapa Zona Ketersediaan dan tetap tersedia.

Untuk Amazon RDS, Amazon Aurora, Amazon Redshift, Amazon EKS, atau Amazon ECS, Multi-AZ adalah opsi konfigurasi. AWS dapat mengarahkan lalu lintas ke instans yang sehat jika failover dimulai. Tindakan failover ini dapat diambil oleh AWS atau sebagaimana diperlukan oleh pelanggan

Untuk instans-instans Amazon EC2 atau tugas-tugas Amazon Redshift, Amazon ECS, atau pod Amazon EKS, pilihlah Zona Ketersediaan yang akan menjadi tujuan deployment. Untuk beberapa desain, Elastic Load Balancing menyediakan solusi untuk mendeteksi instans yang ada di zona yang tidak sehat, dan kemudian merutekan lalu lintas ke zona yang sehat. Penyeimbang Beban Elastis dapat merutekan lalu lintas ke komponen di pusat data on-premise Anda.

Untuk failover lalu lintas Multi-Wilayah, pengalihan rute dapat memanfaatkan Amazon Route 53, Pengontrol Pemulihan Aplikasi Amazon, AWS Global Accelerator, Route 53 Private DNS for VPC, atau CloudFront untuk menyediakan cara menentukan domain internet dan menetapkan kebijakan perutean, termasuk pemeriksaan kondisi, untuk merutekan lalu lintas ke Wilayah yang berkondisi baik. AWS Global Accelerator menyediakan alamat IP statis yang berfungsi sebagai titik masuk tetap ke aplikasi Anda, lalu merutekan ke titik akhir di Wilayah AWS yang Anda pilih, menggunakan jaringan global AWS, bukan internet, demi performa dan keandalan yang lebih baik.

Langkah-langkah implementasi

- Buat desain failover untuk semua aplikasi dan layanan yang sesuai. Isolasi setiap komponen arsitektur dan buat desain failover yang memenuhi RTO dan RPO untuk masing-masing komponen.
- Konfigurasikan lingkungan yang lebih rendah (seperti pengembangan atau pengujian) dengan semua layanan yang diharuskan memiliki rencana failover. Lakukan deployment solusi dengan menggunakan infrastruktur sebagai kode (IaC) untuk memastikan kemampuan pengulangan.
- Konfigurasikan lokasi pemulihan, misalnya Wilayah kedua, untuk mengimplementasikan dan menguji desain failover. Jika perlu, sumber daya untuk pengujian dapat dikonfigurasi secara sementara untuk membatasi biaya-biaya tambahan.
- Tentukan rencana failover mana yang akan diotomatisasi oleh AWS, yang dapat diotomatisasi oleh proses DevOps, dan mana yang mungkin dilakukan secara manual. Dokumentasikan dan ukur RTO dan RPO dari masing-masing layanan.
- Buatlah sebuah playbook failover dan sertakan semua langkah untuk melakukan failover pada setiap sumber daya, aplikasi, dan layanan.

- Buatlah sebuah playbook failback dan sertakan semua langkah untuk melakukan failback (dengan pengaturan waktu) pada setiap sumber daya, aplikasi, dan layanan
- Buatlah sebuah rencana untuk memulai dan melatih playbook. Gunakan simulasi dan pengujian kekacauan untuk menguji langkah-langkah dan otomatisasi playbook.
- Untuk gangguan lokasi (seperti Zona Ketersediaan atau Wilayah AWS), pastikan Anda memiliki sistem untuk melakukan failover ke sumber daya yang sehat di lokasi-lokasi yang tidak terkena gangguan. Periksa kuota, tingkat penskalaan otomatis, dan sumber daya yang berjalan sebelum pengujian failover.

Sumber daya

Praktik terbaik Well-Architected terkait:

- REL13 - Merencanakan DR
- REL10 - Menggunakan isolasi kesalahan untuk melindungi beban kerja Anda

Dokumen terkait:

- Menetapkan Target RTO dan RPO
- Failover menggunakan perutean tertimbang Route 53
- Pemulihan Bencana dengan Pengontrol Pemulihan Aplikasi Amazon
- EC2 dengan penskalaan otomatis
- Deployment EC2 - Multi-AZ
- Deployment ECS – Multi-AZ
- Beralih lalu lintas menggunakan Pengontrol Pemulihan Aplikasi Amazon
- Lambda dengan Penyeimbang Beban Aplikasi dan Failover
- Replikasi ACM dan Failover
- Replikasi Penyimpanan Parameter dan Failover
- Replikasi lintas wilayah ECR dan Failover
- Konfigurasi replikasi lintas wilayah manajer rahasia
- Mengaktifkan replikasi lintas wilayah untuk EFS dan Failover
- Replikasi Lintas Wilayah EFS dan Failover
- Failover Jaringan

- Failover titik akhir S3 menggunakan MRAP
- Membuat replikasi lintas wilayah untuk S3
- Panduan untuk Failover Lintas Wilayah dan Graceful Failback di AWS
- Failover menggunakan akselerator global multi-wilayah
- Failover dengan DRS

Contoh terkait:

- Pemulihan Bencana di AWS
- Pemulihan Bencana Elastis di AWS

REL11-BP03 Melakukan otomatisasi pemulihan di semua lapisan

Setelah kegagalan dideteksi, gunakan kemampuan otomatis untuk melakukan tindakan perbaikan. Degradasi dapat dipulihkan secara otomatis melalui mekanisme servis internal atau memerlukan sumber daya untuk dimulai ulang atau dihapus melalui tindakan remediasi.

Untuk aplikasi yang dikelola secara mandiri dan perbaikan lintas-Wilayah, desain pemulihan dan proses perbaikan otomatis dapat ditarik dari praktik terbaik yang ada sekarang.

Kemampuan untuk memulai ulang atau menghapus sumber daya adalah alat yang penting untuk melakukan remediasi kegagalan. Salah satu praktik terbaiknya adalah membuat layanan menjadi stateless, jika memungkinkan. Praktik ini mencegah hilangnya data atau ketersediaan pada saat sumber daya dimulai ulang. Di cloud, Anda dapat (dan umumnya harus) mengganti seluruh sumber daya (misalnya, instans komputasi atau fungsi nirserver) sebagai bagian dari proses mulai ulang. Tindakan memulai ulang itu sendiri adalah cara yang mudah dan andal untuk pulih dari kegagalan. Ada berbagai jenis kegagalan yang terjadi di dalam beban kerja. Kegagalan dapat terjadi di perangkat keras, perangkat lunak, komunikasi, dan operasi.

Memulai ulang atau mencoba ulang juga berlaku untuk permintaan-permintaan jaringan. Terapkan pendekatan pemulihan yang sama terhadap peristiwa waktu habis jaringan maupun kegagalan dependensi di mana dependensi menunjukkan kesalahan. Kedua peristiwa tersebut memiliki efek yang serupa terhadap sistem, sehingga alih-alih berupaya untuk menjadikan masing-masing sebagai kasus spesial, terapkan strategi serupa berupa coba ulang terbatas dengan penundaan eksponensial dan jitter. Kemampuan untuk memulai ulang sebuah adalah mekanisme pemulihan yang disertakan dalam komputasi berorientasi pemulihan dan arsitektur klaster ketersediaan tinggi.

Hasil yang diinginkan: Tindakan otomatis dilakukan untuk meremediasi deteksi kegagalan.

Anti-pola umum:

- Menyediakan sumber daya tanpa penskalaan otomatis.
- Melakukan deployment aplikasi di instans atau kontainer secara terpisah.
- Melakukan deployment aplikasi yang tidak dapat dilakukan ke beberapa lokasi tanpa menggunakan pemulihan otomatis.
- Memulihkan secara manual aplikasi yang gagal dipulihkan oleh penskalaan otomatis dan pemulihan otomatis.
- Tidak ada otomatisasi untuk failover basis data.
- Tidak ada metode otomatis untuk mengalihkan rute lalu lintas ke titik akhir baru.
- Tidak ada replikasi penyimpanan.

Manfaat menerapkan praktik terbaik ini: Pemulihan otomatis dapat mengurangi waktu rata-rata pemulihan dan meningkatkan ketersediaan Anda.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Tinggi

Panduan implementasi

Desain untuk Amazon EKS atau layanan Kubernetes lainnya harus mencakup replika minimum dan maksimum atau set stateful dan penyesuaian ukuran klaster dan grup simpul minimum. Mekanisme ini menyediakan jumlah minimum sumber daya pemrosesan yang tersedia secara terus-menerus sambil secara otomatis memulihkan kegagalan apa pun menggunakan bidang kontrol Kubernetes.

Pola desain yang diakses melalui penyeimbang beban menggunakan klaster komputasi harus memanfaatkan grup Auto Scaling (penskalaan otomatis). Elastic Load Balancing (ELB) secara otomatis mendistribusikan lalu lintas aplikasi yang masuk di beberapa target dan perangkat virtual di satu atau beberapa Zona Ketersediaan (AZ).

Desain berbasis komputasi terklaster yang tidak menggunakan penyeimbangan beban harus dirancang ukurannya untuk kehilangan setidaknya satu simpul. Dengan begitu, layanan dapat terus berjalan dalam kapasitas yang kemungkinan lebih rendah saat layanan memulihkan simpul baru. Contoh layanannya adalah Mongo, DynamoDB Accelerator, Amazon Redshift, Amazon EMR, Cassandra, Kafka, MSK-EC2, Couchbase, ELK, dan Amazon OpenSearch Service. Banyak dari layanan-layanan ini dapat dirancang dengan fitur pemulihan otomatis tambahan. Beberapa teknologi

klaster harus menghasilkan peringatan atas hilangnya sebuah simpul yang memicu alur kerja otomatis atau manual untuk membuat ulang simpul baru. Alur kerja ini dapat diotomatisasi dengan menggunakan AWS Systems Manager untuk meremediasi masalah dengan cepat.

Amazon EventBridge dapat digunakan untuk memantau dan memfilter peristiwa seperti alarm CloudWatch atau perubahan status di layanan AWS lainnya. Berdasarkan informasi peristiwa, layanan ini kemudian dapat menginvokasi AWS Lambda, Systems Manager Automation, atau target lainnya untuk menjalankan logika remediasi kustom pada beban kerja Anda. Amazon EC2 Auto Scaling dapat dikonfigurasi untuk memeriksa kondisi instans EC2. Jika instans tidak berada dalam status berjalan, atau jika status sistem adalah terganggu, Amazon EC2 Auto Scaling akan menganggap instans tersebut sebagai instans tidak sehat dan kemudian meluncurkan instans pengganti. Untuk penggantian skala besar (seperti hilangnya seluruh Zona Ketersediaan), sebaiknya gunakan stabilitas statis karena ketersediaannya yang tinggi.

Langkah-langkah implementasi

- Gunakan grup Auto Scaling (penskalaan otomatis) untuk melakukan deployment tingkatan di sebuah beban kerja. Penskalaan otomatis dapat melakukan pemulihan mandiri pada aplikasi tanpa status, dan menambahkan atau menghapus kapasitas.
- Untuk instans komputasi yang disebutkan sebelumnya, gunakan penyeimbangan beban dan pilih jenis penyeimbang beban yang sesuai.
- Pertimbangkan pemulihan untuk Amazon RDS. Dengan instans siaga, konfigurasikan untuk failover otomatis ke instans siaga tersebut. Untuk Amazon RDS Read Replica, alur kerja otomatis diperlukan untuk membuat replika baca primer.
- Implementasikan pemulihan otomatis pada instans EC2 dengan aplikasi ter-deploy yang tidak dapat di-deploy di beberapa lokasi, dan dapat mentoleransi boot ulang setelah kegagalan. Pemulihan otomatis dapat digunakan untuk mengganti perangkat keras yang mengalami kegagalan dan memulai ulang instans ketika aplikasi tidak dapat diterapkan di beberapa lokasi. Metadata instans dan alamat IP terkait akan disimpan, begitu juga dengan volume EBS dan titik pemasangan (mount) ke Amazon Elastic File System atau File Systems untuk Lustre dan Windows. Dengan menggunakan AWS OpsWorks, Anda dapat mengonfigurasi pemulihan otomatis instans EC2 pada tingkat lapisan.
- Implementasikan pemulihan otomatis dengan menggunakan AWS Step Functions dan AWS Lambda ketika Anda tidak dapat menggunakan penskalaan otomatis atau pemulihan otomatis, atau ketika pemulihan otomatis gagal. Ketika Anda tidak dapat menggunakan penskalaan otomatis, dan tidak dapat menggunakan pemulihan otomatis atau pemulihan otomatis gagal, Anda dapat mengotomatiskan pemulihan dengan menggunakan AWS Step Functions dan AWS Lambda.

- Amazon EventBridge dapat digunakan untuk memantau dan memfilter peristiwa seperti alarm CloudWatch atau perubahan status di layanan AWS lainnya. Berdasarkan informasi peristiwa, layanan ini kemudian dapat menginvokasi AWS Lambda (atau target-target lainnya) untuk menjalankan logika remediasi kustom pada beban kerja Anda.

Sumber daya

Praktik-praktik terbaik terkait:

- Definisi Ketersediaan
- REL11-BP01 Memantau semua komponen beban kerja untuk mendeteksi kegagalan

Dokumen terkait:

- Cara Kerja AWS Auto Scaling
- Pemulihan Otomatis Amazon EC2
- Amazon Elastic Block Store (Amazon EBS)
- Amazon Elastic File System (Amazon EFS)
- Apa itu Amazon FSx for Lustre?
- Apa itu Amazon FSx for Windows File Server?
- AWS OpsWorks: Menggunakan Auto Healing untuk Mengganti Instans yang Gagal
- Apa itu AWS Step Functions?
- Apa itu AWS Lambda?
- Apa itu Amazon EventBridge?
- Menggunakan Alarm Amazon CloudWatch
- Failover Amazon RDS
- SSM - Systems Manager Automation
- Praktik Terbaik Arsitektur Tangguh

Video terkait:

- Penyediaan Secara Otomatis dan Menskalakan Layanan OpenSearch
- Failover Otomatis Amazon RDS

Contoh terkait:

- Lokakarya Failover Amazon RDS

Alat terkait:

- CloudWatch
- X-Ray CloudWatch

REL11-BP04 Mengandalkan bidang data dan bukan bidang kontrol selama pemulihan

bidang kontrol menyediakan API administratif yang digunakan untuk membuat, membaca, dan mendeskripsikan, memperbarui, menghapus, dan mencantumkan (CRUDL) sumber daya, sedangkan bidang data menangani lalu lintas layanan sehari-hari. Saat mengimplementasikan respons pemulihan atau mitigasi terhadap peristiwa yang berpotensi berdampak pada ketahanan, fokuslah pada penggunaan operasi bidang kontrol dalam jumlah minim untuk memulihkan, mengubah skala, mengembalikan, memperbaiki, atau melakukan failover layanan. Tindakan bidang data harus menggantikan aktivitas apa pun selama terjadi peristiwa degradasi ini.

Misalnya, berikut ini adalah semua tindakan bidang kontrol: meluncurkan instans komputasi baru, membuat penyimpanan blok, dan mendeskripsikan layanan antrean. Saat Anda meluncurkan instans komputasi, bidang kontrol harus melakukan beberapa tugas seperti menemukan temuan host fisik dengan kapasitas, mengalokasikan antarmuka jaringan, menyiapkan volume penyimpanan blok lokal, menghasilkan kredensial, dan menambahkan aturan keamanan. Orkestrasi bidang kontrol cenderung rumit.

Hasil yang diinginkan: Ketika sumber daya memasuki keadaan terganggu, sistem mampu pulih secara otomatis atau manual dengan mengalihkan lalu lintas dari sumber daya yang terganggu ke sumber daya yang sehat.

Anti-pola umum:

- Ketergantungan pada pengubahan catatan DNS untuk merutekan ulang lalu lintas.
- Ketergantungan pada operasi penskalaan bidang kontrol untuk menggantikan komponen yang terganggu karena sumber daya yang disediakan tidak memadai.

- Mengandalkan tindakan bidang kontrol ekstensif multi-API dan multi layanan untuk meremediasi kategori gangguan apa pun.

Manfaat menerapkan praktik terbaik ini: Peningkatan tingkat keberhasilan untuk remediasi otomatis dapat mengurangi waktu rata-rata pemulihan Anda dan meningkatkan ketersediaan beban kerja.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Sedang: Untuk jenis degradasi layanan tertentu, bidang kontrol terpengaruh. Ketergantungan pada penggunaan bidang kontrol secara ekstensif untuk remediasi dapat meningkatkan waktu pemulihan (RTO) dan rata-rata waktu untuk pulih (MTTR).

Panduan implementasi

Untuk membatasi tindakan bidang data, lakukan penilaian atas setiap layanan untuk menemukan tindakan-tindakan yang diperlukan untuk memulihkan layanan.

Manfaatkan Pengendali Pemulihan Aplikasi Amazon untuk menggeser lalu lintas DNS. Fitur-fitur ini akan terus-menerus memantau kemampuan aplikasi Anda untuk pulih dari kegagalan, dan memampukan Anda untuk mengontrol pemulihan aplikasi di beberapa Wilayah AWS, Zona Ketersediaan, dan on-premise.

Kebijakan perutean Route 53 menggunakan bidang kontrol, jadi jangan mengandalkannya untuk pemulihan. Bidang data Route 53 menjawab kueri DNS, dan melakukan serta mengevaluasi pemeriksaan kondisi. Mereka didistribusikan secara global dan dirancang untuk perjanjian tingkat layanan (SLA) ketersediaan 100%.

Konsol dan API manajemen Route 53 di mana Anda membuat, memperbarui, dan menghapus sumber daya Route 53 dijalankan di bidang kontrol yang didesain untuk memprioritaskan durabilitas dan konsistensi tinggi yang Anda perlukan ketika mengelola DNS. Untuk mencapai hal ini, bidang kontrol ditempatkan di satu Wilayah: AS Timur (Virginia Utara). Walaupun kedua sistem dibangun agar sangat andal, bidang kontrol tidak disertakan dalam SLA. Mungkin ada peristiwa langka di mana desain tangguh bidang data memungkinkannya untuk mempertahankan ketersediaan sedangkan bidang kontrol tidak. Untuk mekanisme failover dan pemulihan bencana, Anda harus menggunakan fungsi bidang data untuk memberikan keandalan yang sebaik mungkin.

Rancang infrastruktur komputasi Anda agar mencapai stabilitas statis, sehingga Anda dapat menghindari penggunaan bidang kontrol selama insiden. Misalnya, jika Anda menggunakan instans Amazon EC2, jangan menyediakan instans baru secara manual atau menginstruksikan Grup Auto Scaling untuk menambahkan instans sebagai respons. Untuk tingkat ketahanan tertinggi, berikan

kapasitas yang cukup di klaster yang digunakan untuk failover. Jika ambang kapasitas ini harus dibatasi, tetapkan throttling pada keseluruhan sistem menyeluruh untuk membatasi lalu lintas total yang mencapai set sumber daya terbatas.

Untuk layanan-layanan seperti Amazon DynamoDB, Amazon API Gateway, penyeimbang beban dan nirserver AWS Lambda, penggunaan layanan-layanan tersebut memanfaatkan bidang data. Namun, pembuatan fungsi baru, penyeimbang beban, API gateway, atau tabel DynamoDB adalah tindakan bidang kontrol dan harus diselesaikan sebelum degradasi sebagai persiapan peristiwa dan latihan tindakan failover. Untuk Amazon RDS, tindakan bidang data memungkinkan akses ke data.

Untuk informasi selengkapnya tentang bidang data, bidang kontrol, dan bagaimana AWS membangun layanan untuk memenuhi target ketersediaan tinggi, silakan lihat Stabilitas statis menggunakan Zona Ketersediaan.

Pahami operasi mana yang ada di bidang data dan mana yang ada di bidang kontrol.

Langkah-langkah implementasi

Untuk setiap beban kerja yang perlu dipulihkan setelah terjadinya peristiwa degradasi, lakukan evaluasi terhadap runbook failover, desain ketersediaan tinggi, desain perbaikan otomatis, atau rencana pemulihan sumber daya HA. Identifikasi setiap tindakan yang mungkin dianggap sebagai tindakan bidang kontrol.

Pertimbangkan mengubah tindakan kontrol ke tindakan bidang data:

- Auto Scaling (bidang kontrol) menjadi sumber daya Amazon EC2 yang telah diskalakan sebelumnya (bidang data)
- Penskalaan instans Amazon EC2 (bidang kontrol) menjadi penskalaan AWS Lambda (bidang data)
- Nilai desain apa pun menggunakan Kubernetes dan sifat tindakan bidang kontrol. Menambahkan pod adalah tindakan bidang data di Kubernetes. Tindakan harus dibatasi ke penambahan pod dan bukan ke penambahan simpul. Menggunakan simpul yang disediakan secara berlebihan adalah metode yang lebih disukai untuk membatasi tindakan bidang kontrol

Pertimbangkan pendekatan alternatif yang memungkinkan tindakan-tindakan bidang data untuk memengaruhi remediasi yang sama.

- Perubahan Catatan Route 53 (bidang kontrol) atau Pengontrol Pemulihan Aplikasi Amazon (bidang data)

- Pemeriksaan kondisi Route 53 untuk pembaruan yang lebih otomatis

Pertimbangkan beberapa layanan di Wilayah sekunder, jika layanan sangat penting, untuk memungkinkan lebih banyak tindakan bidang kontrol dan bidang data di Wilayah yang tidak terdampak.

- Amazon EC2 Auto Scaling atau Amazon EKS di Wilayah primer dibandingkan dengan Amazon EC2 Auto Scaling atau Amazon EKS di Wilayah sekunder dan merutekan lalu lintas ke Wilayah sekunder (tindakan bidang kontrol)
- Membuat replika baca di primer sekunder atau mencoba tindakan yang sama di Wilayah primer (tindakan bidang kontrol)

Sumber daya

Praktik-praktik terbaik terkait:

- Definisi Ketersediaan
- REL11-BP01 Memantau semua komponen beban kerja untuk mendeteksi kegagalan

Dokumen terkait:

- Partner APN: partner yang dapat membantu dengan otomatisasi toleransi kesalahan Anda
- AWS Marketplace: produk yang dapat digunakan untuk toleransi kesalahan
- Amazon Builders' Library: Menghindari kelebihan beban dalam sistem terdistribusi dengan mengontrol layanan yang lebih kecil
- API Amazon DynamoDB (bidang kontrol dan bidang data)
- Eksekusi AWS Lambda (dibagi menjadi bidang kontrol dan bidang data)
- Bidang Data AWS Elemental MediaStore
- Membangun aplikasi yang sangat tangguh dengan menggunakan Pengontrol Pemulihan Aplikasi Amazon, Bagian 1: Tumpukan Wilayah Tunggal
- Membangun aplikasi yang sangat tangguh dengan menggunakan Pengontrol Pemulihan Aplikasi Amazon, Bagian 2: Tumpukan Multi-Wilayah
- Membuat Mekanisme Pemulihan Bencana Menggunakan Amazon Route 53
- Apa itu Pengontrol Pemulihan Aplikasi Amazon?

- Bidang kontrol dan bidang data Kubernetes

Video terkait:

- Kembali ke Dasar - Menggunakan Stabilitas Statis
- Membangun beban kerja multi-lokasi yang tangguh menggunakan layanan global AWS

Contoh terkait:

- Memperkenalkan Pengontrol Pemulihan Aplikasi Amazon
- Amazon Builders' Library: Menghindari kelebihan beban dalam sistem terdistribusi dengan mengontrol layanan yang lebih kecil
- Membangun aplikasi yang sangat tangguh dengan menggunakan Pengontrol Pemulihan Aplikasi Amazon, Bagian 1: Tumpukan Wilayah Tunggal
- Membangun aplikasi yang sangat tangguh dengan menggunakan Pengontrol Pemulihan Aplikasi Amazon, Bagian 2: Tumpukan Multi-Wilayah
- Stabilitas statis menggunakan Zona Ketersediaan

Alat terkait:

- Amazon CloudWatch
- AWS X-Ray

REL11-BP05 Menggunakan stabilitas statis untuk mencegah perilaku bimodal

Beban kerja harus stabil secara statis dan hanya beroperasi dalam mode normal tunggal. Perilaku bimodal adalah ketika beban kerja Anda menunjukkan perilaku yang berbeda dalam mode normal dan mode kegagalan.

Misalnya, Anda mungkin mencoba untuk melakukan pemulihan dari kegagalan Zona Ketersediaan dengan meluncurkan instans baru di Zona Ketersediaan yang berbeda. Hal ini dapat menyebabkan respons bimodal saat berada dalam mode kegagalan. Sebagai gantinya, Anda harus membangun beban kerja yang stabil secara statis dan beroperasi dalam satu mode saja. Dalam contoh ini, instans-instans tersebut seharusnya telah disediakan di Zona Ketersediaan kedua sebelum terjadinya

kegagalan. Desain stabilitas statis ini memastikan beban kerja hanya beroperasi dalam satu mode tunggal.

Hasil yang diinginkan: Beban kerja tidak menunjukkan perilaku bimodal selama mode normal dan mode kegagalan.

Anti-pola umum:

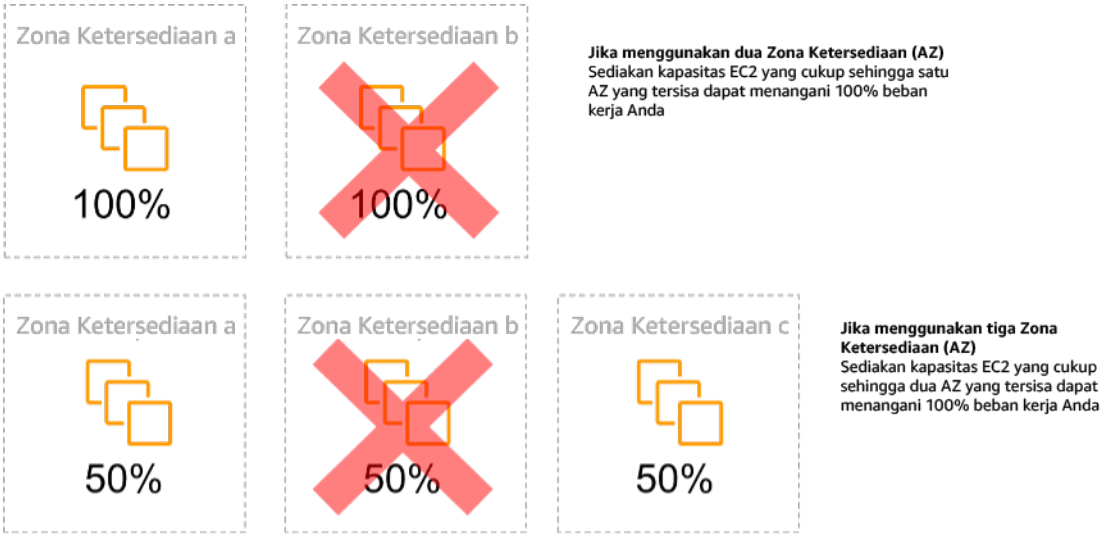
- Asumsikan bahwa sumber daya selalu dapat disediakan terlepas dari ruang lingkup kegagalannya.
- Mencoba memperoleh sumber daya secara dinamis selama terjadi kegagalan.
- Tidak menyediakan sumber daya yang memadai di seluruh zona atau Wilayah sampai terjadi kegagalan.
- Mempertimbangkan desain stabil statis hanya untuk sumber daya komputasi.

Manfaat menerapkan praktik terbaik ini: Beban kerja yang berjalan dengan desain yang stabil secara statis mampu memiliki hasil-hasil yang dapat diprediksi selama terjadi peristiwa normal dan kegagalan.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Sedang

Panduan implementasi

Perilaku bimodal terjadi ketika beban kerja Anda menunjukkan perilaku yang berbeda dalam mode normal dan mode gagal, (misalnya, mengandalkan peluncuran instans baru jika Zona Ketersediaan gagal). Contoh perilaku bimodal adalah ketika desain Amazon EC2 yang stabil menyediakan instans yang cukup di setiap Zona Ketersediaan untuk menangani beban dari beban kerja jika satu AZ disingkirkan. Penyeimbang Beban Elastis atau kesehatan Amazon Route 53 akan memeriksa untuk mengalihkan beban dari instans yang terganggu. Setelah lalu lintas dipindahkan, gunakan AWS Auto Scaling untuk mengganti instans secara tak selaras dari zona yang gagal dan luncurkan di zona sehat. Stabilitas statis untuk deployment komputasi (seperti kontainer atau instans EC2) akan menghasilkan keandalan yang paling tinggi.



Stabilitas statis instans EC2 di seluruh Zona Ketersediaan

Ini harus ditimbang berdasarkan biaya untuk model ini serta nilai bisnis untuk memelihara beban kerja berdasarkan semua kasus ketahanan. Menyediakan kapasitas komputasi yang lebih sedikit dan mengandalkan peluncuran instans baru apabila terjadi kegagalan memang lebih murah, tetapi untuk kegagalan-kegagalan berskala besar (seperti gangguan pada Zona Ketersediaan atau Regional), pendekatan ini kurang efektif karena bergantung pada bidang operasional serta sumber daya yang tersedia di zona atau Wilayah yang tidak terdampak.

Solusi Anda harus mengukur keandalan berdasarkan kebutuhan biaya untuk beban kerja Anda. Arsitektur stabilitas statis berlaku untuk berbagai arsitektur termasuk instans komputasi yang tersebar di Zona Ketersediaan, desain replika baca basis data, desain klaster Kubernetes (Amazon EKS), dan arsitektur failover multi-Wilayah.

Anda juga dapat menerapkan desain yang lebih stabil secara statis dengan menggunakan lebih banyak sumber daya di setiap zona. Dengan menggunakan lebih banyak zona, Anda mengurangi jumlah komputasi tambahan yang Anda perlukan untuk stabilitas statis.

Contoh perilaku bimodal adalah batas waktu jaringan yang dapat menyebabkan sebuah sistem untuk mencoba melakukan refresh status konfigurasi seluruh sistem. Ini akan menambahkan beban yang tak terduga ke komponen lain dan dapat menyebabkan komponen tersebut mengalami kegagalan, sehingga menimbulkan konsekuensi-konsekuensi lain yang tak diharapkan. Putaran umpan balik negatif ini memengaruhi ketersediaan beban kerja Anda. Sebagai gantinya, Anda dapat membangun sistem yang stabil secara statis dan beroperasi dalam satu mode saja. Desain yang stabil secara statis akan melakukan tugas yang konstan, dan selalu menyegarkan status konfigurasi dengan irama

yang teratur. Ketika sebuah panggilan gagal, beban kerja akan menggunakan nilai yang sebelumnya di-cache, dan memulai alarm.

Contoh perilaku bimodal lainnya adalah memperbolehkan klien untuk melewati cache beban kerja Anda ketika kegagalan terjadi. Ini mungkin terlihat seperti solusi yang mengakomodasi kebutuhan klien, tetapi hal ini secara signifikan dapat mengubah permintaan di beban kerja Anda dan kemungkinan akan mengakibatkan kegagalan.

Lakukan penilaian beban kerja kritis untuk menentukan beban kerja apa yang memerlukan jenis desain ketahanan ini. Untuk beban kerja yang dianggap kritis, setiap komponen aplikasi harus ditinjau. Contoh jenis layanan yang memerlukan evaluasi stabilitas statis adalah:

- Komputasi: Amazon EC2, EKS-EC2, ECS-EC2, EMR-EC2
- Basis Data: Amazon Redshift, Amazon RDS, Amazon Aurora
- Penyimpanan: Amazon S3 (Zona Tunggal), Amazon EFS (dudukan), Amazon FSx (dudukan)
- Penyeimbang beban: Menggunakan desain tertentu

Langkah-langkah implementasi

- Bangun sistem yang stabil secara statis dan hanya beroperasi dalam satu mode saja. Dalam hal ini, sediakan cukup instans di setiap Zona Ketersediaan atau Wilayah untuk menangani kapasitas beban kerja jika satu Zona Ketersediaan atau Wilayah dihapus. Berbagai layanan dapat digunakan untuk perutean ke sumber daya yang sehat, seperti:
  - Perutean DNS Lintas Wilayah
  - Perutean MRAP Amazon S3 MultiRegion
  - AWS Global Accelerator
  - Pengontrol Pemulihan Aplikasi Amazon
- Konfigurasikan replika baca basis data untuk memperhitungkan hilangnya satu instans utama atau replika baca. Jika lalu lintas dilayani oleh replika baca, maka kuantitas di setiap Zona Ketersediaan dan setiap Wilayah harus sama dengan kebutuhan keseluruhan jika terjadi kegagalan zona atau Wilayah.
- Konfigurasikan data kritis di dalam penyimpanan Amazon S3 yang dirancang agar stabil secara statis untuk data yang disimpan jika terjadi kegagalan Zona Ketersediaan. Jika kelas penyimpanan Amazon S3 One Zone-IA digunakan, ini tidak boleh dianggap stabil secara statis, karena hilangnya zona tersebut akan meminimalkan akses ke data yang disimpan.

- Penyeimbang beban terkadang dikonfigurasi secara salah atau secara bawaan untuk melayani satu Zona Ketersediaan tertentu. Dalam hal ini, desain yang stabil secara statis mungkin adalah menyebarkan beban kerja di beberapa AZ dalam desain yang lebih kompleks. Desain asli dapat digunakan untuk mengurangi lalu lintas antar zona karena alasan keamanan, latensi, atau biaya.

Sumber daya

Praktik terbaik Well-Architected terkait:

- Definisi Ketersediaan
- REL11-BP01 Memantau semua komponen beban kerja untuk mendeteksi kegagalan
- REL11-BP04 Mengandalkan bidang data dan bukan bidang kontrol selama pemulihan

Dokumen terkait:

- Meminimalkan Ketergantungan dalam Rencana Pemulihan Bencana
- Amazon Builders' Library: Stabilitas statis menggunakan Zona Ketersediaan
- Batas Isolasi Kesalahan
- Stabilitas statis menggunakan Zona Ketersediaan
- RDS Multi-Zona
- Meminimalkan Ketergantungan dalam Rencana Pemulihan Bencana
- Perutean DNS Lintas Wilayah
- Perutean MRAP Amazon S3 MultiRegion
- AWS Global Accelerator
- Pengontrol Pemulihan Aplikasi Amazon
- Amazon S3 Zona Tunggal
- Penyeimbangan Beban Lintas Zona

Video terkait:

- Stabilitas statis di AWS: AWS re:Invent 2019: Memperkenalkan Amazon Builders' Library (DOP328)

REL11-BP06 Mengirimkan notifikasi ketika peristiwa memengaruhi ketersediaan

Notifikasi dikirimkan setelah pelanggaran ambang batas terdeteksi, bahkan apabila peristiwa yang menyebabkan masalah tersebut sudah diatasi secara otomatis.

Pemulihan otomatis menjadikan beban kerja Anda andal. Namun demikian, kemampuan ini juga menyembunyikan masalah dasar yang perlu diatasi. Implementasikan pemantauan peristiwa yang baik agar Anda dapat mendeteksi setiap pola masalah, termasuk masalah-masalah yang ditangani oleh pemulihan otomatis, sehingga Anda dapat mengatasi akar penyebab masalahnya.

Sistem yang tangguh dirancang sedemikian rupa sehingga setiap terjadi peristiwa degradasi langsung dikomunikasikan kepada tim yang tepat. Notifikasi ini harus dikirim melalui satu atau banyak saluran komunikasi.

Hasil yang diinginkan: Pemberitahuan langsung dikirim ke tim operasi ketika ambang batas dilanggar, seperti tingkat kesalahan, latensi, atau metrik indikator performa utama (KPI) penting lainnya, sehingga masalah ini diselesaikan sesegera mungkin dan dampak terhadap pengguna dapat dicegah atau diminimalkan.

Anti-pola umum:

- Mengirimkan terlalu banyak alarm.
- Mengirimkan alarm yang tidak dapat ditindaklanjuti.
- Mengatur ambang alarm terlalu tinggi (terlalu sensitif) atau terlalu rendah (kurang sensitif).
- Tidak mengirimkan alarm untuk dependensi eksternal.
- Tidak mempertimbangkan kegagalan abu-abu saat merancang pemantauan dan alarm.
- Melakukan otomatisasi pemulihan, tetapi tidak memberikan notifikasi kepada tim yang tepat bahwa pemulihan diperlukan.

Manfaat menerapkan praktik terbaik ini: Notifikasi pemulihan membuat tim operasional dan bisnis menyadari adanya degradasi layanan sehingga mereka dapat segera bereaksi untuk meminimalkan waktu deteksi rata-rata (MTTD) dan waktu perbaikan rata-rata (MTTR). Notifikasi peristiwa pemulihan juga menjamin bahwa Anda tidak mengabaikan masalah yang jarang terjadi.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Sedang. Kegagalan mengimplementasikan mekanisme pemantauan dan notifikasi peristiwa secara tepat dapat

mengakibatkan terjadinya kegagalan dalam mendeteksi pola masalah, termasuk masalah yang ditangani oleh pemulihan otomatis. Sebuah tim hanya akan menyadari adanya degradasi sistem ketika pengguna menghubungi layanan pelanggan atau secara kebetulan.

Panduan implementasi

Saat menetapkan strategi pemantauan, alarm yang dipicu adalah sebuah peristiwa umum. Peristiwa ini kemungkinan berisi pengidentifikasi untuk alarm, status alarm (seperti `IN ALARM` dan `OK`), dan detail tentang apa yang memicunya. Dalam banyak kasus, sebuah peristiwa alarm seharusnya dideteksi dan email notifikasi dikirimkan. Ini adalah contoh tindakan pada alarm. Notifikasi alarm sangat penting dalam hal observabilitas karena notifikasi ini memberi tahu orang yang tepat bahwa ada masalah. Namun demikian, ketika tindakan terhadap peristiwa sudah matang di dalam solusi observabilitas Anda, tindakan tersebut dapat secara otomatis memperbaiki masalah tanpa memerlukan campur tangan manusia.

Setelah alarm pemantauan KPI ditetapkan, peringatan seharusnya dikirimkan ke tim yang tepat ketika ambang batas terlampaui. Peringatan tersebut juga dapat digunakan untuk memicu proses otomatis yang akan mencoba memperbaiki degradasi.

Untuk pemantauan ambang batas yang lebih kompleks, alarm gabungan harus dipertimbangkan. Alarm gabungan menggunakan sejumlah alarm pemantauan KPI untuk membuat peringatan berdasarkan logika bisnis operasional. Alarm CloudWatch dapat dikonfigurasi untuk mengirimkan email, atau untuk mencatatkan log insiden di dalam sistem pelacakan insiden pihak ketiga menggunakan integrasi Amazon SNS atau Amazon EventBridge.

Langkah-langkah implementasi

Buat berbagai jenis alarm berdasarkan cara yang digunakan untuk memantau beban kerja, seperti:

- Alarm aplikasi digunakan untuk mendeteksi ketika ada bagian dari beban kerja Anda yang tidak berfungsi dengan baik.
- Alarm infrastruktur menunjukkan kapan Anda harus menskalakan sumber daya. Alarm dapat ditampilkan secara visual di dasbor, mengirimkan peringatan melalui Amazon SNS atau email, dan menggunakan Penskalaan Otomatis untuk menaikkan atau menurunkan skala sumber daya beban kerja.
- Alarm statis dapat dibuat untuk memantau ketika sebuah metrik melanggar ambang batas statis selama periode evaluasi tertentu.
- Alarm gabungan dapat memperhitungkan alarm-alarm kompleks dari berbagai sumber.

- Setelah alarm dibuat, buatlas peristiwa-peristiwa notifikasi yang sesuai. Anda dapat langsung menginvokasi sebuah Amazon SNS API untuk mengirim notifikasi dan menautkan otomatisasi apa pun untuk remediasi atau komunikasi.
- Terus dapatkan informasi tentang penurunan layanan terkait AWS Health. Buat notifikasi peristiwa AWS Health sesuai keperluan yang dikirim ke saluran email dan obrolan melalui Notifikasi Pengguna AWS serta integrasikan secara programatis dengan alat pemantauan dan peringatan Anda melalui Amazon EventBridge.

Sumber daya

Praktik terbaik Well-Architected terkait:

- Definisi Ketersediaan

Dokumen terkait:

- Membuat Alarm Berdasarkan Ambang Batas Statis
- Apa itu Amazon EventBridge?
- Apa itu Amazon Simple Notification Service?
- Memublikasikan Metrik Kustom
- Menggunakan Alarm Amazon CloudWatch
- Menyiapkan alarm CloudWatch Composite
- Yang baru di Observabilitas AWS di re:Invent 2022

Alat terkait:

- CloudWatch
- X-Ray CloudWatch

REL11-BP07 Merancang produk Anda agar memenuhi target-target ketersediaan dan perjanjian tingkat layanan (SLA) waktu aktif

Rancang produk Anda agar memenuhi target-target ketersediaan dan perjanjian tingkat layanan (SLA) waktu aktif. Jika Anda memublikasi atau secara pribadi menyetujui target-target ketersediaan

atau SLA yang berlaku, verifikasikan bahwa proses operasional dan arsitektur Anda didesain untuk mendukungnya.

Hasil yang diinginkan: Setiap aplikasi memiliki target yang ditentukan untuk ketersediaan dan SLA untuk metrik kinerja, yang dapat Anda pantau dan pelihara untuk memenuhi hasil bisnis.

Anti-pola umum:

- Mendesain dan melakukan deployment beban kerja tanpa menetapkan SLA apa pun.
- Metrik SLA ditetapkan ke tingkat tinggi tanpa rasional atau persyaratan bisnis.
- Menetapkan SLA tanpa memperhitungkan dependensi dan SLA yang mendasarinya.
- Desain aplikasi dibuat tanpa mempertimbangkan Model Tanggung Jawab Bersama untuk Ketangguhan.

Manfaat menerapkan praktik terbaik ini: Merancang aplikasi berdasarkan target ketahanan utama akan membantu Anda dalam memenuhi tujuan bisnis dan harapan pelanggan. Tujuan-tujuan ini membantu mendorong proses desain aplikasi yang mengevaluasi berbagai macam teknologi dan mempertimbangkan beragam kompromi.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Sedang

Panduan implementasi

Desain aplikasi harus memperhitungkan berbagai rangkaian persyaratan yang didapatkan dari tujuan bisnis, operasional, dan finansial. Dalam persyaratan operasional, beban kerja harus memiliki target-target metrik ketangguhan tertentu sehingga dapat dipantau dan dukung dengan sesuai. Metrik-metrik ketangguhan tidak boleh ditetapkan atau didapatkan setelah melakukan deployment beban kerja. Metrik-metrik ketangguhan tersebut harus ditetapkan selama fase desain dan membantu memandu berbagai keputusan dan kompromi.

- Setiap beban kerja harus memiliki rangkaian metrik-metrik ketangguhannya sendiri. Metrik-metrik tersebut mungkin berbeda dari aplikasi bisnis yang lain.
- Mengurangi dependensi dapat memiliki dampak positif pada ketersediaan. Setiap beban kerja harus mempertimbangkan dependensinya serta SLA-nya. Secara umum, pilihlah dependensi dengan target ketersediaan yang setara dengan atau lebih besar dari target beban kerja Anda.
- Pertimbangkan desain yang mengunakan penggabungan longgar sehingga beban kerja Anda dapat beroperasi dengan benar meskipun ada gangguan dependensi, apabila mungkin.

- Kurangi dependensi bidang kontrol, terutama selama pemulihan atau degradasi. Evaluasi desain yang secara statis stabil untuk beban kerja yang penting bagi misi. Gunakan penghematan sumber daya untuk meningkatkan ketersediaan dependensi tersebut di sebuah beban kerja.
- Observabilitas dan instrumentasi sangat penting untuk mencapai SLA dengan mengurangi Waktu Rata-Rata ke Deteksi (MTTD) dan Waktu Rata-Rata ke Perbaikan (MTTR).
- Kegagalan lebih jarang (MTBF lebih lama), waktu deteksi kegagalan lebih pendek (MTTD lebih singkat), dan waktu perbaikan lebih singkat (MTTR lebih singkat) adalah tiga faktor yang digunakan untuk meningkatkan ketersediaan di sistem terdistribusi.
- Menetapkan dan memenuhi metrik-metrik ketangguhan untuk beban kerja merupakan fondasi dari desain yang efektif. Desain tersebut harus memperhitungkan berbagai kompromi terkait kompleksitas desain, dependensi layanan, performa, penskalaan, dan biaya.

Langkah-langkah implementasi

- Tinjau dan dokumentasikan desain beban kerja sambil mempertimbangkan pertanyaan-pertanyaan berikut:
  - Di mana bidang kontrol digunakan di beban kerja?
  - Bagaimana beban kerja mengimplementasikan toleransi kesalahan?
  - Apa saja pola desain untuk penskalaan, penskalaan otomatis, redundansi, dan komponen dengan ketersediaan tinggi?
  - Apa saja persyaratan untuk ketersediaan dan konsistensi data?
  - Apakah ada pertimbangan untuk penghematan sumber daya atau stabilitas statis sumber daya?
  - Apa saja dependensi layanan?
- Tetapkan metrik-metrik SLA berdasarkan arsitektur beban kerja sambil bekerja sama dengan para pemangku kepentingan. Pertimbangkan SLA semua dependensi yang digunakan oleh beban kerja.
- Setelah target SLA ditetapkan, optimalkan arsitektur untuk memenuhi SLA.
- Setelah desain yang akan memenuhi SLA dibuat, implementasikan perubahan operasional, otomatisasi proses, dan runbook yang juga akan memiliki fokus pada pengurangan MTTD dan MTTR.
- Setelah di-deploy, pantau dan buat laporan SLA.

Sumber daya

Praktik-praktik terbaik terkait:

- REL03-BP01 Pilih cara mengelompokkan beban kerja Anda
- REL10-BP01 Melakukan deployment beban kerja ke beberapa lokasi
- REL11-BP01 Memantau semua komponen beban kerja untuk mendeteksi kegagalan
- REL11-BP03 Melakukan otomatisasi pemulihan di semua lapisan
- REL12-BP04 Menguji ketahanan menggunakan chaos engineering
- REL13-BP01 Menetapkan sasaran pemulihan untuk waktu henti dan kehilangan data
- Memahami kesehatan beban kerja

Dokumen terkait:

- Ketersediaan dengan redundansi
- Pilar keandalan - Ketersediaan
- Mengukur ketersediaan
- Batas Isolasi Kesalahan AWS
- Model Tanggung Jawab Bersama untuk Ketangguhan
- Stabilitas statis menggunakan Zona Ketersediaan
- Perjanjian Tingkat Layanan (SLA) AWS
- Panduan untuk Arsitektur Berbasis Sel pada AWS
- Infrastruktur AWS
- Laporan resmi Pola Ketangguhan Multi-AZ Lanjutan

Layanan terkait:

- Amazon CloudWatch
- AWS Config
- AWS Trusted Advisor

Uji keandalan

Setelah Anda mendesain beban kerja Anda agar tangguh terhadap tekanan produksi, pengujian adalah satu-satunya cara untuk memastikan beban kerja akan beroperasi sesuai desain, dan memberikan ketangguhan yang Anda harapkan.

Lakukan pengujian untuk memvalidasi bahwa beban kerja Anda memenuhi persyaratan fungsional dan non-fungsional, karena bug atau bottleneck kinerja dapat memengaruhi keandalan beban kerja Anda. Uji ketangguhan beban kerja Anda untuk membantu Anda menemukan bug tersembunyi yang hanya muncul dalam produksi. Lakukan pengujian ini secara rutin.

Praktik terbaik

- REL12-BP01 Menggunakan playbook untuk menyelidiki kegagalan
- REL12-BP02 Menjalankan analisis setelah insiden
- REL12-BP03 Menguji persyaratan skalabilitas dan kinerja
- REL12-BP04 Menguji ketahanan menggunakan chaos engineering
- REL12-BP05 Mengadakan game day secara rutin

REL12-BP01 Menggunakan playbook untuk menyelidiki kegagalan

Dokumentasikan proses penyelidikan di dalam playbook agar dapat memberikan respons yang cepat dan konsisten terhadap skenario kegagalan yang tidak benar-benar dipahami. Playbook adalah langkah-langkah yang telah ditetapkan di awal untuk mengidentifikasi faktor yang menyebabkan skenario kegagalan. Hasil dari setiap langkah proses digunakan untuk menentukan langkah berikutnya yang harus diambil sampai masalah teridentifikasi atau dieskalasi.

Palybook ini adalah perencanaan proaktif yang harus Anda lakukan, agar Anda dapat mengambil tindakan reaktif secara efektif. Ketika skenario kegagalan yang tidak tercakup dalam playbook dialami di lingkungan produksi, tangani masalah terlebih dahulu (put our the fire). Lalu lihat kembali langkah-langkah yang telah Anda ambil untuk mengatasi masalah tersebut dan gunakan itu semua untuk menambahkan entri baru dalam playbook.

Perhatikan, playbook digunakan untuk merespons insiden tertentu, sedangkan runbook digunakan untuk mencapai hasil tertentu. Sering kali, runbook digunakan untuk untuk aktivitas rutin, dan playbook digunakan untuk merespons peristiwa non-rutin.

Anti-pola umum:

- Berencana untuk melakukan deployment beban kerja tanpa mengetahui proses untuk mendiagnosis masalah atau merespons insiden.
- Keputusan yang tidak direncanakan tentang sistem mana saja yang dikumpulkan log dan metriknya saat menyelidiki peristiwa.
- Tidak mempertahankan metrik dan peristiwa cukup lama agar dapat mengambil data.

Manfaat menjalankan praktik terbaik ini: Merekam playbook akan memastikan bahwa proses dapat diikuti secara konsisten. Melakukan kodifikasi pada playbook dapat membatasi munculnya kesalahan dari aktivitas manual. Melakukan otomatisasi pada playbook dapat menghemat waktu respons peristiwa dengan menghilangkan keharusan campur tangan anggota tim atau memberikan informasi tambahan ketika campur tangan mereka dimulai.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Tinggi

Panduan implementasi

- Gunakan playbook untuk mengidentifikasi masalah. Playbook adalah proses-proses yang didokumentasikan untuk menyelidiki masalah. Dokumentasikan proses penyelidikan di playbook agar dapat memberikan respons yang cepat dan konsisten terhadap skenario kegagalan. Playbook harus memuat informasi dan panduan yang dapat digunakan oleh orang yang cukup terampil untuk mengumpulkan informasi, mengidentifikasi potensi sumber kegagalan, mengisolasi kesalahan, dan menentukan faktor penyebabnya (lakukan analisis pasca-insiden).
  - Implementasikan playbook sebagai kode. Jalankan operasi sebagai kode dengan membuat skrip playbook Anda untuk memastikan konsistensi dan mengurangi kesalahan yang disebabkan proses manual. Playbook dapat terdiri dari beberapa skrip sesuai dengan banyaknya langkah yang diperlukan untuk mengidentifikasi faktor-faktor penyebab masalah. Aktivitas runbook dapat diinvokasi atau dijalankan sebagai bagian dari aktivitas playbook, atau mempercepat eksekusi playbook untuk merespons peristiwa yang teridentifikasi.
    - Lakukan otomatisasi pada playbook operasional Anda dengan Systems Manager AWS
    - Run Command Systems Manager AWS
    - AWS Systems Manager Automation
    - Apa itu AWS Lambda?
    - Apa itu Amazon EventBridge?
    - Menggunakan Alarm Amazon CloudWatch

Sumber daya

Dokumen terkait:

- AWS Systems Manager Automation
- Run Command Systems Manager AWS
- Lakukan otomatisasi pada playbook operasional Anda dengan Systems Manager AWS

- Menggunakan Alarm Amazon CloudWatch
- Menggunakan Canary (Amazon CloudWatch Synthetics)
- Apa itu Amazon EventBridge?
- Apa itu AWS Lambda?

Contoh terkait:

- Melakukan otomatisasi operasi dengan Playbook dan Runbook

REL12-BP02 Menjalankan analisis setelah insiden

Lakukan peninjauan terhadap peristiwa-peristiwa yang memengaruhi pelanggan, dan identifikasi faktor yang berkontribusi serta tindakan pencegahannya. Gunakan informasi ini untuk mengembangkan langkah-langkah mitigasi untuk meminimalkan atau mencegah kemungkinan terjadi lagi. Kembangkan prosedur untuk respons efektif dan cepat. Komunikasikan faktor-faktor yang berkontribusi dan tindakan-tindakan korektif yang diperlukan, yang disesuaikan dengan audiens target. Buatlah sebuah metode untuk mengomunikasikan penyebab ini ke pihak-pihak lain sesuai keperluan.

Lakukan penilaian untuk mengidentifikasi alasan mengapa pengujian yang ada tidak dapat menemukan masalahnya. Menambahkan pengujian untuk kasus ini jika pengujian belum ada.

Hasil yang diinginkan: Tim Anda memiliki pendekatan yang konsisten dan disepakati untuk menangani analisis pasca-insiden. Salah satu mekanismenya adalah proses koreksi kesalahan (COE). Proses COE akan membantu tim Anda untuk mengidentifikasi, memahami, dan mengatasi akar penyebab insiden, sekaligus membangun mekanisme dan pagar pembatas untuk membatasi kemungkinan insiden yang sama terjadi lagi.

Anti-pola umum:

- Menemukan temuan tentang faktor-faktor yang berkontribusi, tetapi tidak terus-menerus mencari lebih dalam berusaha mencari masalah potensial dan pendekatan lainnya untuk memitigasi.
- Hanya mengidentifikasi penyebab kesalahan manusia, dan tidak memberikan pelatihan atau otomatisasi apa pun yang dapat mencegah kesalahan manusia.
- Fokus menyalahkan, bukan memahami akar penyebabnya, sehingga tercipta budaya ketakutan dan menghambat komunikasi yang terbuka

- Tidak berbagi wawasan, yang membuat temuan-temuan analisis insiden hanya diketahui kelompok kecil saja, sehingga orang lain tidak dapat belajar dari pengalaman tersebut
- Tidak ada mekanisme yang digunakan untuk merekam pengetahuan institusional, sehingga wawasan yang berharga hilang karena pelajaran yang didapat tidak dipelihara dalam bentuk praktik terbaik yang diperbarui secara berkelanjutan dan mengakibatkan insiden berulang dengan akar penyebab yang sama atau serupa

Manfaat menjalankan praktik terbaik ini: Dengan melakukan analisis setelah insiden dan membagikan hasilnya, beban kerja lain akan dapat memitigasi risiko jika beban kerja sudah mengimplementasikan faktor yang berkontribusi yang sama, sehingga mitigasi atau pemulihan otomatis dapat diimplementasikan sebelum insiden terjadi.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Tinggi

Panduan implementasi

Analisis setelah insiden yang baik memberikan peluang untuk mengusulkan solusi-solusi umum terhadap masalah dengan pola arsitektur yang digunakan di tempat lainnya dalam sistem.

Dokumentasi dan penanganan masalah merupakan landasan proses COE. Sebaiknya tentukan cara standar untuk mendokumentasikan akar penyebab masalah kritis, dan memastikan penyebab tersebut ditinjau dan ditangani. Tetapkan kepemilikan yang jelas untuk proses analisis setelah insiden. Tunjuk individu atau tim penanggung jawab yang akan mengawasi penyelidikan dan tindak lanjut insiden.

Dorong budaya yang berfokus pada pembelajaran dan peningkatan, bukan menyalahkan. Tekankan bahwa tujuannya adalah untuk mencegah insiden di kemudian hari, bukan untuk menghukum individu.

Kembangkan prosedur yang jelas untuk melakukan analisis setelah insiden. Prosedur ini harus menguraikan langkah-langkah yang harus diambil, informasi yang akan dikumpulkan, dan pertanyaan-pertanyaan penting yang harus dicari jawabannya saat melakukan analisis. Selidiki insiden secara menyeluruh, tidak hanya pada penyebab langsung guna mengidentifikasi akar penyebab masalah dan faktor penyumbangnya. Gunakan teknik-teknik seperti lima alasan untuk memahami lebih dalam masalah-masalah mendasar.

Buatlah repositori pelajaran yang didapat dari analisis insiden. Pengetahuan institusional ini dapat digunakan sebagai referensi untuk insiden dan upaya pencegahan ke depannya. Bagikan temuan dan wawasan dari analisis yang dilakukan setelah insiden, dan pertimbangkan untuk mengadakan

pertemuan peninjauan pasca insiden yang terbuka untuk semua (open-invite) untuk membahas pelajaran yang didapatkan.

Langkah-langkah implementasi

- Saat melakukan analisis pasca-insiden, pastikan untuk tidak menyalahkan siapa pun dalam proses tersebut. Dengan begitu, orang-orang yang terlibat dalam insiden tersebut bersikap rasional terhadap tindakan korektif yang diusulkan dan mendorong penilaian mandiri yang jujur serta kolaborasi di seluruh tim.
- Tentukan cara terstandardisasi untuk mendokumentasikan masalah-masalah kritis. Contoh struktur untuk dokumen tersebut adalah sebagai berikut:
  - Apa yang terjadi?
  - Apa dampaknya terhadap para pelanggan dan bisnis Anda?
  - Apa akar penyebabnya?
  - Data apa yang Anda miliki untuk mendukung hal ini?
    - Misalnya, metrik dan grafik
  - Apa implikasi pilar kritis, terutama keamanan?
    - Saat merancang beban kerja, Anda memilah pilar-pilar sesuai dengan konteks bisnis Anda. Keputusan bisnis ini dapat mendorong prioritas rekayasa Anda. Anda dapat melakukan optimalisasi untuk mengurangi biaya dengan mengorbankan keandalan dalam lingkungan pengembangan, atau, untuk solusi yang sangat penting, Anda dapat melakukan optimalisasi keandalan dengan biaya yang lebih tinggi. Keamanan selalu menjadi hal yang didahulukan dan diutamakan, karena Anda harus melindungi pelanggan Anda.
  - Pelajaran apa hal yang Anda dapatkan?
  - Tindakan-tindakan korektif apa yang Anda ambil?
    - Item tindakan
    - Item terkait
- Buat prosedur-prosed operasi terstandardisasi yang jelas untuk melakukan analisis pasca insiden.
- Siapkan proses pelaporan insiden terstandardisasi. Dokumentasikan semua insiden secara komprehensif, termasuk laporan insiden awal, log, komunikasi, dan tindakan-tindakan yang diambil saat insiden berlangsung.
- Ingatlah bahwa sebuah insiden tidak harus berupa terhentinya sistem (outage). Insiden juga bisa berupa kejadian yang hampir menyebabkan henti sistem (near-miss), atau performa sistem yang tidak sesuai harapan meski tetap memenuhi fungsi bisnisnya.

- Terus tingkatkan proses analisis pasca insiden Anda berdasarkan umpan balik dan pelajaran yang dipetik.
- Rekam temuan-temuan utama dalam sistem manajemen pengetahuan, dan pertimbangkan pola apa pun yang perlu ditambahkan ke dalam panduan developer atau daftar periksa sebelum deployment.

Sumber daya

Dokumen terkait:

- Mengapa Anda harus mengembangkan koreksi kesalahan (COE)

Video terkait:

- Pendekatan Amazon untuk gagal dengan sukses
- AWS re:Invent 2021 - Amazon Builders' Library: Keunggulan Operasional di Amazon

REL12-BP03 Menguji persyaratan skalabilitas dan kinerja

Gunakan teknik-teknik seperti pengujian beban untuk memvalidasi bahwa beban kerja memenuhi persyaratan kinerja dan penskalaan.

Di dalam cloud, Anda dapat membuat sebuah lingkungan pengujian dalam skala produksi untuk beban kerja Anda sesuai permintaan. Alih-alih mengandalkan lingkungan pengujian dengan skala yang diturunkan, sehingga dapat membuat prediksi perilaku produksi tidak akurat, Anda dapat menggunakan cloud untuk menyediakan lingkungan pengujian yang sangat mirip dengan lingkungan produksi yang Anda harapkan. Lingkungan ini membantu Anda menguji dengan simulasi yang lebih akurat dari kondisi nyata yang dihadapi aplikasi Anda.

Selain upaya pengujian performa, penting juga untuk memvalidasi bahwa sumber daya dasar, pengaturan penskalaan, kuota layanan, dan desain ketahanan Anda beroperasi sebagaimana mestinya saat menerima beban. Pendekatan holistik ini memverifikasi bahwa penskalaaan dan performa aplikasi bisa diandalkan sesuai kebutuhan, bahkan dalam kondisi yang paling berat.

Hasil yang diinginkan: Beban kerja Anda tetap memiliki perilaku yang diharapkan bahkan saat mengalami beban puncak. Anda secara proaktif mengatasi masalah terkait performa yang mungkin timbul seiring aplikasi bertumbuh dan berkembang.

Anti-pola umum:

- Anda menggunakan lingkungan pengujian yang tidak mencerminkan lingkungan produksi.
- Anda memperlakukan pengujian beban sebagai aktivitas satu kali yang terpisah, bukan bagian terintegrasi dari pipeline integrasi berkelanjutan (CI) deployment.
- Anda tidak menentukan persyaratan performa yang jelas dan terukur, seperti target waktu respons, throughput, dan skalabilitas.
- Anda melakukan pengujian dengan skenario beban yang tidak realistis atau tidak memadai, dan Anda gagal menguji beban puncak, lonjakan mendadak, dan beban tinggi yang berkelanjutan.
- Anda tidak melakukan pengujian tekanan pada beban kerja hingga melebihi batas beban yang diharapkan.
- Anda menggunakan alat pengujian beban dan pembuatan profil performa yang tidak memadai atau tidak tepat.
- Anda tidak memiliki sistem pemantauan dan peringatan yang komprehensif untuk melacak metrik performa dan mendeteksi anomali.

Manfaat menjalankan praktik terbaik ini:

- Pengujian beban membantu Anda mengidentifikasi potensi hambatan performa dalam sistem Anda sebelum masalah ini masuk ke produksi. Saat Anda menyimulasikan lalu lintas dan beban kerja tingkat produksi, Anda dapat mengidentifikasi di area mana sistem Anda mungkin kesulitan menangani beban, seperti waktu respons yang lambat, keterbatasan sumber daya, atau kegagalan sistem.
- Dengan menguji sistem Anda dalam berbagai kondisi beban, Anda dapat lebih memahami persyaratan sumber daya yang diperlukan untuk mendukung beban kerja Anda. Informasi ini dapat membantu Anda membuat keputusan yang matang untuk alokasi sumber daya dan mencegah penyediaan sumber daya yang berlebih atau kurang.
- Untuk mengidentifikasi potensi titik kegagalan, Anda dapat mengamati performa beban kerja Anda dalam kondisi beban tinggi. Informasi ini membantu Anda meningkatkan keandalan dan ketahanan beban kerja Anda dengan menerapkan mekanisme toleransi kesalahan, strategi failover, dan tindakan redundansi yang sesuai.
- Anda mengidentifikasi dan mengatasi masalah performa lebih dini, sehingga Anda dapat terhindari dari konsekuensi yang merugikan seperti pemadaman sistem, waktu respons yang lambat, dan ketidakpuasan pengguna.

- Data performa mendetail dan informasi pembuatan profil yang dikumpulkan selama pengujian dapat membantu Anda memecahkan masalah performa yang mungkin muncul dalam produksi. Hal ini dapat menghasilkan respons dan penyelesaian insiden yang lebih cepat, sehingga mengurangi dampak bagi pengguna dan operasi organisasi Anda.
- Dalam industri tertentu, pengujian performa proaktif dapat membantu beban kerja Anda memenuhi standar kepatuhan, sehingga mengurangi risiko denda atau masalah hukum.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Tinggi

Panduan implementasi

Langkah pertama adalah menentukan strategi pengujian komprehensif yang mencakup semua aspek penskalaan dan persyaratan performa. Untuk memulai, tentukan dengan jelas sasaran tingkat layanan (SLO) beban kerja Anda berdasarkan kebutuhan bisnis Anda, seperti throughput, histogram latensi, dan tingkat kesalahan. Selanjutnya, rancang serangkaian pengujian yang dapat menyimulasikan berbagai skenario beban yang berkisar dari penggunaan rata-rata hingga lonjakan mendadak dan beban puncak berkelanjutan, dan verifikasi bahwa perilaku beban kerja memenuhi SLO Anda. Pengujian ini harus diotomatisasi dan diintegrasikan ke dalam pipeline integrasi dan deployment berkelanjutan Anda untuk menemukan regresi performa lebih dini dalam proses pengembangan.

Untuk menguji penskalaan dan performa secara efektif, berinvestasilah pada alat dan infrastruktur yang tepat. Hal ini termasuk alat pengujian beban yang dapat menghasilkan lalu lintas pengguna yang realistis, alat profil performa untuk mengidentifikasi hambatan, dan solusi pemantauan untuk melacak metrik utama. Yang penting, Anda harus memverifikasi bahwa lingkungan pengujian Anda sangat mirip dengan lingkungan produksi dalam hal infrastruktur dan kondisi lingkungan untuk mendapatkan hasil pengujian seakurat mungkin. Untuk mempermudah mereplikasi dan menskalakan pengaturan mirip produksi dengan andal, gunakan infrastruktur sebagai kode dan aplikasi berbasis kontainer.

Pengujian penskalaan dan performa adalah proses yang berkelanjutan, tidak hanya dilakukan satu kali saja. Implementasikan pemantauan dan peringatan komprehensif untuk melacak performa aplikasi dalam produksi, serta gunakan data ini untuk terus menyempurnakan strategi pengujian dan upaya pengoptimalan Anda. Analisis data performa secara rutin untuk mengidentifikasi masalah yang muncul, menguji strategi penskalaan baru, dan mengimplementasikan pengoptimalan untuk meningkatkan efisiensi dan keandalan aplikasi. Ketika Anda mengadopsi pendekatan iteratif dan terus belajar dari data produksi, Anda bisa memverifikasi bahwa aplikasi Anda dapat beradaptasi

dengan permintaan pengguna yang bervariasi dan menjaga ketahanan serta performa yang optimal dari waktu ke waktu.

Langkah-langkah implementasi

1. Tentukan persyaratan performa yang jelas dan terukur, seperti target waktu respons, throughput, dan skalabilitas. Persyaratan ini harus didasarkan pada pola penggunaan beban kerja, ekspektasi pengguna, dan kebutuhan bisnis Anda.
2. Pilih dan konfigurasikan alat pengujian beban yang dapat secara akurat meniru pola beban dan perilaku pengguna di lingkungan produksi Anda.
3. Siapkan lingkungan pengujian yang sangat mirip dengan lingkungan produksi, termasuk kondisi infrastruktur dan lingkungan, agar hasil pengujian bisa lebih akurat.
4. Buat rangkaian pengujian yang mencakup berbagai skenario, mulai dari pola penggunaan rata-rata hingga beban puncak, lonjakan mendadak, dan beban tinggi yang berkelanjutan. Integrasikan pengujian ke dalam pipeline integrasi dan deployment berkelanjutan Anda untuk menemukan regresi performa lebih dini dalam proses pengembangan.
5. Lakukan pengujian beban untuk menyimulasikan lalu lintas pengguna di dunia nyata dan memahami bagaimana aplikasi Anda berperilaku dalam berbagai kondisi beban. Untuk melakukan pengujian tekanan terhadap aplikasi Anda, lampaui beban yang diharapkan dan amati perilakunya, seperti penurunan waktu respons, kehabisan sumber daya, atau kegagalan sistem, sehingga membantu mengidentifikasi titik puncak aplikasi Anda dan menentukan strategi penskalaan Anda. Evaluasi skalabilitas beban kerja Anda dengan meningkatkan beban secara bertahap, dan ukur dampak performa untuk mengidentifikasi batas penskalaan serta merencanakan kebutuhan kapasitas ke depannya.
6. Implementasikan pemantauan dan peringatan komprehensif untuk melacak metrik performa, mendeteksi anomali, dan memulai tindakan penskalaan atau pemberitahuan ketika ambang batas terlampaui.
7. Terus pantau dan analisis data performa untuk mengidentifikasi area yang dapat ditingkatkan lebih lanjut. Lakukan iterasi terhadap strategi pengujian dan upaya pengoptimalan Anda.

Sumber daya

Praktik-praktik terbaik terkait:

- REL01-BP04 Memantau dan mengelola kuota
- REL06-BP01 Memantau semua komponen untuk beban kerja (Pembuatan)

- REL06-BP03 Mengirimkan notifikasi (Pemrosesan dan pembuatan alarm waktu nyata)

Dokumen terkait:

- Aplikasi pengujian beban
- Pengujian Beban Terdistribusi di AWS
- Pemantauan Performa Aplikasi
- Kebijakan Pengujian Amazon EC2

Contoh terkait:

- Pengujian Beban Terdistribusi di AWS (GitHub)

Alat terkait:

- Amazon CodeGuru Profiler
- Amazon CloudWatch RUM
- Apache JMeter
- K6
- Vegeta
- Hey
- ab
- wrk
- Pengujian Beban Terdistribusi di AWS

REL12-BP04 Menguji ketahanan menggunakan chaos engineering

Jalankan eksperimen chaos secara rutin di lingkungan yang berada dalam atau sedekat mungkin dengan produksi untuk memahami bagaimana sistem Anda merespons kondisi yang merugikan.

Hasil yang diinginkan:

Ketahanan beban kerja diverifikasi secara rutin dengan menerapkan chaos engineering dalam bentuk eksperimen injeksi kesalahan atau injeksi beban tak terduga. Selain itu, terdapat pengujian ketahanan yang memvalidasi perilaku yang diharapkan yang diketahui dari beban kerja Anda selama

berlangsungnya sebuah peristiwa. Gabungkan chaos engineering dan pengujian ketahanan agar Anda meyakini bahwa beban kerja dapat bertahan dari kegagalan komponen dan dapat pulih dari gangguan tak terduga dengan dampak minimal hingga tanpa dampak.

Anti-pola umum:

- Menentukan desain untuk mendapatkan ketahanan, tetapi tidak memastikan bagaimana fungsi beban kerja secara keseluruhan saat terjadi kesalahan.
- Tidak pernah bereksperimen dalam kondisi dunia nyata dan dengan beban yang diharapkan.
- Tidak memperlakukan eksperimen Anda sebagai kode atau memeliharanya melalui siklus pengembangan.
- Tidak menjalankan eksperimen chaos baik sebagai bagian dari pipeline CI/CD Anda maupun di luar deployment.
- Tidak menggunakan analisis pasca-insiden terdahulu saat menentukan kesalahan mana yang akan digunakan dalam eksperimen.

Manfaat menjalankan praktik terbaik ini: Menginjeksi kesalahan untuk memverifikasi ketahanan beban kerja Anda akan membuat Anda percaya bahwa prosedur pemulihan yang dimiliki oleh desain Anda yang tangguh akan berfungsi efektif jika terjadi kesalahan nyata.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Sedang

Panduan implementasi

Chaos engineering memberi tim Anda kemampuan untuk terus menginjeksi gangguan (simulasi) dunia nyata dengan cara yang terkontrol di tingkat penyedia layanan, infrastruktur, beban kerja, dan komponen, dengan dampak minimal atau tanpa dampak terhadap pelanggan Anda. Hal ini akan memungkinkan tim Anda belajar dari kesalahan serta mengamati, mengukur, dan meningkatkan ketahanan beban kerja Anda, serta memvalidasi bahwa peringatan akan diluncurkan dan tim mendapatkan notifikasi jika terjadi suatu peristiwa.

Jika dilakukan terus-menerus, chaos engineering dapat menunjukkan kekurangan dalam beban kerja Anda yang, jika dibiarkan tidak ditangani, dapat berdampak negatif pada ketersediaan dan operasi.

> **Note**
>
> Chaos engineering adalah bidang ilmu yang bereksperimen pada suatu sistem guna membangun kepercayaan pada kemampuan sistem untuk bertahan dari kondisi gangguan dalam produksi. – Prinsip-prinsip Chaos Engineering

Jika sistem mampu bertahan dari gangguan ini, eksperimen chaos harus dipertahankan sebagai pengujian regresi otomatis. Dengan demikian, eksperimen chaos harus dilakukan sebagai bagian dari siklus hidup pengembangan sistem (SDLC) Anda dan sebagai bagian dari pipeline CI/CD Anda.

Untuk memastikan bahwa beban kerja Anda dapat bertahan dari kegagalan komponen, lakukan injeksi peristiwa dunia nyata sebagai bagian dari eksperimen Anda. Misalnya, lakukan eksperimen dengan kehilangan instans Amazon EC2 atau failover instans basis data Amazon RDS utama, lalu verifikasi bahwa beban kerja Anda tidak terpengaruh (atau hanya sedikit terpengaruh). Gunakan kombinasi kesalahan komponen untuk membuat simulasi peristiwa yang mungkin disebabkan oleh gangguan di Zona Ketersediaan.

Untuk kesalahan tingkat aplikasi (seperti crash), Anda dapat memulai dengan stressor seperti kehabisan memori dan daya CPU.

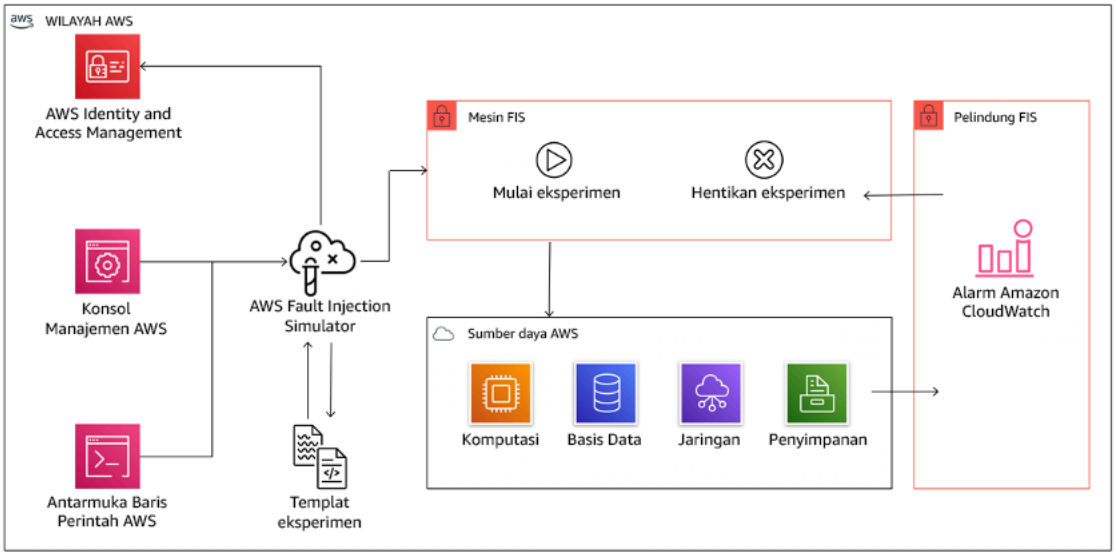
Untuk melakukan validasi mekanisme fallback atau failover untuk dependensi eksternal karena gangguan jaringan yang terputus-putus, komponen Anda harus menyimulasikan peristiwa tersebut dengan memblokir akses ke penyedia pihak ketiga selama durasi tertentu yang dapat berlangsung dari hitungan detik hingga jam.

Mode degradasi lainnya dapat menyebabkan berkurangnya fungsionalitas dan respons yang lambat, sehingga sering kali mengakibatkan gangguan terjadi pada layanan Anda. Degradasi ini umumnya disebabkan oleh terjadinya peningkatan latensi pada layanan yang sangat penting dan komunikasi jaringan yang tidak dapat diandalkan (paket yang tidak dikirim). Eksperimen dengan kesalahan ini, termasuk efek jaringan seperti latensi, pesan yang tidak terkirim, dan kegagalan DNS, dapat mencakup ketidakmampuan untuk mengganti nama, menjangkau layanan DNS, atau membuat koneksi ke layanan yang dependen.

Alat chaos engineering:

AWS Fault Injection Service (AWS FIS) sebuah adalah layanan terkelola penuh untuk menjalankan eksperimen injeksi kesalahan yang dapat digunakan sebagai bagian dari pipeline CD Anda, atau di luar pipeline. AWS FIS adalah pilihan yang baik untuk digunakan selama game day chaos

engineering. Ini mendukung pengenalan kesalahan bersamaan di berbagai jenis sumber daya termasuk Amazon EC2, Amazon Elastic Container Service (Amazon ECS), Amazon Elastic Kubernetes Service (Amazon EKS), dan Amazon RDS. Kesalahan ini termasuk terhentinya sumber daya, memaksa failover, membebani CPU atau memori, throttling, latensi, dan kehilangan paket. Karena layanan ini terintegrasi dengan Amazon CloudWatch Alarms, Anda dapat mengatur kondisi berhenti sebagai pagar pembatas untuk melakukan rollback jika eksperimen menyebabkan dampak tak terduga.



AWS Fault Injection Service terintegrasi dengan sumber daya AWS untuk memungkinkan Anda menjalankan eksperimen injeksi kesalahan untuk beban kerja Anda.

Ada juga beberapa opsi pihak ketiga untuk melakukan eksperimen injeksi kesalahan. Ini termasuk alat-alat sumber terbuka seperti Chaos Toolkit, Chaos Mesh, dan Litmus Chaos, serta opsi komersial seperti Gremlin. Untuk memperluas cakupan kesalahan yang dapat diinjeksikan pada AWS, AWS FIS terintegrasi dengan Chaos Mesh dan Litmus Chaos, sehingga memungkinkan Anda mengoordinasikan alur kerja injeksi kesalahan di antara beberapa alat. Misalnya, Anda dapat menjalankan pengujian tekanan pada CPU dari sebuah pod dengan menggunakan kesalahan Chaos Mesh atau Litmus sambil menghentikan sebagian simpul klaster yang dipilih secara acak menggunakan tindakan kesalahan AWS FIS.

Langkah-langkah implementasi

1. Tentukan kesalahan mana yang akan digunakan untuk eksperimen.

   Lakukan penilaian desain beban kerja Anda untuk mengetahui ketahanannya. Desain semacam itu (dibuat menggunakan praktik terbaik Kerangka Well-Architected) memperhitungkan risiko

berdasarkan dependensi kritis, peristiwa masa lalu, masalah yang diketahui, dan persyaratan kepatuhan. Buat daftar yang berisi setiap elemen desain yang dimaksudkan untuk menjaga ketahanan dan kesalahan yang akan dimitigasi oleh elemen desain tersebut. Untuk informasi selengkapnya tentang cara membuat daftar tersebut, lihat laporan resmi Peninjauan Kesiapan Operasional yang memandu Anda tentang cara membuat proses untuk mencegah terulangnya insiden sebelumnya. Proses Analisis Mode dan Efek Kegagalan (FMEA) memberikan Anda kerangka kerja untuk melakukan analisis tingkat komponen terhadap kegagalan dan bagaimana dampaknya terhadap beban kerja Anda. FMEA diuraikan secara lebih rinci oleh Adrian Cockcroft dalam Mode Kegagalan dan Ketahanan Berkelanjutan.

2. Tetapkan prioritas untuk setiap kesalahan.

   Mulailah dengan kategorisasi yang umum seperti tinggi, sedang, atau rendah. Untuk menilai prioritas, pertimbangkan frekuensi kesalahan dan dampak kegagalan terhadap beban kerja secara keseluruhan.

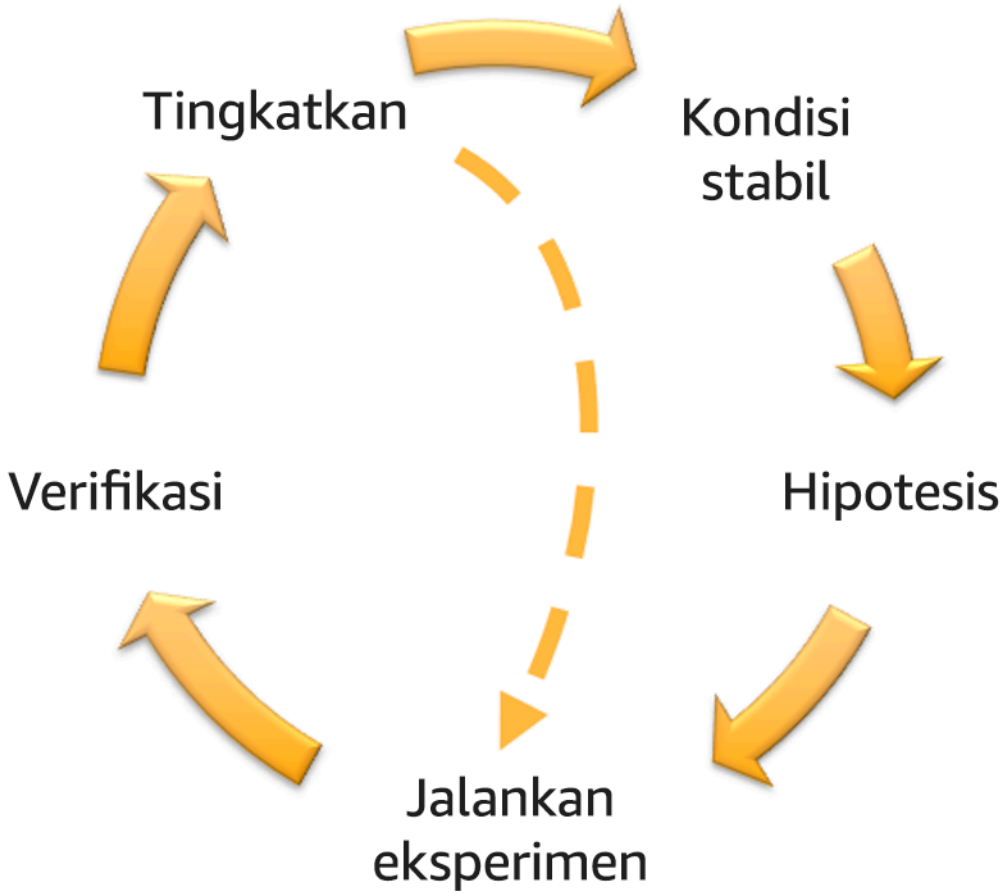
   Saat mempertimbangkan frekuensi kesalahan tertentu, lakukan analisis pada data terdahulu untuk beban kerja ini jika tersedia. Jika tidak tersedia, gunakan data dari beban kerja lain yang berjalan di lingkungan yang serupa.

   Ketika mempertimbangkan dampak dari kesalahan tertentu, makin besar cakupan kesalahan, biasanya makin besar dampaknya. Pertimbangkan juga desain dan tujuan beban kerja. Misalnya, kemampuan untuk mengakses penyimpanan data sumber sangat krusial untuk beban kerja yang melakukan transformasi dan analisis data. Dalam hal ini, Anda akan memprioritaskan eksperimen untuk kesalahan akses, serta akses yang di-throttling dan penyisipan latensi.

   Analisis pasca-insiden adalah sumber data yang baik untuk memahami frekuensi dan dampak mode kegagalan.

   Gunakan prioritas yang ditetapkan untuk menentukan kesalahan mana yang akan digunakan terlebih dahulu dalam eksperimen beserta urutannya agar dapat mengembangkan eksperimen injeksi kesalahan baru.

3. Untuk setiap eksperimen yang Anda lakukan, gunakan roda chaos engineering dan ketahanan berkelanjutan pada gambar berikut.



Roda chaos engineering dan ketahanan berkelanjutan yang menggunakan metode ilmiah dari Adrian Hornsby.

a. Definisikan kondisi stabil sebagai output terukur dari beban kerja yang menunjukkan perilaku normal.

   Beban kerja Anda menunjukkan kondisi stabil jika beroperasi dengan andal dan seperti yang diharapkan. Oleh karena itu, validasikan bahwa beban kerja Anda berkondisi baik sebelum menentukan kondisi stabil. Dalam kondisi stabil, bukan berarti tidak akan ada dampak pada beban kerja saat terjadi kesalahan, karena sejumlah kesalahan tertentu mungkin berada dalam batas yang dapat diterima. Kondisi stabil adalah acuan dasar yang akan Anda amati selama eksperimen, yang akan menunjukkan anomali jika hipotesis yang Anda tentukan pada langkah berikutnya tidak berjalan seperti yang diharapkan.

Misalnya, kondisi stabil sistem pembayaran dapat didefinisikan sebagai pemrosesan 300 TPS dengan tingkat keberhasilan 99% dan waktu round-trip 500 md.

b. Bentuk hipotesis tentang bagaimana beban kerja akan bereaksi terhadap kesalahan.

Hipotesis yang baik didasarkan pada bagaimana beban kerja diharapkan akan memitigasi kesalahan untuk mempertahankan kondisi stabil. Hipotesis menyatakan bahwa dengan kesalahan jenis tertentu, sistem atau beban kerja akan terus berkondisi stabil karena beban kerja ini dirancang dengan mitigasi tertentu. Jenis spesifik kesalahan dan mitigasi harus ditentukan dalam hipotesis.

Templat berikut dapat digunakan untuk hipotesis (tetapi pernyataan lain juga dapat diterima):

> **Note**
>
> Jika terjadi *`kesalahan tertentu`*, *`nama beban kerja`* akan *`menjelaskan mitigasi kontrol`* untuk mempertahankan *`dampak metrik bisnis atau teknis`*.

Misalnya:

- Jika 20% dari total simpul dalam grup simpul Amazon EKS dihapus, Transaction Create API akan terus melayani persentil ke-99 dari permintaan dalam waktu kurang dari 100 md (kondisi stabil). Simpul Amazon EKS akan pulih dalam waktu lima menit, dan pod akan dijadwalkan dan memproses lalu lintas dalam waktu delapan menit setelah dimulainya eksperimen. Peringatan akan diaktifkan dalam waktu tiga menit.
- Jika terjadi kegagalan instans Amazon EC2 tunggal, pemeriksaan kondisi Penyeimbangan Beban Elastis akan mengakibatkan Penyeimbangan Beban Elastis hanya mengirim permintaan ke instans berkondisi baik yang tersisa, sedangkan Amazon EC2 Auto Scaling mengganti instans yang gagal, sehingga mempertahankan peningkatan kesalahan sisi server (5xx) sebanyak kurang dari 0,01% (kondisi stabil).
- Jika instans basis data Amazon RDS utama gagal, beban kerja pengumpulan data Rantai Pasokan akan melakukan failover dan terhubung ke instans basis data Amazon RDS yang siaga untuk mempertahankan kesalahan baca atau tulis basis data selama kurang dari 1 menit (kondisi stabil).

c. Jalankan eksperimen dengan menginjeksikan kesalahan.

Eksperimen secara default harus memiliki kemampuan fail-safe dan ditoleransi oleh beban kerja. Jika Anda tahu bahwa beban kerja akan gagal, jangan jalankan eksperimen. Chaos engineering harus digunakan untuk menemukan “known-unknown” atau “unknown-unknown”. Known-unknown adalah hal-hal yang Anda ketahui, tetapi tidak sepenuhnya paham, dan unknown-unknown adalah hal-hal yang tidak Anda ketahui atau pahami sepenuhnya. Bereksperimen dengan beban kerja yang Anda tahu dalam kondisi rusak tidak akan memberi Anda wawasan baru. Eksperimen Anda harus direncanakan dengan cermat, memiliki cakupan dampak yang jelas, dan menyediakan mekanisme rollback yang dapat diterapkan jika terjadi gangguan tak terduga. Jika uji tuntas Anda menunjukkan bahwa beban kerja Anda dapat bertahan dalam eksperimen, lanjutkan eksperimen. Ada beberapa opsi untuk menginjeksikan kesalahan. Untuk beban kerja aktif pada AWS, AWS FIS menyediakan banyak simulasi kesalahan standar yang disebut tindakan. Anda juga dapat menentukan tindakan kustom yang berjalan di AWS FIS menggunakan dokumen AWS Systems Manager.

Kami tidak menyarankan penggunaan skrip kustom untuk eksperimen chaos, kecuali jika skrip tersebut memiliki kemampuan untuk memahami status terkini beban kerja, mampu menghasilkan log, dan menyediakan mekanisme untuk rollback dan kondisi berhenti jika memungkinkan.

Kerangka kerja atau kumpulan alat efektif yang mendukung chaos engineering harus melacak kondisi terkini eksperimen, menghasilkan log, dan menyediakan mekanisme rollback untuk mendukung pelaksanaan eksperimen yang terkontrol. Mulailah dengan layanan andal seperti AWS FIS yang akan memungkinkan Anda melakukan eksperimen dengan cakupan yang jelas dan mekanisme keamanan yang melakukan rollback jika eksperimen menimbulkan gangguan tak terduga. Untuk mempelajari tentang berbagai eksperimen yang lebih luas menggunakan AWS FIS, lihat juga lab Aplikasi Tangguh dan Well-Architected dengan Chaos Engineering. Selain itu, AWS Resilience Hub akan menganalisis beban kerja Anda dan membuat eksperimen yang dapat Anda pilih untuk diterapkan dan dijalankan di AWS FIS.

> **Note**
>
> Untuk setiap eksperimen, pahami dengan jelas cakupan dan dampaknya. Kami merekomendasikan bahwa kesalahan harus disimulasikan terlebih dahulu di lingkungan nonproduksi sebelum dijalankan dalam produksi.

Eksperimen harus berjalan dalam produksi di bawah beban dunia nyata menggunakan deployment canary yang memutar kontrol dan deployment sistem eksperimental, jika memungkinkan. Menjalankan eksperimen selama waktu sepi adalah praktik yang baik untuk mengurangi potensi dampak saat pertama kali bereksperimen dalam produksi. Selain itu, jika menggunakan lalu lintas pelanggan yang sebenarnya akan menimbulkan terlalu banyak risiko, Anda dapat menjalankan eksperimen menggunakan lalu lintas sintetis di infrastruktur produksi terhadap deployment kontrol dan eksperimental. Jika tidak dapat menggunakan produksi, jalankan eksperimen di lingkungan praproduksi yang semirip mungkin dengan produksi.

Anda harus membuat dan memantau pagar pembatas untuk memastikan eksperimen tidak memengaruhi lalu lintas produksi atau sistem lain di luar batas yang dapat diterima. Tetapkan kondisi berhenti untuk menghentikan eksperimen jika mencapai ambang batas pada metrik pagar pembatas yang Anda tentukan. Hal ini harus mencakup metrik untuk kondisi stabil beban kerja, serta metrik berdasarkan komponen yang diinjeksi dengan kesalahan. Monitor sintetis (juga dikenal sebagai canary pengguna) adalah salah satu metrik yang biasanya harus Anda sertakan sebagai proksi pengguna. Hentikan kondisi untuk AWS FIS didukung sebagai bagian dari templat eksperimen, sehingga memungkinkan maksimal lima kondisi berhenti per templat.

Salah satu prinsip chaos adalah meminimalkan cakupan eksperimen dan dampaknya:

Meskipun harus ada kelonggaran untuk beberapa dampak negatif dalam jangka pendek, Chaos Engineer bertanggung jawab dan berkewajiban untuk memastikan gangguan dari eksperimen diminimalkan dan dikendalikan.

Metode untuk memverifikasi cakupan dan dampak potensial adalah dengan melakukan eksperimen di lingkungan nonproduksi terlebih dahulu, memverifikasi bahwa ambang batas untuk kondisi berhenti diaktifkan seperti yang diharapkan selama eksperimen dan kemampuan pengamatan (observabilitas) diterapkan untuk menemukan pengecualian, bukan langsung bereksperimen dalam produksi.

Saat menjalankan eksperimen injeksi kesalahan, verifikasikan bahwa semua pihak yang bertanggung jawab sudah mengetahui informasi yang jelas. Berkomunikasilah dengan tim yang sesuai seperti tim operasi, tim keandalan layanan, dan dukungan pelanggan untuk memberi tahu mereka kapan eksperimen akan dijalankan dan apa yang diharapkan. Berikan alat komunikasi kepada berbagai tim ini untuk memberi tahu tim tertentu yang menjalankan eksperimen jika muncul efek yang merugikan.

Anda harus memulihkan beban kerja dan sistem yang mendasarinya kembali ke kondisi awal yang diketahui berfungsi baik. Sering kali, desain beban kerja yang tangguh akan pulih sendiri. Namun, beberapa desain yang salah atau eksperimen yang gagal dapat membuat beban kerja Anda berada dalam kondisi kegagalan yang tidak terduga. Pada akhir eksperimen, Anda harus menyadari hal ini dan memulihkan beban kerja dan sistem. Dengan AWS FIS, Anda dapat mengatur konfigurasi rollback (juga disebut post action) dalam parameter tindakan. Post action mengembalikan target ke keadaan sebelum tindakan dijalankan. Baik diotomatiskan (seperti menggunakan AWS FIS) maupun manual, post action ini harus menjadi bagian dari playbook yang menjelaskan cara mendeteksi dan menangani kegagalan.

d. Verifikasikan hipotesisnya.

   Prinsip Chaos Engineering memberikan panduan tentang cara memverifikasi kondisi stabil beban kerja Anda:

   Fokus pada output terukur dari suatu sistem, bukan atribut internal sistem. Pengukuran output tersebut selama periode waktu yang singkat merupakan proksi untuk kondisi stabil sistem. Throughput sistem secara keseluruhan, tingkat kesalahan, dan persentil latensi semuanya dapat menjadi metrik penting yang merepresentasikan perilaku kondisi stabil. Dengan berfokus pada pola perilaku sistemik selama eksperimen, chaos engineering memverifikasi bahwa sistem berfungsi, bukan mencoba memvalidasi cara kerjanya.

   Dalam dua contoh sebelumnya, kami menyertakan metrik kondisi stabil dengan peningkatan kesalahan sisi server (5xx) sebanyak kurang dari 0,01% serta kesalahan baca dan tulis basis data selama kurang dari satu menit.

   Kesalahan 5xx adalah metrik yang baik karena merupakan konsekuensi dari mode kegagalan yang akan dialami langsung oleh klien yang menggunakan beban kerja. Pengukuran kesalahan basis data cocok digunakan sebagai konsekuensi langsung dari kesalahan, tetapi juga harus dilengkapi dengan pengukuran dampak klien seperti permintaan pelanggan yang gagal atau kesalahan yang muncul bagi klien. Selain itu, sertakan pemantauan sintetis (juga dikenal sebagai user canary) pada API atau URI apa pun yang diakses langsung oleh klien yang menggunakan beban kerja Anda.

e. Tingkatkan desain beban kerja agar memiliki ketahanan.

   Jika kondisi stabil tidak dipertahankan, selidiki cara desain beban kerja dapat ditingkatkan untuk mengurangi kesalahan, dengan menerapkan praktik terbaik dari pilar Keandalan Well-Architected AWS. Panduan dan sumber daya tambahan dapat ditemukan di Perpustakaan

Pembangun AWS, yang menghosting artikel tentang cara meningkatkan pemeriksaan kondisi Anda atau menggunakan percobaan ulang dengan backoff dalam kode aplikasi Anda, antara lain.

Setelah perubahan ini diterapkan, jalankan eksperimen lagi (ditunjukkan dengan garis putus-putus pada roda chaos engineering) untuk mengetahui keefektifannya. Jika langkah verifikasi menunjukkan bahwa hipotesisnya benar, beban kerja akan berada dalam kondisi stabil, dan siklusnya berlanjut.

4. Jalankan eksperimen secara rutin.

   Eksperimen chaos adalah sebuah siklus, dan eksperimen harus dijalankan secara rutin sebagai bagian dari chaos engineering. Setelah beban kerja memenuhi hipotesis eksperimen, eksperimen harus diotomatiskan untuk terus berjalan sebagai bagian regresi dalam alur CI/CD Anda. Untuk mempelajari cara melakukan ini, lihat blog ini tentang cara menjalankan eksperimen AWS FIS menggunakan AWS CodePipeline. Laboratorium tentang eksperimen AWS FIS berulang dalam pipeline CI/CD ini memungkinkan Anda untuk bekerja langsung.

   Eksperimen injeksi kesalahan juga merupakan bagian dari game day (lihat REL12-BP05 Mengadakan game day secara rutin). Game day menyimulasikan kegagalan atau peristiwa untuk memverifikasi sistem, proses, dan respons tim. Tujuannya adalah untuk benar-benar menerapkan tindakan yang perlu dilakukan oleh tim seolah memang terjadi peristiwa yang tidak diharapkan.

5. Rekam dan simpan hasil eksperimen.

   Hasil eksperimen injeksi kesalahan harus direkam dan dijadikan persisten. Sertakan semua data yang diperlukan (seperti waktu, beban kerja, dan kondisi) agar dapat menganalisis hasil dan tren eksperimen nantinya. Contoh hasilnya dapat mencakup tangkapan layar dasbor, dump CSV dari basis data metrik Anda, atau catatan ketik manual yang berisi peristiwa dan pengamatan dari eksperimen. Pencatatan eksperimen dengan AWS FIS dapat menjadi bagian dari pengambilan data ini.

Sumber daya

Praktik-praktik terbaik terkait:

- REL08-BP03 Mengintegrasikan pengujian ketahanan sebagai bagian dari deployment Anda
- REL13-BP03 Menguji implementasi pemulihan bencana untuk memvalidasi implementasi

Dokumen terkait:

- Apa itu AWS Fault Injection Service?
- Apa itu AWS Resilience Hub?
- Prinsip-prinsip Chaos Engineering
- Chaos Engineering: Merencanakan eksperimen pertama Anda
- Rekayasa Ketahanan: Belajar Menerima Kegagalan
- Kisah Chaos Engineering
- Menghindari fallback dalam sistem terdistribusi
- Deployment Canary untuk Eksperimen Chaos

Video terkait:

- AWS re:Invent 2020: Menguji ketahanan menggunakan chaos engineering (ARC316)
- AWS re:Invent 2019: Meningkatkan ketahanan dengan chaos engineering (DOP309-R1)
- AWS re:Invent 2019: Melakukan chaos engineering di dunia nirserver (CMY301)

Alat terkait:

- AWS Fault Injection Service
- AWS Marketplace: Platform Chaos Engineering Gremlin
- Chaos Toolkit
- Chaos Mesh
- Litmus

REL12-BP05 Mengadakan game day secara rutin

Lakukan game day untuk melatih prosedur Anda secara teratur dalam merespons kejadian dan gangguan yang memengaruhi beban kerja. Libatkan tim yang sama yang akan bertanggung jawab untuk menangani skenario produksi. Latihan-latihan ini membantu menerapkan langkah untuk mencegah dampak pada pengguna yang disebabkan oleh peristiwa produksi. Ketika Anda melatih prosedur respons Anda dalam kondisi realistis, Anda dapat mengidentifikasi dan mengatasi setiap kesenjangan atau kelemahan sebelum kejadian nyata terjadi.

Game day menyimulasikan peristiwa di lingkungan serupa produksi untuk menguji sistem, proses, dan respons tim. Tujuannya adalah untuk melakukan tindakan yang sama yang perlu dilakukan oleh tim seolah-olah peristiwa yang tidak diharapkan benar-benar terjadi. Latihan ini akan membantu Anda memahami sisi mana yang perlu ditingkatkan dan membantu mengembangkan pengalaman organisasi dalam menangani peristiwa dan gangguan. Hal ini harus dilakukan secara teratur sehingga tim Anda akan mengembangkan kebiasaan tentang cara memberikan respons.

Game day mempersiapkan tim untuk menangani peristiwa produksi dengan lebih percaya diri. Tim yang terlatih dengan baik lebih mampu mendeteksi dan merespons berbagai skenario dengan cepat. Hal ini akan jauh meningkatkan kesiapan dan postur ketahanan.

Hasil yang diinginkan: Anda menjalankan game day ketahanan secara konsisten dan terjadwal. Game day ini dianggap sebagai hal normal yang sama pentingnya dengan kegiatan bisnis lainnya. Organisasi Anda telah membangun budaya kesiapsiagaan, dan ketika masalah produksi terjadi, tim Anda siap untuk merespons secara efektif, menyelesaikan masalah secara efisien, dan memitigasi dampak terhadap pelanggan.

Anti-pola umum:

- Anda mendokumentasikan prosedur, tetapi tidak pernah mengadakan latihannya.
- Anda tidak melibatkan pengambil keputusan bisnis dalam latihan pengujian.
- Anda menjalankan game day, tetapi Anda tidak memberi tahu semua pemangku kepentingan yang relevan.
- Anda hanya fokus pada kegagalan teknis, tetapi Anda tidak melibatkan pemangku kepentingan bisnis.
- Anda tidak menerapkan pelajaran yang dipetik dari game day ke dalam proses pemulihan Anda.
- Anda menyalahkan tim atas kegagalan atau bug.

Manfaat menjalankan praktik terbaik ini:

- Meningkatkan keterampilan respons: Pada game day, tim mempraktikkan tugas mereka dan menguji mekanisme komunikasi mereka selama peristiwa simulasi, sehingga menciptakan respons yang lebih terkoordinasi dan efisien dalam situasi produksi.
- Mengidentifikasi dan mengatasi dependensi: Lingkungan yang kompleks sering kali melibatkan dependensi yang rumit antara berbagai sistem, layanan, dan komponen. Game day dapat membantu Anda mengidentifikasi dan mengatasi dependensi ini, dan memverifikasi bahwa sistem

dan layanan penting Anda tercakup dengan benar dalam prosedur runbook Anda dan dapat dinaikkan skalanya atau dipulihkan dengan segera.

- Menumbuhkan budaya ketahanan: Game day dapat membantu menumbuhkan pola pikir ketahanan dalam suatu organisasi. Ketika Anda melibatkan tim lintas fungsi dan pemangku kepentingan, latihan ini mempromosikan kesadaran, kolaborasi, dan pemahaman bersama tentang pentingnya ketahanan di seluruh organisasi.
- Peningkatan dan adaptasi berkelanjutan: Game day reguler membantu Anda untuk terus menilai dan menyesuaikan strategi ketahanan Anda, sehingga membuatnya tetap relevan dan efektif dalam menghadapi keadaan yang berubah-ubah.
- Meningkatkan kepercayaan pada sistem: Game day yang berhasil dapat membantu Anda membangun kepercayaan pada kemampuan sistem untuk bertahan dan pulih dari gangguan.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Sedang

Panduan implementasi

Setelah Anda merancang dan mengimplementasikan langkah-langkah ketahanan yang diperlukan, lakukan game day untuk memvalidasi bahwa semuanya berfungsi sesuai rencana dalam produksi. Game day, terutama yang pertama, harus melibatkan semua anggota tim. Semua pemangku kepentingan dan peserta harus diberi tahu terlebih dahulu tentang tanggal, waktu, dan skenario simulasi.

Selama game day, tim yang terlibat menyimulasikan berbagai peristiwa dan skenario potensial sesuai dengan prosedur yang ditentukan. Para peserta memantau dengan cermat dan menilai dampak dari peristiwa simulasi ini. Jika sistem beroperasi sesuai rancangan, deteksi otomatis, penskalaan, dan mekanisme pemulihan mandiri seharusnya aktif dan hanya berdampak sedikit atau sama sekali tidak berdampak pada pengguna. Jika tim menemukan dampak negatif, mereka melakukan rollback pengujian dan memperbaiki masalah yang diidentifikasi, baik melalui cara otomatis atau intervensi manual yang didokumentasikan dalam runbook yang berlaku.

Untuk terus meningkatkan ketahanan, penting untuk mendokumentasikan dan menerapkan pelajaran yang dipetik. Proses ini disebut siklus umpan balik yang secara sistematis mengumpulkan wawasan dari game day dan menggunakannya untuk meningkatkan sistem, proses, dan kemampuan tim.

Untuk membantu Anda menciptakan skenario tiruan dunia nyata berupa kegagalan komponen atau layanan sistem secara tak terduga, injeksikan kesalahan tersimulasi sebagai latihan game day. Tim dapat menguji ketahanan dan toleransi kesalahan sistem mereka dan menyimulasikan respons insiden serta proses pemulihan mereka di lingkungan terkontrol.

Di AWS, game day Anda dapat dilakukan dengan replika lingkungan produksi Anda menggunakan infrastruktur sebagai kode. Melalui proses ini, Anda dapat menguji di lingkungan yang aman yang sangat mirip dengan lingkungan produksi Anda. Pertimbangkan AWS Fault Injection Service untuk membuat beberapa skenario kegagalan yang berbeda. Gunakan layanan seperti Amazon CloudWatch dan AWS X-Ray untuk memantau perilaku sistem selama game day. Gunakan AWS Systems Manager untuk mengelola serta menjalankan playbook, dan gunakan AWS Step Functions untuk melakukan orkestrasi alur kerja game day berulang.

Langkah-langkah implementasi

- Buat program game day: Kembangkan program terstruktur yang menentukan frekuensi, ruang lingkup, dan tujuan game day. Libatkan pemangku kepentingan utama dan ahli bidang studi dalam merencanakan dan menjalankan latihan ini.
- Siapkan game day:
  1. Identifikasi layanan paling kritis bagi bisnis yang menjadi fokus game day. Buat katalog dan petakan orang, proses, dan teknologi yang mendukung layanan tersebut.
  2. Tetapkan agenda untuk game day, dan persiapkan tim yang terlibat untuk berpartisipasi dalam acara tersebut. Siapkan layanan otomatisasi Anda untuk menyimulasikan skenario yang direncanakan dan menjalankan proses pemulihan yang sesuai. Layanan AWS seperti AWS Fault Injection Service, AWS Step Functions, dan AWS Systems Manager dapat membantu Anda mengotomatiskan berbagai aspek game day, seperti injeksi kesalahan dan inisiasi tindakan pemulihan.
- Jalankan simulasi Anda: Pada game day, jalankan skenario yang direncanakan. Amati dan dokumentasikan bagaimana orang, proses, dan teknologi bereaksi terhadap peristiwa simulasi.
- Lakukan peninjauan pasca-latihan: Setelah game day, lakukan sesi retrospektif untuk meninjau pelajaran yang dipetik. Identifikasi area untuk perbaikan dan tindakan apa pun yang diperlukan untuk meningkatkan ketahanan operasional. Dokumentasikan temuan Anda, dan lacak setiap perubahan yang diperlukan untuk meningkatkan strategi ketahanan dan kesiapan Anda untuk menyelesaikannya.

Sumber daya

Praktik-praktik terbaik terkait:

- REL12-BP01 Menggunakan playbook untuk menyelidiki kegagalan
- REL12-BP04 Menguji ketahanan menggunakan chaos engineering

- OPS04-BP01 Identifikasikan indikator performa utama
- OPS07-BP03 Menggunakan runbook untuk menjalankan prosedur
- OPS10-BP01 Menggunakan proses untuk manajemen peristiwa, insiden, dan masalah

Dokumen terkait:

- Apa itu GameDay AWS?

Video terkait:

- AWS re:Invent 2023 - Berlatihlah seperti bermain: Bagaimana Amazon meningkatkan ketahanan ke tingkat yang lebih tinggi

Contoh terkait:

- AWS Lokakarya - Menghadapi berbagai tantangan: Menerapkan kekacauan terkendali untuk menciptakan sistem yang tangguh
- Membuat Game Day Anda Sendiri untuk Mendukung Ketahanan Operasional

Rencanakan Pemulihan Bencana (DR)

Memiliki cadangan dan komponen beban kerja berlebih adalah permulaan dari strategi DR Anda. RTO dan RPO adalah sasaran pemulihan beban kerja Anda. Tetapkan ini berdasarkan kebutuhan bisnis. Implementasikan sebuah strategi untuk memenuhi tujuan-tujuan ini, sambil mempertimbangkan lokasi dan fungsi data dan sumber daya beban kerja. Probabilitas gangguan dan biaya pemulihan juga merupakan faktor penting yang akan membantu menginformasikan nilai bisnis dari penyediaan pemulihan bencana untuk beban kerja.

Ketersediaan dan Pemulihan Bencana sama-sama mengandalkan praktik terbaik, seperti pemantauan kegagalan, deployment ke beberapa lokasi, dan failover otomatis. Namun demikian, Ketersediaan berfokus pada komponen beban kerja, sedangkan Pemulihan Bencana berfokus pada salinan terpisah (discrete) dari keseluruhan beban kerja. Pemulihan Bencana memiliki tujuan yang berbeda dari Ketersediaan, yang berfokus pada waktu pemulihan setelah terjadi sebuah bencana.

Praktik terbaik

- REL13-BP01 Menetapkan sasaran pemulihan untuk waktu henti dan kehilangan data

- REL13-BP02 Menggunakan strategi pemulihan untuk memenuhi sasaran pemulihan
- REL13-BP03 Menguji implementasi pemulihan bencana untuk memvalidasi implementasi
- REL13-BP04 Mengelola penyimpangan konfigurasi di lokasi atau Wilayah Pemulihan Bencana (DR)
- REL13-BP05 Mengotomatiskan pemulihan

REL13-BP01 Menetapkan sasaran pemulihan untuk waktu henti dan kehilangan data

Kegagalan dapat merugikan bisnis Anda dalam banyak hal. Pertama, kegagalan dapat menyebabkan gangguan layanan (waktu henti). Kedua, kegagalan dapat menyebabkan data hilang, tidak konsisten, atau usang. Untuk memandu cara Anda merespons dan memulihkan diri dari kegagalan, tentukan Sasaran Waktu Pemulihan (RTO) dan Sasaran Titik Pemulihan (RPO) untuk setiap beban kerja. Sasaran Waktu Pemulihan (RTO) adalah jeda waktu maksimum yang dapat diterima antara gangguan layanan dan pemulihan layanan. Sasaran Titik Pemulihan (RPO) adalah waktu maksimum yang dapat diterima sejak titik pemulihan data terakhir.

Hasil yang diinginkan: Setiap beban kerja memiliki RTO dan RPO yang ditentukan berdasarkan pertimbangan teknis dan dampak bisnis.

Anti-pola umum:

- Anda belum menentukan sasaran pemulihan.
- Anda memilih sasaran pemulihan tanpa pertimbangan matang.
- Anda memilih sasaran pemulihan yang terlalu longgar dan tidak memenuhi sasaran bisnis.
- Anda belum mengevaluasi dampak waktu henti dan kehilangan data.
- Anda memilih sasaran pemulihan yang tidak realistis, seperti nol waktu pemulihan atau nol kehilangan data, yang mungkin tidak dapat dicapai untuk konfigurasi beban kerja Anda.
- Anda memilih sasaran pemulihan yang lebih ketat daripada sasaran bisnis yang sesungguhnya. Hal ini memaksakan implementasi pemulihan yang lebih mahal dan lebih rumit dibandingkan yang dibutuhkan beban kerja.
- Anda memilih sasaran pemulihan yang tidak kompatibel dengan sasaran beban kerja dependen.
- Anda tidak mempertimbangkan persyaratan peraturan dan kepatuhan.

Manfaat menjalankan praktik terbaik ini: Ketika Anda menetapkan RTO dan RPO untuk beban kerja Anda, Anda menetapkan tujuan yang jelas dan terukur untuk pemulihan berdasarkan kebutuhan bisnis Anda. Setelah Anda menetapkan sasaran-sasaran tersebut, Anda dapat membuat rencana pemulihan bencana (DR) yang disesuaikan untuk memenuhinya.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Tinggi

Panduan implementasi

Buat matriks atau lembar kerja untuk membantu memandu perencanaan pemulihan bencana Anda. Dalam matriks Anda, buat kategori atau tingkatan beban kerja yang berbeda berdasarkan dampak bisnisnya (seperti kritis, tinggi, sedang, dan rendah) serta RTO dan RPO terkait yang ditargetkan untuk setiap kategori. Matriks berikut memberikan contoh (perhatikan bahwa nilai RTO dan RPO Anda mungkin berbeda) yang dapat Anda ikuti:

Matriks Pemulihan Bencana						
		Sasaran Titik Pemulihan				
		< 1 Menit	< 1 Jam	< 6 Jam	< 1 Hari	+ 1 Hari
Sasaran Waktu Pemulihan	< 10 Menit	Kritis	Kritis	Tinggi	Sedang	Sedang
	< 2 Jam	Kritis	Tinggi	Sedang	Sedang	Rendah
	< 8 Jam	Tinggi	Sedang	Sedang	Rendah	Rendah
	< 24 Jam	Sedang	Sedang	Rendah	Rendah	Rendah
	Lebih dari 24 Jam	Sedang	Rendah	Rendah	Rendah	Rendah

Contoh matriks pemulihan bencana

Untuk setiap beban kerja, selidiki dan pahami dampak waktu henti serta kehilangan data pada bisnis Anda. Dampaknya biasanya meningkat seiring dengan waktu henti dan kehilangan data, tetapi bentuk dampaknya dapat berbeda berdasarkan jenis beban kerja. Misalnya, waktu henti hingga satu jam mungkin berdampak rendah, tetapi setelah itu, dampaknya bisa meningkat dengan cepat. Dampak dapat berupa banyak hal, termasuk dampak keuangan (seperti kehilangan pendapatan), dampak reputasi (termasuk hilangnya kepercayaan pelanggan), dampak operasional (seperti gaji yang terlambat atau penurunan produktivitas), dan risiko peraturan. Jika sudah, tetapkan beban kerja ke tingkatan yang sesuai.

Pertimbangkan pertanyaan-pertanyaan berikut ketika Anda menganalisis dampak kegagalan:

1. Berapa lama waktu maksimum beban kerja bisa tidak tersedia sebelum menimbulkan dampak yang signifikan pada bisnis?
2. Seberapa besar dampak, dan seperti apa bentuknya, yang akan ditimbulkan pada bisnis dari suatu gangguan beban kerja? Pertimbangkan semua jenis dampak, termasuk keuangan, reputasi, operasional, dan peraturan.
3. Berapakah jumlah data maksimum yang dapat hilang atau tidak dapat dipulihkan sebelum menimbulkan dampak yang signifikan pada bisnis?
4. Apakah data yang hilang bisa dibuat lagi dari sumber lain (juga dikenal sebagai data turunan)? Jika demikian, pertimbangkan juga RPO dari semua data sumber yang digunakan untuk membuat ulang data beban kerja.
5. Apa saja sasaran pemulihan dan ekspektasi ketersediaan beban kerja yang bergantung pada sasaran pemulihan ini (hilir)? Sasaran beban kerja Anda harus dapat dicapai mengingat kemampuan pemulihan dependensi hilirnya. Pertimbangkan kemungkinan solusi alternatif atau mitigasi untuk dependensi hilir yang dapat meningkatkan kemampuan pemulihan beban kerja ini.
6. Apa saja sasaran pemulihan dan ekspektasi ketersediaan beban kerja yang bergantung pada sasaran pemulihan ini (hulu)? Sasaran beban kerja hulu mungkin mengharuskan beban kerja ini memiliki kemampuan pemulihan yang lebih ketat daripada yang terlihat pada awalnya.
7. Apakah ada sasaran pemulihan yang berbeda berdasarkan jenis insiden? Misalnya, Anda mungkin memiliki RTO dan RPO yang berbeda, bergantung pada apakah insiden tersebut memengaruhi suatu Zona Ketersediaan atau seluruh Wilayah.
8. Apakah sasaran pemulihan Anda berubah selama peristiwa atau waktu tertentu dalam setahun? Misalnya, Anda mungkin memiliki RTO dan RPO yang berbeda di sekitar musim belanja liburan, acara olahraga, obral khusus, dan peluncuran produk baru.
9. Bagaimana sasaran pemulihan selaras dengan strategi pemulihan bencana lini bisnis dan organisasi yang mungkin Anda miliki?
10. Apakah ada konsekuensi hukum atau kontrak yang perlu dipertimbangkan? Misalnya, apakah Anda secara kontraktual berkewajiban untuk menyediakan layanan dengan RTO atau RPO tertentu? Konsekuensi apa yang mungkin Anda terima jika tidak memenuhi kewajiban tersebut?
11. Apakah Anda diwajibkan untuk menjaga integritas data untuk memenuhi persyaratan peraturan atau kepatuhan?

Lembar kerja berikut dapat membantu evaluasi Anda untuk setiap beban kerja. Anda dapat mengubah lembar kerja ini sesuai kebutuhan spesifik Anda, seperti memasukkan pertanyaan tambahan.

Langkah 2: Pertanyaan utama	Berlaku untuk beban kerja?	RTO beban kerja	RPO beban kerja	Penyesuaian RTO.	Penyesuaian RPO.	Instruksi
[1] waktu maksimum beban kerja dapat mengalami waktu henti						diukur pada waktunya dari mulai pemadaman hingga pemulihan
[2] jumlah maksimum data yang bisa hilang						diukur pada waktunya sejak set data dapat dipulihkan berkualitas baik terakhir diketahui
[3a] dependensi hulu						masukkan tujuan pemulihan hilir yang paling ketat
[3b] dependensi hilir						masukkan tujuan pemulihan hilir yang paling longgar
[3a] dependensi hulu yang direkonsiliasi						Jika nilai hulu kurang dari nilai saat ini dan nilai hilir lebih besar, bekerjalah dengan
[3b] dependensi hilir yang direkonsiliasi						dependensi untuk merekonsiliasi dan masukkan nilai yang direkonsiliasi di sini
[3] dependensi						turunkan nilai untuk memenuhi dependensi hulu atau naikkan berdasarkan kemampuan dependensi hilir
Langkah 2: Pertanyaan tambahan						Tunjukkan apakah pertanyaan berlaku. Jika tidak, lewati
RTO/RPO dasar						Turunkan nilai RTO dan RPO dari atas ke sini
[4] tipe pemadaman	[ ]Y / [ ]N					Masukkan tujuan pemulihan untuk tipe peristiwa dengan persyaratan paling ketat
[5] tujuan berbasis waktu khusus	[ ]Y / [ ]N					Masukkan tujuan pemulihan untuk waktu-waktu dengan persyaratan paling ketat
[6] pelanggan terganggu	[ ]Y / [ ]N					Buat grafik pelanggan yang terkena dampak saat fungsi mengalami waktu henti atau data hilang. Gunakan grafik tersebut untuk memasukkan RTO dan RPO maksimum yang diizinkan berdasarkan dampak pelanggan
[7] dampak reputasi	[ ]Y / [ ]N					Bekerjalah dengan bisnis untuk menentukan RTO dan RPO berdasarkan dampak terhadap reputasi
[8] dampak operasi	[ ]Y / [ ]N					Masukkan RTO dan RPO maksimum berdasarkan dampak operasional
[9] penyelarasan organisasi	[ ]Y / [ ]N					Masukkan RTO dan RPO maksimum untuk beban kerja tipe ini sesuai LOB dan persyaratan organisasi
[10] kewajiban kontrak	[ ]Y / [ ]N					Masukkan RTO dan RPO maksimum berdasarkan kewajiban kontrak
[11] kepatuhan peraturan	[ ]Y / [ ]N					Masukkan RTO dan RPO maksimum berdasarkan kepatuhan peraturan yang berlaku
target berdasarkan pertanyaan tambahan						Ambil nilai minimum (nilai yang lebih ketat) dari Q 4-11 dan masukkan di sini
target yang disesuaikan						Jika tujuan di baris di atas tidak dapat diakomodasi, bekerjalah dengan pemangku kepentingan untuk mengurai hambatan, dan masukkan jumlah minimum baru di sini
RTO/RPO yang disesuaikan						Masukkan nilai RPO/RTO acuan, atau target yang disesuaikan, mana saja yang lebih rendah
Langkah 3						
Peta ke kategori atau tingkat yang ditetapkan sebelumnya						Sesuaikan kedua nilai ke bawah (lebih ketat) agar selaras dengan tingkat yang ditetapkan terdekat

Lembar kerja

Langkah-langkah implementasi

1. Identifikasi pemangku kepentingan bisnis dan tim teknis yang bertanggung jawab atas setiap beban kerja, dan libatkan mereka.
2. Buat kategori atau tingkat kekritisan untuk dampak beban kerja di organisasi Anda. Contoh kategori meliputi kritis, tinggi, sedang, dan rendah. Untuk setiap kategori, pilih RTO dan RPO yang mencerminkan tujuan serta persyaratan bisnis Anda.
3. Tetapkan salah satu kategori dampak yang Anda buat di langkah sebelumnya ke setiap beban kerja. Untuk memutuskan bagaimana suatu beban kerja dipetakan ke suatu kategori, pertimbangkan pentingnya beban kerja bagi bisnis serta dampak gangguan atau kehilangan data, lalu gunakan pertanyaan di atas untuk memandu Anda. Hal ini menghasilkan RTO dan RPO untuk setiap beban kerja.
4. Pertimbangkan RTO dan RPO untuk setiap beban kerja yang ditentukan pada langkah sebelumnya. Libatkan tim bisnis dan teknis beban kerja untuk menentukan apakah sasaran harus disesuaikan. Misalnya, pemangku kepentingan bisnis dapat menentukan bahwa target yang

lebih ketat diperlukan. Atau, tim teknis dapat menentukan bahwa target harus dimodifikasi untuk membuatnya dapat dicapai dengan sumber daya yang tersedia dan keterbatasan teknologi.

Sumber daya

Praktik-praktik terbaik terkait:

- REL09-BP04 Melakukan pemulihan data secara berkala untuk memverifikasi integritas dan proses pencadangan
- REL12-BP01 Menggunakan playbook untuk menyelidiki kegagalan
- REL13-BP02 Menggunakan strategi pemulihan untuk memenuhi sasaran pemulihan
- REL13-BP03 Menguji implementasi pemulihan bencana untuk memvalidasi implementasi

Dokumen terkait:

- AWS Blog Arsitektur : Seri Pemulihan Bencana
- Pemulihan Bencana Beban Kerja di AWS: Pemulihan di Cloud (Laporan Resmi AWS)
- Mengelola kebijakan ketangguhan dengan AWS Resilience Hub
- Partner APN: partner yang dapat membantu Anda melakukan pemulihan bencana
- AWS Marketplace: produk yang dapat digunakan untuk pemulihan bencana

Video terkait:

- AWS re:Invent 2018: Pola Arsitektur untuk Aplikasi Aktif-Aktif Multi-Wilayah
- Pemulihan Bencana Beban Kerja di AWS

REL13-BP02 Menggunakan strategi pemulihan untuk memenuhi sasaran pemulihan

Tentukan strategi pemulihan bencana (DR) yang memenuhi sasaran pemulihan beban kerja. Pilih strategi seperti pencadangan dan pemulihan, standby (aktif/pasif), atau aktif/aktif.

Hasil yang diinginkan: Strategi DR ditentukan dan diimplementasikan untuk setiap beban kerja agar beban kerja dapat mencapai sasaran DR. Strategi DR antara beban kerja menggunakan pola yang dapat digunakan kembali (seperti strategi yang telah dijelaskan sebelumnya),

Anti-pola umum:

- Mengimplementasikan prosedur pemulihan yang tidak konsisten untuk beban kerja dengan sasaran DR yang serupa.
- Membiarkan strategi DR diimplementasikan secara ad-hoc saat bencana terjadi.
- Tidak memiliki rencana untuk pemulihan bencana.
- Dependensi pada operasi bidang kontrol selama pemulihan.

Manfaat menjalankan praktik terbaik ini:

- Dengan strategi pemulihan yang ditentukan, Anda dapat menggunakan prosedur tes dan peralatan umum.
- Menggunakan strategi pemulihan yang ditentukan akan meningkatkan penyebaran pengetahuan antara tim dan implementasi DR pada beban kerja milik mereka.

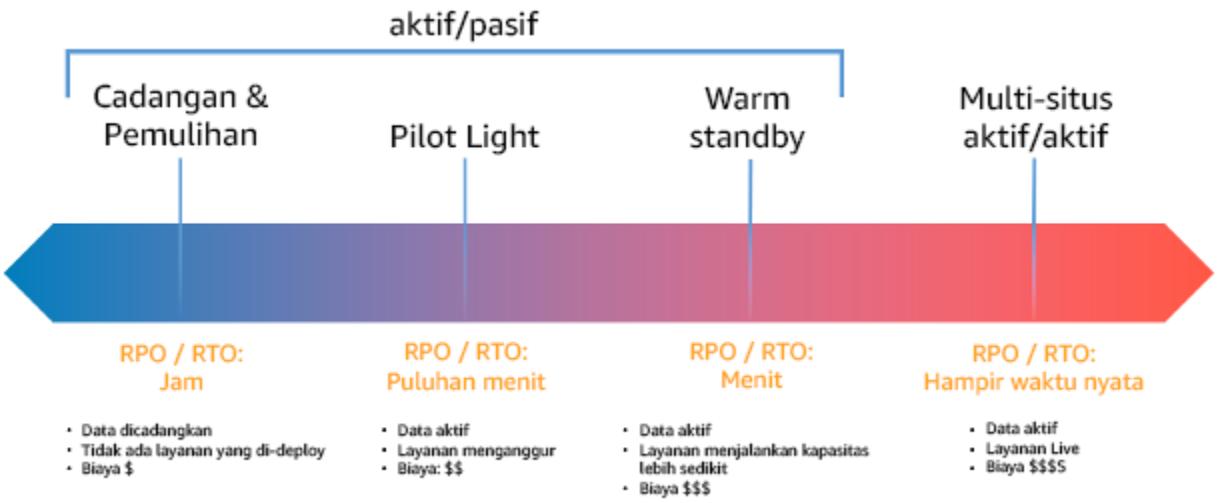
Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Tinggi. Tanpa strategi DR yang direncanakan, diimplementasikan, dan diuji, Anda akan kesulitan mencapai sasaran pemulihan ketika bencana terjadi.

Panduan implementasi

Strategi DR mengandalkan kemampuan untuk mempertahankan beban kerja di situs pemulihan jika lokasi utama tidak dapat menjalankan beban kerja. Sasaran pemulihan yang paling umum adalah RTO dan RPO, seperti yang didiskusikan dalam REL13-BP01 Menetapkan sasaran pemulihan untuk waktu henti dan kehilangan data.

Strategi DR di beberapa Zona Ketersediaan (AZ) dalam Wilayah AWS tunggal, dapat menyediakan mitigasi bencana seperti kebakaran, banjir, dan pemadaman listrik besar-besaran. Anda dapat menggunakan strategi DR yang menggunakan beberapa Wilayah jika memang perlu mengimplementasikan perlindungan terhadap peristiwa yang membuat beban kerja tidak dapat dijalankan di Wilayah AWS.

Anda harus memilih salah satu dari strategi berikut saat merancang strategi DR di beberapa Wilayah. Mereka terdaftar dalam urutan peningkatan biaya dan kompleksitas, dan penurunan urutan RTO dan RPO. Wilayah Pemulihan mengacu pada Wilayah AWS selain dari yang utama yang digunakan untuk beban kerja Anda.



Gambar 17: Strategi pemulihan bencana (DR)

- Cadangkan dan pulihkan (RPO dalam hitungan jam, RTO dalam 24 jam atau kurang): Cadangkan data dan aplikasi Anda ke dalam Wilayah pemulihan. Menggunakan pencadangan otomatis atau berkelanjutan dapat mengaktifkan pemulihan titik waktu (PITR), yang dalam beberapa kasus dapat menurunkan RPO hingga 5 menit dalam beberapa kasus. Saat terjadi bencana, Anda akan melakukan deployment infrastruktur (menggunakan infrastruktur sebagai kode untuk mengurangi RTO), melakukan deploymennt kode, dan memulihkan data yang dicadangkan untuk memulihkan dari bencana di Wilayah pemulihan.
- Pilot light (RPO dalam menit, RTO dalam kelipatan sepuluh menit): Sediakan salinan infrastruktur beban kerja inti di Wilayah pemulihan. Replikasikan data ke Wilayah pemulihan dan buat cadangan di sana. Sumber daya yang diperlukan untuk mendukung replikasi dan pencadangan data, misalnya basis data dan penyimpanan objek, selalu aktif. Elemen lainnya seperti server aplikasi atau komputasi nirserver tidak di-deploy, tetapi dapat dibuat saat dibutuhkan dengan kode aplikasi dan konfigurasi yang diperlukan.
- Warm standby (RPO dalam hitungan detik, RTO dalam hitungan menit): Mengaktifkan versi yang diturunkan tetapi berfungsi sepenuhnya dari beban kerja yang selalu dijalankan di Wilayah pemulihan. Sistem yang vital untuk bisnis sepenuhnya digandakan dan selalu aktif, tetapi dengan armada yang diturunkan skalanya. Data direplikasi dan berada dalam Wilayah pemulihan. Ketika pemulihan diperlukan, sistem dinaikkan skalanya dengan cepat untuk menangani beban produksi. Semakin warm standby dinaikkan skalanya, akan semakin rendah pengandalan RTO dan bidang kontrol. Saat skala sesuai sepenuhnya, ini disebut sebagai hot standby.

- Multi-Wulayah (multi-lokasi) aktif-aktif (RPO mendekati nol, RTO berpotensi nol): Beban kerja Anda di-deploy ke, dan aktif menangani lalu lintas dari, beberapa Wilayah AWS. Strategi ini perlu menyinkronkan data di seluruh Wilayah. Konflik potensial yang disebabkan oleh menulis catatan yang sama di dua replika wilayah yang berbeda harus dihindari atau ditangani, karena bisa menjadi kompleks. Replikasi data bermanfaat untuk sinkronisasi data dan akan melindungi Anda terhadap beberapa jenis bencana, tetapi tidak melindungi terhadap kerusakan atau kehilangan data kecuali solusi juga disertai opsi untuk pemulihan titik waktu.

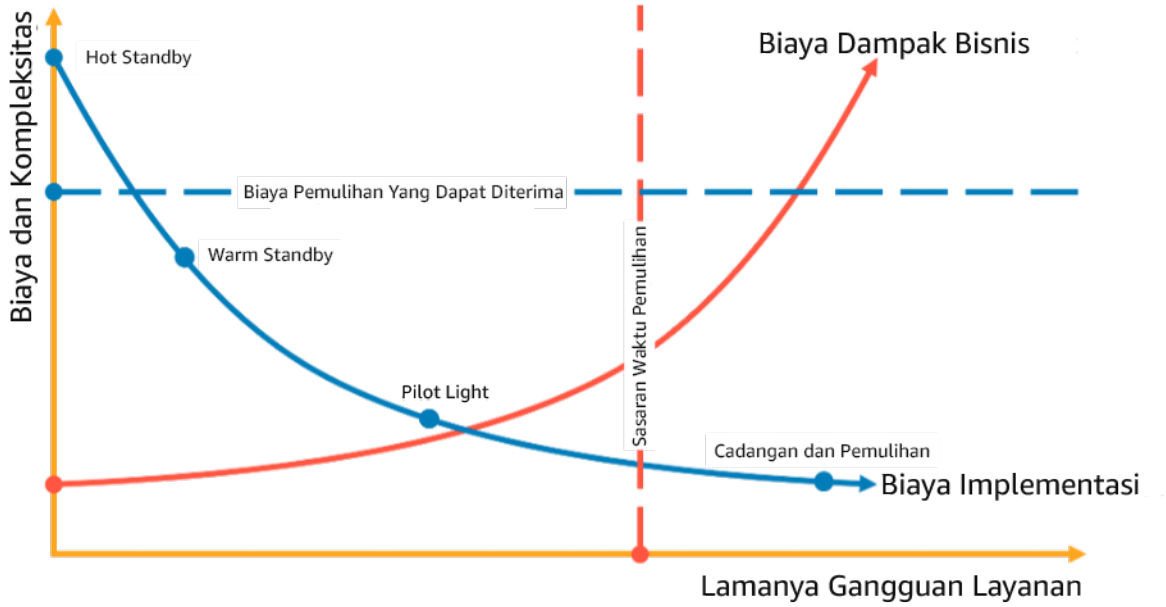
> **Note**
>
> Perbedaan antara pilot light dan warm standby terkadang sulit dimengerti. Keduanya menyertakan lingkungan di Wilayah pemulihan dengan salinan aset wilayah utama. Perbedaannya adalah pilot light tidak dapat memproses permintaan tanpa lebih dulu melakukan tindakan tambahan, sedangkan warm standby dapat menangani lalu lintas (pada kapasitas yang dikurangi) dengan cepat. Pilot light mengharuskan Anda mengaktifkan server, menaikkan skala, dan mungkin mengharuskan Anda melakukan deployment infrastruktur tambahan (bukan inti). Sementara itu, warm standby hanya meminta Anda untuk menaikkan skala (semuanya sudah di-deploy dan dijalankan). Pilih berdasarkan kebutuhan RTO dan RPO Anda.
>
> Apabila ada kekhawatiran tentang biaya, dan Anda ingin mencapai sasaran RPO dan RTO yang serupa dengan yang ditetapkan dalam strategi warm standby, Anda dapat mempertimbangkan solusi cloud-native, seperti AWS Elastic Disaster Recovery, yang akan mengambil pendekatan pilot light dan menawarkan target RPO dan RTO lebih baik.

Langkah-langkah implementasi

1. Tentukan strategi DR yang akan memenuhi persyaratan pemulihan untuk beban kerja ini.

   Saat memilih strategi DR, Anda harus memilih antara mengurangi waktu henti dan kehilangan data (RTO dan RPO) dan meningkatkan biaya dan kompleksitas untuk mengimplementasikan strategi, atau sebaliknya. Sebaiknya hindari strategi yang lebih sulit dari yang dibutuhkan, karena hal ini akan menambah biaya yang tidak perlu.

   Misalnya, dalam diagram berikut, bisnis telah menentukan RTO maksimum yang diizinkan serta batas yang dapat digunakan pada strategi pemulihan layanan. Berdasarkan sasaran bisnis, strategi DR pilot light atau warm standby akan memenuhi kriteria biaya dan RTO.



Gambar 18: Pemilihan strategi DR berdasarkan RTO dan biaya

Untuk mempelajari lebih lanjut, lihat Business Continuity Plan (BCP).

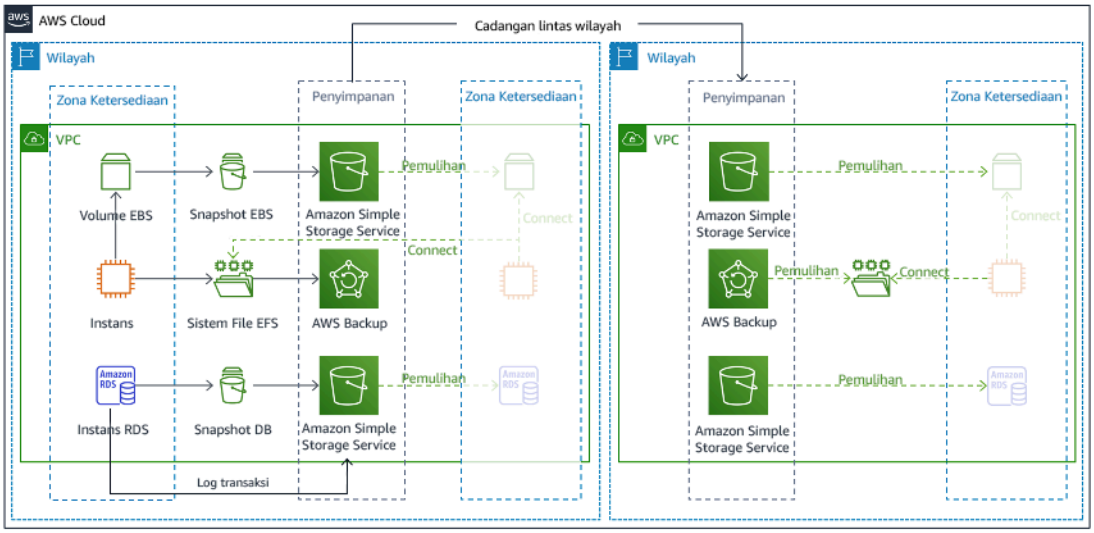
2. Tinjau pola tentang bagaimana strategi DR yang dipilih dapat diimplementasikan.

   Langkah ini digunakan untuk memahami cara Anda mengimplementasikan strategi yang dipilih. Strategi dijelaskan menggunakan Wilayah AWS sebagai situs utama dan pemulihan. Namun, Anda juga dapat memilih untuk menggunakan Zona Ketersediaan dalam Wilayah tunggal sebagai strategi DR, yang menggunakan beberapa elemen dari berbagai strategi tersebut.

   Dalam langkah berikut ini, Anda dapat menerapkan strategi pada beban kerja spesifik Anda.

   Pencadangan dan pemulihan

   Cadangkan dan pulihkan adalah strategi yang tidak terlalu kompleks untuk diimplementasikan, tetapi akan memerlukan waktu dan usaha lebih untuk mengembalikan beban kerja, sehingga RTO dan RPO menjadi lebih tinggi. Sebaiknya selalu buat cadangan data, dan salin cadangan tersebut ke situs lain (misalnya Wilayah AWS lain).

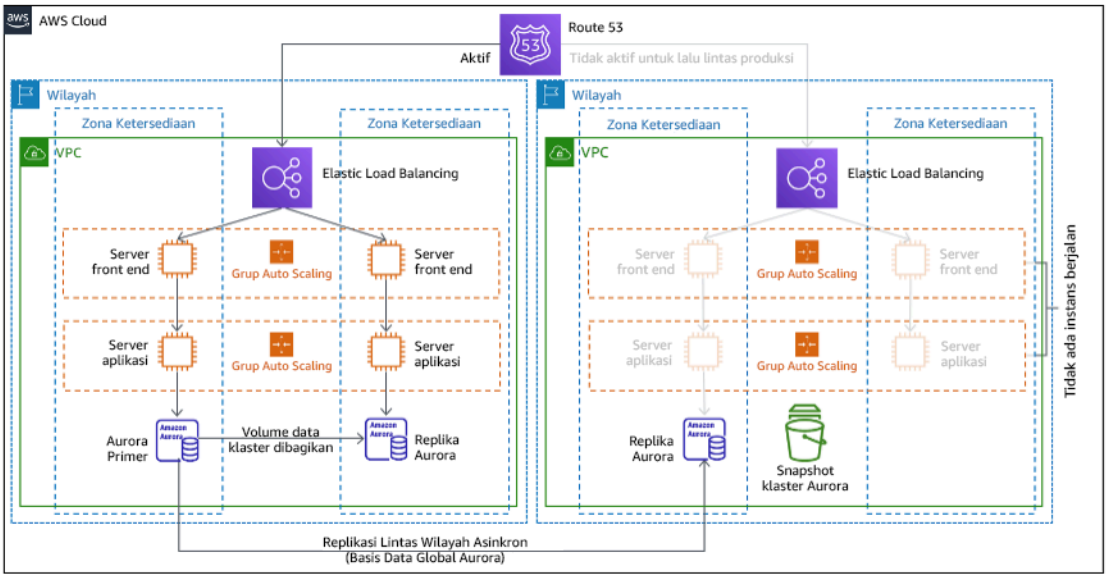


Gambar 19: Arsitektur pencadangan dan pemulihan

Untuk rincial lebih lanjut pada strategi ini lihat [Arsitektur Pemulihan Bencana (DR) di AWS, Bagian II: Pencadangan dan Pemulihan dengan Pemulihan Cepat].

Pilot light

Dengan pendekatan pilot light, Anda mereplikasi data dari Wilayah utama ke Wilayah pemulihan. Sumber daya inti yang digunakan untuk infrastruktur beban kerja di-deploy di Wilayah pemulihan. Namun, sumber daya tambahan dan dependensi lainnya masih diperlukan untuk membuat tumpukan fungsional ini. Misalnya, dalam gambar 20, tidak ada instans komputasi yang di-deploy.

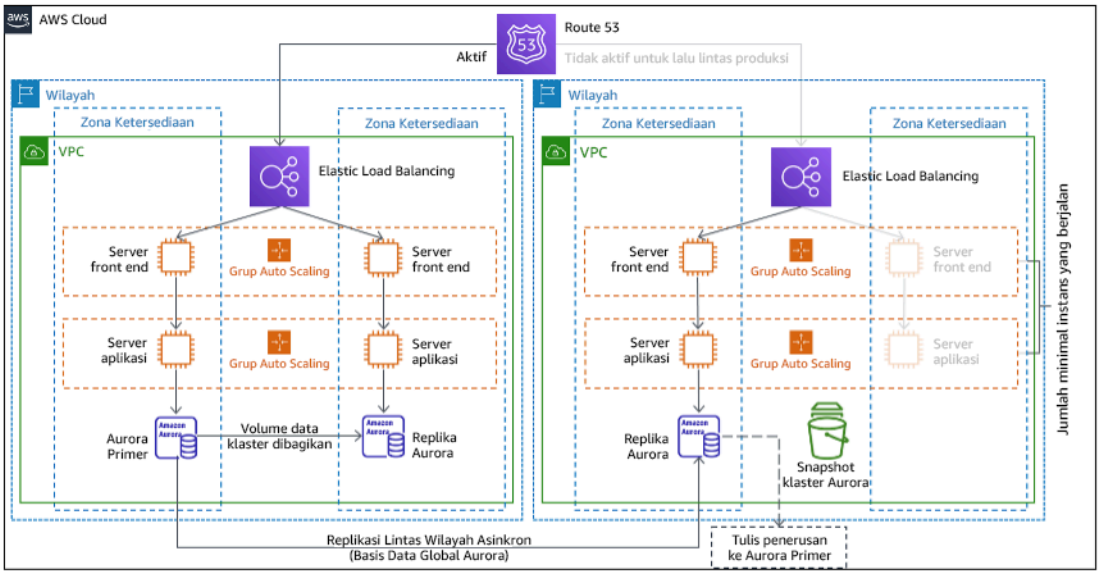


Gambar 20: Arsitektur pilot light

Untuk detail lebih lanjut tentang strategi ini, lihat Arsitektur Pemulihan Bencana (DR) di AWS, Bagian III: Pilot Light and Warm Standby.

Warm standby

Pendekatan warm standby melibatkan memastikan ada salinan lingkungan produksi yang skalanya diturunkan tetapi berfungsi sepenuhnya di Wilayah lainnya. Pendekatan ini memperpanjang konsep pilot light dan mempercepat waktu pemulihan karena beban kerja selalu aktif di Wilayah lainnya. Jika Wilayah pemulihan di-deploy pada kapasitas penuh, hal ini disebut dengan hot standby.



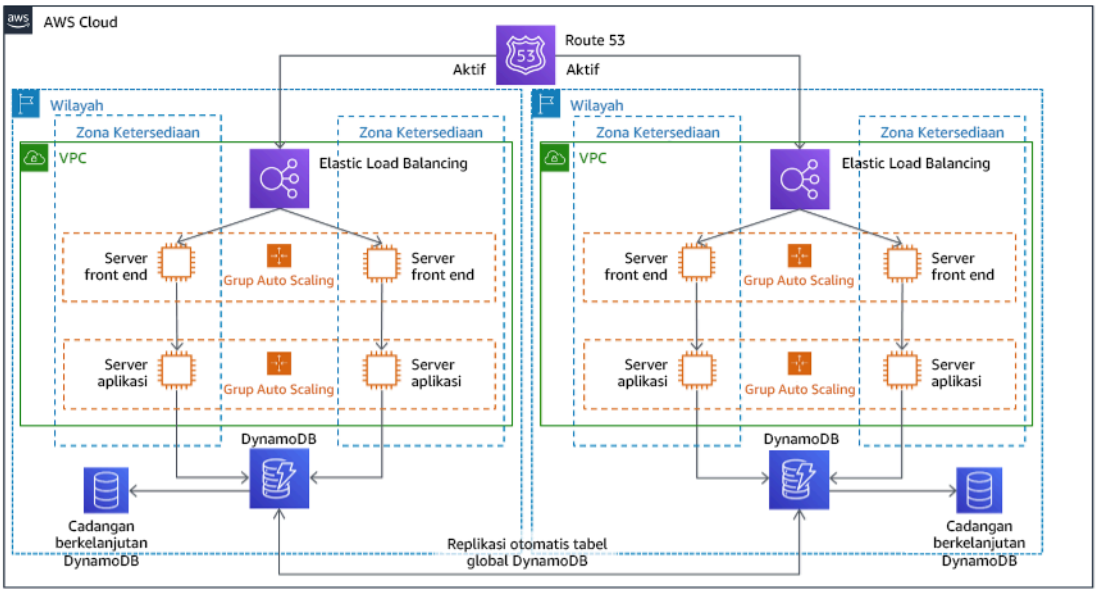
Gambar 21: Arsitektur warm standby

Saat menggunakan warm standby atau pilot light, Anda perlu menaikkan skala sumber daya di Wilayah pemulihan. Untuk memverifikasi kapasitas yang tersedia bila diperlukan, pertimbangkan penggunaan untuk reservasi kapasitas untuk instans EC2. Jika menggunakan AWS Lambda, konkurensi yang disediakan dapat menyediakan lingkungan runtime sehingga siap untuk segera merespons invokasi fungsi Anda.

Untuk detail lebih lanjut tentang strategi ini, lihat Arsitektur Pemulihan Bencana (DR) di AWS, Bagian III: Pilot Light and Warm Standby.

Multi-situs aktif/aktif

Anda dapat menjalankan beban kerja secara berkelanjutan di beberapa Wilayah sebagai bagian dari strategi multi-situs aktif/aktif. Multi-situs aktif/aktif menjalankan lalu lintas dari semua wilayah ke wilayah tempatnya di-deploy. Konsumen dapat memilih strategi ini untuk alasan selain dari DR. Strategi ini dapat digunakan untuk meningkatkan ketersediaan, atau saat melakukan deployment beban kerja ke audiens global (untuk menempatkan titik akhir lebih dekat dengan pengguna dan/atau melakukan deployment tumpukan yang dilokalkan untuk audiens di wilayah tersebut). Sebagai strategi DR, jika beban kerja tidak dapat didukung di salah satu dari Wilayah AWS tempatnya di-deploy, Wilayah tersebut dievakuasi, dan Wilayah sisanya digunakan untuk mempertahankan ketersediaan. Multi-situs aktif/aktif adalah strategi DR yang paling sulit dioperasikan, dan sebaiknya hanya dipilih saat persyaratan bisnis mengharuskannya.



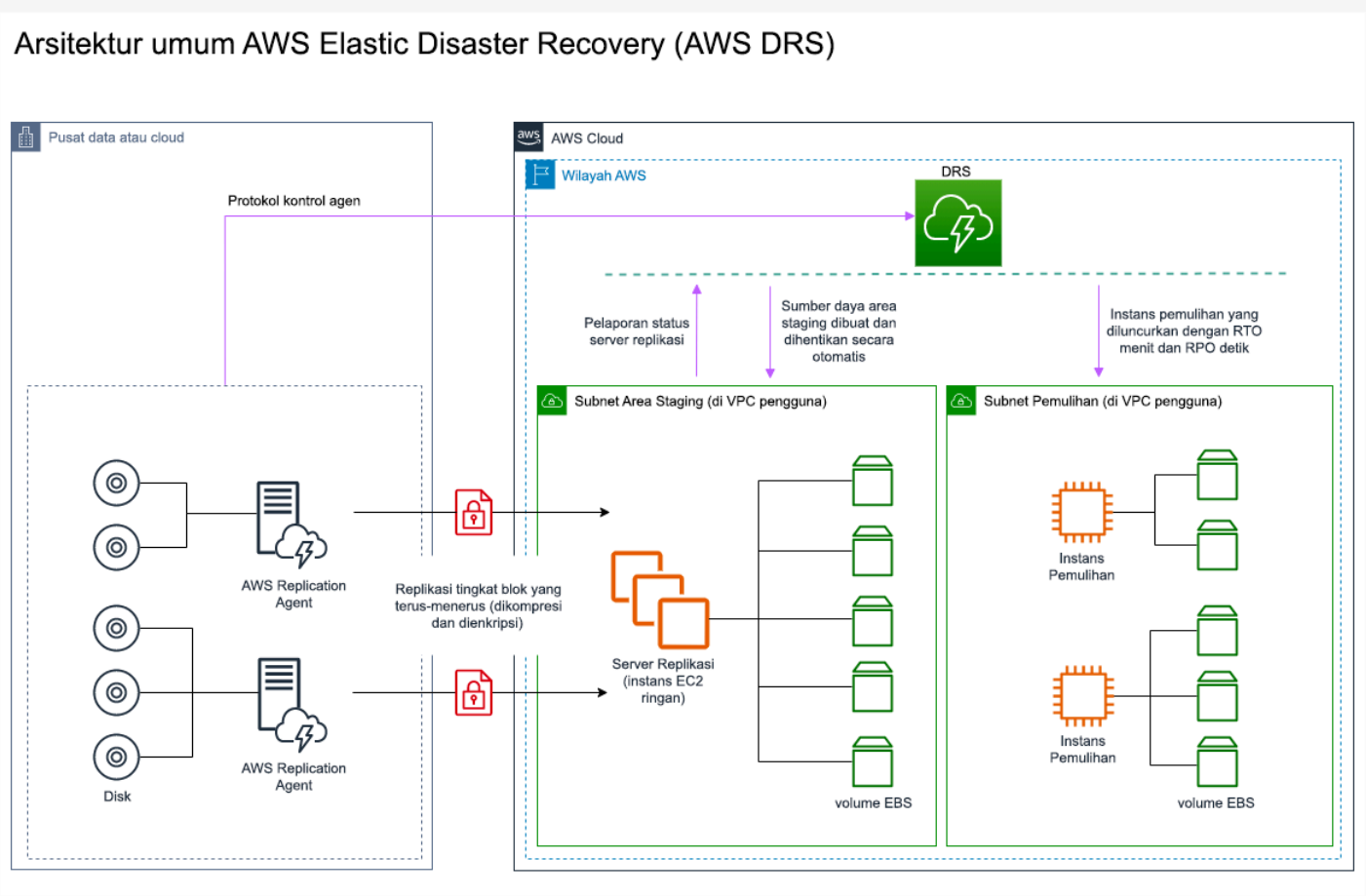
Gambar 22: Arsitektur multi-situs aktif/aktif

Untuk detail lebih lanjut tentang strategi ini, lihat Arsitektur Pemulihan Bencana (DR) di AWS, Bagian IV: Multi-situs Aktif/Aktif.

AWS Elastic Disaster Recovery

Jika Anda mempertimbangkan lampu pilot atau strategi siaga hangat untuk pemulihan bencana, AWS Elastic Disaster Recovery dapat memberikan pendekatan alternatif dengan manfaat yang lebih baik. Elastic Disaster Recovery dapat menawarkan target RPO dan RTO yang mirip dengan siaga hangat, tetapi mempertahankan pendekatan lampu pilot berbiaya rendah. Elastic Disaster Recovery mereplikasi data Anda dari wilayah utama Anda ke Wilayah pemulihan Anda,

menggunakan perlindungan data berkelanjutan untuk mencapai RPO yang diukur dalam hitungan detik dan RTO yang dapat diukur dalam hitungan menit. Hanya sumber daya yang diperlukan untuk mereplikasi data yang di-deploy di wilayah pemulihan, sehingga menekan biaya tetap rendah, serupa dengan strategi pilot light. Ketika menggunakan Elastic Disaster Recovery, layanan mengoordinasi dan mengatur pemulihan sumber daya komputasi ketika dimulai sebagai bagian dari failover atau latihan.



Gambar 23: arsitektur AWS Elastic Disaster Recovery

Praktik tambahan untuk melindungi data

Dengan semua strategi, Anda juga harus melakukan mitigasi terhadap bencana data. Replikasi data berkelanjutan melindungi Anda terhadap beberapa jenis bencana, tetapi tidak melindungi terhadap kerusakan atau kehilangan data kecuali strategi juga disertai penentuan versi data yang disimpan atau opsi pemulihan titik waktu. Selain replika, Anda juga harus mencadangkan data yang direplikasi di situs pemulihan untuk membuat pencadangan titik waktu.

Menggunakan beberapa Zona Ketersediaan (AZ) dalam Wilayah AWS tunggal

Saat menggunakan beberapa AZ dalam Wilayah tunggal, implementasi DR Anda menggunakan beberapa elemen dari strategi di atas. Anda harus terlebih dahulu membuat arsitektur ketersediaan tinggi (HA) menggunakan beberapa AZ yang ditampilkan dalam Gambar 23. Arsitektur ini menggunakan pendekatan aktif/aktif multi-situs, karena instans Amazon EC2 dan Penyeimbang Beban Elastis memiliki sumber daya yang digunakan di beberapa AZ yang secara aktif memberikan permintaan. Arsitektur juga menunjukkan hot standby, di mana jika instans Amazon RDS utama gagal (atau AZ itu sendiri gagal), maka instans siaga dipromosikan ke primer.

Gambar 24: Arsitektur Multi-AZ

Selain arsitektur HA ini, Anda perlu menambahkan cadangan data yang dibutuhkan untuk menjalankan beban kerja. Ini sangat penting untuk data yang dibatasi pada satu zona seperti volume Amazon EBS atau cluster Amazon Redshift. Jika sebuah AZ gagal, Anda perlu memulihkan data ini ke AZ lainnya. Jika memungkinkan, Anda perlu menyalin cadangan data ke Wilayah AWS sebagai lapisan perlindungan tambahan.

Pendekatan alternatif yang kurang umum untuk DR Wilayah tunggal dan Multi-AZ diilustrasikan dalam postingan blog, Membangun aplikasi yang sangat tangguh menggunakan Pengontrol Pemulihan Aplikasi Amazon, Bagian 1: tumpukan Wilayah Tunggal. Strategi yang digunakan di sini adalah mempertahankan isolasi sebanyak mungkin di antara AZ, seperti bagaimana Wilayah dioperasikan. Dengan menggunakan strategi alternatif ini, Anda dapat memilih pendekatan aktif/aktif atau aktif/pasif.

> **Note**
>
> Beberapa beban kerja memiliki persyaratan residensi data peraturan. Jika ini diterapkan untuk beban kerja di lokalitas yang saat ini hanya memiliki satu Wilayah AWS, maka multi-Wilayah tidak akan sesuai untuk kebutuhan bisnis. Strategi multi-AZ memberikan perlindungan yang baik terhadap sebagian besar bencana.

3. Evaluasikan sumber daya beban kerja, dan seperti apa konfigurasinya di Wilayah pemulihan sebelum failover (selama operasi normal).

   Untuk infrastruktur dan sumber daya AWS, gunakan infrastruktur sebagai kode seperti AWS CloudFormation atau alat pihak ketiga seperti Hashicorp Terraform. Untuk melakukan deployment di beberapa akun dan Wilayah dengan operasi tunggal, Anda dapat menggunakan AWS CloudFormation StackSets. Untuk strategi Multi-situs aktif/aktif dan Hot Standby, infrastruktur yang di-deploy di Wilayah pemulihan memiliki sumber daya yang sama seperti Wilayah utama. Untuk strategi Pilot Light dan Warm Standby, infrastruktur yang di-deploy memerlukan tindakan tambahan agar berubah menjadi siap produksi. Dengan menggunakan parameter dan logika bersyarat CloudFormation, Anda dapat mengontrol apakah tumpukan yang diterapkan aktif atau siaga dengan satu templat. Ketika menggunakan Elastic Disaster Recovery, layanan akan mereplikasi dan mengatur pemulihan konfigurasi aplikasi dan sumber daya komputasi.

   Semua strategi DR mengharuskan sumber data dicadangkan di dalam Wilayah AWS, dan kemudian cadangan tersebut disalin ke Wilayah pemulihan. AWS Backup akan menyediakan tampilan tersentralisasi di mana Anda dapat mengonfigurasi, menjadwalkan, dan memantau cadangan untuk sumber daya ini. Untuk Pilot Light, Warm Standby, dan Multi-situs aktif/aktif, Anda juga harus mereplikasi data dari Wilayah utama ke sumber daya data di Wilayah pemulihan, seperti instans DB Amazon Relational Database Service (Amazon RDS) atau tabel Amazon DynamoDB. Dengan demikian, sumber data ini aktif dan siap menangani permintaan di Wilayah pemulihan.

   Untuk mempelajari lebih lanjut tentang cara layanan-layanan AWS beroperasi di seluruh Wilayah, lihat seri blog ini tentang Membuat Aplikasi Multi-Wilayah dengan Layanan AWS.

4. Tentukan dan implementasikan cara Anda mempersiapkan Wilayah untuk failover saat dibutuhkan (selama peristiwa bencana).

   Untuk multi-situs aktif/aktif, failover berarti mengevakuasi Wilayah dan mengandalkan Wilayah aktif yang tersisa. Secara umum, Wilayah tersebut siap menerima lalu lintas. Untuk strategi Pilot Light

dan Warm Standby, tindakan pemulihan perlu mencakup deployment sumber daya yang hilang, seperti instans EC2 dalam Gambar 20, juga sumber daya yang hilang lainnya.

Untuk semua strategi di atas, Anda mungkin perlu mengubah instans hanya-baca basis data menjadi instans baca/tulis.

Untuk pencadangan dan pemulihan, pemulihan data dari cadangan menghasilkan sumber daya untuk data tersebut seperti volume EBS, instans RDS DB, dan tabel DynamoDB. Anda juga perlu memulihkan infrastruktur dan melakukan deployment kode. Anda dapat menggunakan AWS Backup untuk memulihkan data di Wilayah pemulihan. Lihat REL09-BP01 Mengidentifikasi dan mencadangkan data yang perlu dicadangkan, atau melakukan reproduksi ulang data dari sumber untuk detail selengkapnya. Membangun kembali infrastruktur termasuk membuat sumber daya seperti instans EC2 selain Amazon Virtual Private Cloud (Amazon VPC), subnet, dan grup keamanan yang diperlukan. Anda dapat mengotomatiskan banyak proses pemulihan. Untuk mempelajari caranya, silakan lihat posting blog ini.

5. Tentukan dan implementasikan cara Anda akan merutekan kembali lalu lintas ke failover saat dibutuhkan (selama peristiwa bencana).

   Operasi failover ini dapat dimulai secara otomatis dan manual. Failover yang dimulai secara otomatis berdasarkan pemeriksaan kondisi atau alarm harus digunakan dengan hati-hati karena failover yang tidak perlu (alarm palsu) dapat dikenakan biaya seperti ketidaktersediaan dan kehilangan data. Oleh karena itu, Failover yang dimulai secara manual sering digunakan. Dalam kasus ini, Anda masih harus mengotomatiskan langkah failover, sehingga inisiasi manual akan seperti menekan tombol.

   Ada beberapa opsi manajemen lalu lintas yang perlu dipertimbangkan saat menggunakan layanan AWS. Salah satu opsinya adalah menggunakan Amazon Route 53. Dengan menggunakan Amazon Route 53, Anda dapat mengaitkan beberapa titik akhir IP di satu Wilayah AWS atau lebih dengan nama domain Route 53. Untuk mengimplementasikan failover yang dimulai secara manual, Anda dapat menggunakan Pengontrol Pemulihan Aplikasi Amazon, yang menyediakan API bidang data dengan ketersediaan yang tinggi untuk mengalihkan lalu lintas ke Wilayah pemulihan. Saat mengimplementasikan failover, gunakan operasi bidang data dan hindari bidang kontrol yang dideskripsikan di REL11-BP04 Mengandalkan bidang data dan bukan bidang kontrol selama pemulihan.

   Untuk mempelajari lebih lanjut tentang ini dan opsi lainnya, lihat bagian ini dari Laporan Resmi Pemulihan Bencana.

6. Rancang rencana terkait bagaimana beban kerja akan failback.

Failback adalah saat Anda mengembalikan operasi beban kerja ke Wilayah utama, setelah bencana berakhir. Penyediaan infrastruktur dan kode untuk Wilayah utama umumnya mengikuti langkah yang sama yang digunakan saat memulai, dengan mengandalkan infrastruktur sebagai kode dan pipeline deployment kode. Tantangan failback adalah mengembalikan penyimpanan data, dan memastikan konsistensi dengan Wilayah pemulihan dalam operasi.

Dalam status failed over, basis data dalam Wilayah pemulihan bersifat waktu nyata dan memiliki data terbaru. Tujuannya adalah untuk menyinkronkan kembali dari Wilayah pemulihan ke Wilayah utama, memastikannya tetap terbaru.

Hal ini dilakukan secara otomatis untuk beberapa layanan AWS. Jika menggunakan tabel global Amazon DynamoDB, meskipun tabel di Wilayah utama menjadi tidak tersedia, saat kembali online, DynamoDB akan melanjutkan penulisan yang tertunda. Jika menggunakan Basis Data Global Amazon Aurora dan menggunakan failover terencana dan terkelola, topologi replikasi Aurora basis data global yang ada dipertahankan. Dengan demikian, instans baca/tulis sebelumnya di Wilayah utama akan menjadi replika dan menerima pembaruan dari Wilayah pemulihan.

Dalam kasus saat ini tidak dibuat otomatis, Anda perlu menetapkan ulang basis data di Wilayah utama sebagai replika dari basis data di Wilayah pemulihan. Dalam banyak kasus, ini akan melibatkan penghapusan basis data utama yang lama dan membuat replika yang baru.

Setelah failover, jika Anda dapat tetap menjalankannya di Wilayah pemulihan, pertimbangkan membuat ini menjadi Wilayah utama yang baru. Anda masih harus melakukan semua langkah di atas untuk membuat Wilayah utama sebelumnya menjadi Wilayah pemulihan. Beberapa organisasi melakukan rotasi terjadwal, menukar Wilayah utama dan pemulihan secara berkala (misalnya setiap tiga bulan).

Semua langkah yang diperlukan untuk failover dan failback harus diperiksa di playbook yang tersedia untuk semua anggota tim dan ditinjau secara berkala.

Ketika menggunakan Elastic Disaster Recovery, layanan akan membantu mengatur dan mengotomatiskan proses failback. Untuk detail selengkapnya, lihat Melakukan failback.

Tingkat upaya untuk Rencana Implementasi: Tinggi

Sumber daya

Praktik-praktik terbaik terkait:

- the section called “REL09-BP01 Mengidentifikasi dan mencadangkan data yang perlu dicadangkan, atau melakukan reproduksi ulang data dari sumber”
- the section called “REL11-BP04 Mengandalkan bidang data dan bukan bidang kontrol selama pemulihan”
- the section called “REL13-BP01 Menetapkan sasaran pemulihan untuk waktu henti dan kehilangan data”

Dokumen terkait:

- Blog Arsitektur AWS: Seri Pemulihan Bencana
- Pemulihan Bencana Beban Kerja di AWS: Pemulihan di Cloud (Laporan Resmi AWS)
- Opsi pemulihan bencana di cloud
- Bangun solusi backend aktif-aktif nirserver multi-wilayah dalam satu jam
- Backend nirserver multi-wilayah — dimuat ulang
- RDS: Mereplikasi Replika Baca di Seluruh Wilayah
- Route 53: Mengonfigurasi failover DNS
- S3: Replika Lintas-Wilayah
- Apa itu AWS Backup?
- Apa itu Pengontrol Pemulihan Aplikasi Amazon?
- AWS Elastic Disaster Recovery
- HashiCorp Terraform: Memulai - AWS
- Partner APN: partner yang dapat membantu Anda mengatasi pemulihan bencana
- AWS Marketplace: produk yang dapat digunakan untuk pemulihan bencana

Video terkait:

- Pemulihan Bencana Beban Kerja di AWS
- AWS re:Invent 2018: Pola Arsitektur untuk Aplikasi Aktif-Aktif Multi-Wilayah (ARC209-R2)
- Memulai AWS Elastic Disaster Recovery | Amazon Web Services

REL13-BP03 Menguji implementasi pemulihan bencana untuk memvalidasi implementasi

Secara rutin uji failover ke situs pemulihan Anda untuk memastikan operasi yang baik dan RTO serta RPO terpenuhi.

Anti-pola umum:

- Tidak pernah melakukan failover di lingkungan produksi.

Manfaat menerapkan praktik terbaik ini: Pengujian rencana pemulihan bencana secara rutin akan memverifikasi bahwa rencana tersebut akan berfungsi saat diperlukan, dan tim Anda tahu cara mengeksekusi strategi.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Tinggi

Panduan implementasi

Pola untuk dihindari adalah mengembangkan jalur pemulihan yang sangat jarang dilakukan. Misalnya, Anda mungkin memiliki penyimpanan data sekunder yang digunakan untuk kueri hanya-baca. Saat Anda menulis ke penyimpanan data dan penyimpanan primer gagal, Anda mungkin ingin melakukan failover ke penyimpanan data sekunder. Jika Anda tidak sering menguji failover ini, Anda mungkin akan mendapati bahwa asumsi Anda tentang kemampuan penyimpanan data sekunder ternyata salah. Kapasitas sekunder, yang selama ini mungkin mencukupi saat terakhir Anda uji, mungkin sudah tidak mampu mentoleransi beban di bawah skenario ini. Pengalaman kami menunjukkan bahwa satu-satunya pemulihan kesalahan yang dapat diterapkan adalah jalur yang sering Anda uji. Inilah alasan memiliki sedikit jalur pemulihan adalah yang terbaik. Anda dapat membuat pola pemulihan dan mengujinya secara rutin. Jika Anda memiliki jalur pemulihan yang kompleks atau kritis, Anda tetap perlu secara rutin melatih kegagalan tersebut dalam lingkungan produksi agar Anda yakin bahwa jalur pemulihan tersebut berfungsi. Pada contoh yang baru saja kita bahas, Anda harus melakukan failover ke penyimpanan siaga secara rutin, terlepas ada tidaknya kebutuhan.

Langkah-langkah implementasi

1. Rekayasa beban kerja Anda untuk pemulihan. Uji jalur pemulihan Anda secara rutin. Komputasi yang berorientasi pada pemulihan mengidentifikasi karakteristik dalam sistem yang meningkatkan pemulihan: isolasi dan redundansi, kemampuan di seluruh sistem untuk membatalkan perubahan, kemampuan untuk memantau dan menentukan kondisi, kemampuan untuk menyediakan

diagnostik, pemulihan otomatis, desain modular, dan kemampuan untuk memulai ulang. Latih jalur pemulihan untuk memverifikasi bahwa Anda dapat menyelesaikan pemulihan dalam waktu yang ditentukan ke status yang ditentukan. Gunakan runbook selama pemulihan ini untuk mendokumentasikan masalah dan menemukan solusinya sebelum pengujian berikutnya.

2. Untuk beban kerja berbasis Amazon EC2, gunakan AWS Elastic Disaster Recovery untuk menerapkan dan meluncurkan instans latihan untuk strategi DR Anda. AWS Elastic Disaster Recovery menyediakan kemampuan untuk menjalankan latihan secara efisien, yang akan membantu Anda mempersiapkan peristiwa failover. Anda juga dapat sering-sering meluncurkan instans menggunakan Pemulihan Bencana Elastis untuk tujuan pengujian dan latihan tanpa mengarahkan ulang lalu lintas.

Sumber daya

Dokumen terkait:

- Partner APN: partner yang dapat membantu Anda mengatasi pemulihan bencana
- AWS Blog Arsitektur : Seri Pemulihan Bencana
- AWS Marketplace: produk yang dapat digunakan untuk pemulihan bencana
- AWS Elastic Disaster Recovery
- Pemulihan Bencana Beban Kerja di AWS: Pemulihan di Cloud (Laporan Resmi AWS)
- AWS Elastic Disaster Recovery Mempersiapkan Failover
- Proyek komputasi Berkeley/Stanford berorientasi pemulihan
- Apa itu Simulator Injeksi Kesalahan AWS?

Video terkait:

- AWS re:Invent 2018: Pola Arsitektur untuk Aplikasi Aktif-Aktif Multi-Wilayah
- AWS re:Invent 2019: Pencadangan dan pemulihan serta solusi pemulihan bencana dengan AWS

REL13-BP04 Mengelola penyimpangan konfigurasi di lokasi atau Wilayah Pemulihan Bencana (DR)

Untuk melakukan prosedur pemulihan bencana (DR) yang berhasil, beban kerja Anda harus dapat kembali beroperasi normal dengan cepat tanpa kehilangan fungsionalitas atau data yang relevan

setelah lingkungan DR aktif dan siap digunakan. Untuk mencapai tujuan ini, penting untuk menjaga konsistensi infrastruktur, data, dan konfigurasi antara lingkungan DR Anda dan lingkungan primer.

Hasil yang diinginkan: Konfigurasi dan data situs pemulihan bencana Anda setara dengan situs primer, sehingga memudahkan pemulihan yang cepat dan menyeluruh saat dibutuhkan.

Anti-pola umum:

- Anda tidak memperbarui lokasi pemulihan ketika ada perubahan pada lokasi primer, sehingga menghasilkan konfigurasi usang yang dapat menghambat upaya pemulihan.
- Anda tidak mempertimbangkan potensi keterbatasan seperti perbedaan layanan antara lokasi primer dan pemulihan, sehingga dapat menyebabkan kegagalan tak terduga selama failover.
- Anda mengandalkan proses manual untuk memperbarui dan menyinkronkan lingkungan DR, sehingga meningkatkan risiko kesalahan manusia dan inkonsistensi.
- Anda gagal mendeteksi penyimpangan konfigurasi, sehingga Anda keliru menganggap bahwa situs DR siap sebelum insiden.

Manfaat menjalankan praktik terbaik ini: Konsistensi antara lingkungan DR dan lingkungan primer secara signifikan meningkatkan kemungkinan keberhasilan pemulihan setelah insiden dan mengurangi risiko prosedur pemulihan yang gagal.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Tinggi

Panduan implementasi

Pendekatan komprehensif untuk manajemen konfigurasi dan kesiapan failover dapat membantu Anda memverifikasi bahwa situs DR secara konsisten diperbarui dan siap untuk mengambil alih jika terjadi kegagalan situs primer.

Untuk mencapai konsistensi antara lingkungan primer dan pemulihan bencana (DR) Anda, validasikan bahwa pipeline pengiriman Anda mendistribusikan aplikasi ke situs primer dan juga situs DR Anda. Terapkan perubahan ke situs DR setelah periode evaluasi yang sesuai (juga dikenal sebagai deployment bertahap) untuk mendeteksi masalah di situs primer dan menghentikan deployment sebelum masalah tersebut menyebar. Terapkan pemantauan untuk mendeteksi peyimpangan konfigurasi, dan lacak perubahan serta kepatuhan di berbagai lingkungan Anda. Lakukan remediasi otomatis di situs DR agar tetap konsisten dan siap untuk mengambil alih jika terjadi insiden.

Langkah-langkah implementasi

1. Validasi bahwa wilayah DR berisi fitur dan layanan AWS yang diperlukan untuk keberhasilan pelaksanaan rencana DR Anda.
2. Gunakan infrastruktur sebagai kode (IaC). Jaga agar templat infrastruktur produksi dan konfigurasi aplikasi Anda tetap akurat, dan terapkan secara teratur ke lingkungan pemulihan bencana Anda. AWS CloudFormation dapat mendeteksi penyimpangan antara konfigurasi yang ditentukan dalam templat CloudFormation Anda dan apa yang sebenarnya di-deploy.
3. Konfigurasikan pipeline CI/CD untuk melakukan deployment pembaruan aplikasi dan infrastruktur ke semua lingkungan, termasuk situs primer dan DR. Solusi CI/CD seperti AWS CodePipeline dapat mengotomatiskan proses deployment, sehingga mengurangi risiko penyimpangan konfigurasi.
4. Lakukan deployment bertahap antara lingkungan primer dan DR. Pendekatan ini memungkinkan pembaruan di-deploy dan diuji terlebih dahulu di lingkungan primer, sehingga masalah di situs primer dapat diisolasi sebelum menyebar ke situs DR. Pendekatan ini mencegah kecacatan didorong ke situs produksi dan DR pada saat yang bersamaan dan menjaga integritas lingkungan DR.
5. Terus pantau konfigurasi sumber daya di lingkungan primer dan DR. Solusi seperti AWS Config dapat membantu menegakkan kepatuhan konfigurasi dan mendeteksi penyimpangan konfigurasi, sehingga membantu menjaga konsistensi konfigurasi di seluruh lingkungan.
6. Implementasikan mekanisme peringatan untuk melacak dan memberitahukan setiap penyimpangan konfigurasi atau gangguan atau keterlambatan replikasi data.
7. Otomatiskan perbaikan penyimpangan konfigurasi yang terdeteksi.
8. Jadwalkan audit dan pemeriksaan kepatuhan rutin untuk memverifikasi keselarasan yang berkelanjutan antara konfigurasi primer dan DR. Peninjauan berkala membantu Anda menjaga kepatuhan terhadap aturan yang ditetapkan dan mengidentifikasi ketidaksesuaian apa pun yang perlu ditangani.
9. Periksa ketidakcocokan dalam kapasitas yang disediakan AWS, kuota layanan, batas throttle, serta perbedaan konfigurasi dan versi.

Sumber daya

Praktik-praktik terbaik terkait:

- REL01-BP01 Mengetahui kuota dan keterbatasan layanan

- REL01-BP02 Mengelola kuota layanan di seluruh akun dan wilayah
- REL01-BP04 Memantau dan mengelola kuota
- REL13-BP03 Menguji implementasi pemulihan bencana untuk memvalidasi implementasi

Dokumen terkait:

- Mengatasi Sumber Daya AWS yang Tidak Patuh oleh Aturan AWS Config
- AWS Systems Manager Automation
- AWS CloudFormation: Mendeteksi perubahan konfigurasi tidak terkelola ke tumpukan dan sumber daya
- AWS CloudFormation: Mendeteksi Penyimpangan di Seluruh Tumpukan CloudFormation
- AWS Systems Manager Automation
- Pemulihan Bencana Beban Kerja di AWS: Pemulihan di Cloud (Laporan Resmi AWS)
- Bagaimana cara mengimplementasikan solusi Manajemen Konfigurasi Infrastruktur di AWS?
- Mengatasi Sumber Daya AWS yang Tidak Patuh oleh Aturan AWS Config

Video terkait:

- AWS re:Invent 2018: Pola Arsitektur untuk Aplikasi Aktif-Aktif Multi-Wilayah (ARC209-R2)

Contoh terkait:

- Registri CloudFormation
- Pemantau Kuota untuk AWS
- Mengimplementasikan perbaikan otomatis atas penyimpangan untuk AWS CloudFormation menggunakan Amazon CloudWatch dan AWS Lambda
- AWS Blog Arsitektur : Seri Pemulihan Bencana
- AWS Marketplace: produk yang dapat digunakan untuk pemulihan bencana
- Melakukan otomatisasi deployment secara aman dan otonom

REL13-BP05 Mengotomatiskan pemulihan

Implementasikan mekanisme pemulihan teruji dan otomatis yang andal, dapat diamati, serta dapat direproduksi untuk mengurangi risiko dan dampak bisnis dari kegagalan.

Hasil yang diinginkan: Anda telah mengimplementasikan alur kerja otomatisasi untuk proses pemulihan yang terdokumentasi dengan baik, terstandardisasi, dan teruji secara menyeluruh. Otomatisasi pemulihan Anda secara otomatis memperbaiki masalah kecil yang menimbulkan risiko rendah kehilangan atau ketidaktersediaan data. Anda dapat dengan cepat menginvokasi proses pemulihan untuk insiden serius, mengamati perilaku perbaikan saat proses tersebut beroperasi, dan menghentikan proses jika Anda mengamati situasi berbahaya atau kegagalan.

Anti-pola umum:

- Anda bergantung pada komponen atau mekanisme yang berada dalam keadaan gagal atau kinerjanya menurun sebagai bagian dari rencana pemulihan Anda.
- Proses pemulihan Anda memerlukan intervensi manual, seperti akses konsol (juga dikenal sebagai click ops).
- Anda secara otomatis memulai prosedur pemulihan dalam situasi yang menimbulkan risiko tinggi kehilangan atau ketidaktersediaan data.
- Anda gagal menyertakan mekanisme untuk membatalkan prosedur pemulihan (seperti kabel Andon atau tombol darurat besar warna merah) yang tidak berfungsi atau yang menimbulkan risiko tambahan.

Manfaat menjalankan praktik terbaik ini:

- Peningkatan keandalan, prediktabilitas, dan konsistensi operasi pemulihan.
- Kemampuan untuk memenuhi sasaran pemulihan yang lebih ketat, termasuk Sasaran Waktu Pemulihan (RTO) dan Sasaran Titik Pemulihan (RPO).
- Mengurangi kemungkinan pemulihan gagal selama suatu insiden.
- Mengurangi risiko kegagalan yang terkait dengan proses pemulihan manual yang rentan terhadap kesalahan manusia.

Tingkat risiko yang terjadi jika praktik terbaik ini tidak diterapkan: Sedang

Panduan implementasi

Untuk menerapkan pemulihan otomatis, Anda memerlukan pendekatan komprehensif yang menggunakan layanan dan praktik terbaik AWS. Untuk memulai, identifikasi komponen penting dan titik kegagalan potensial dalam beban kerja Anda. Kembangkan proses otomatis yang dapat memulihkan beban kerja dan data Anda dari kegagalan tanpa campur tangan manusia.

Kembangkan otomatisasi pemulihan Anda menggunakan prinsip infrastruktur sebagai kode (IaC). Hal ini membuat lingkungan pemulihan Anda konsisten dengan lingkungan sumber dan memungkinkan kontrol versi untuk proses pemulihan Anda. Untuk mengorkestrasi alur kerja pemulihan yang kompleks, pertimbangkan solusi seperti Otomatisasi AWS Systems Manager atau AWS Step Functions.

Otomatisasi proses pemulihan memberikan manfaat yang signifikan dan dapat membantu Anda lebih mudah mencapai Sasaran Waktu Pemulihan (RTO) dan Sasaran Titik Pemulihan (RPO). Namun, otomatisasi tersebut dapat mengalami situasi tak terduga yang dapat membuatnya gagal atau menciptakan risiko baru seperti waktu henti tambahan dan kehilangan data. Untuk mengurangi risiko ini, berikan kemampuan yang dapat dengan cepat menghentikan otomatisasi pemulihan yang sedang berlangsung. Setelah dihentikan, Anda dapat menyelidiki dan mengambil langkah-langkah korektif.

Untuk beban kerja yang didukung, pertimbangkan solusi seperti AWS Elastic Disaster Recovery (AWS DRS) untuk menyediakan failover otomatis. AWS DRS secara terus-menerus mereplikasi mesin Anda (termasuk sistem operasi, konfigurasi status sistem, basis data, aplikasi, dan file) ke dalam area staging di akun Akun AWS target Anda dan Wilayah yang dipilih. Jika terjadi insiden, AWS DRS mengotomatiskan konversi server replika Anda menjadi beban kerja yang disediakan sepenuhnya di Wilayah pemulihan Anda di AWS.

Pemulihan otomatis adalah proses yang perlu dipelihara dan ditingkatkan secara berkelanjutan. Terus uji dan sempurnakan prosedur pemulihan Anda berdasarkan pelajaran yang dipetik, dan tetap ikuti informasi terbaru tentang fitur dan layanan AWS baru yang dapat meningkatkan kemampuan pemulihan Anda.

Langkah-langkah implementasi

1. Rencanakan pemulihan otomatis
    a. Lakukan tinjauan menyeluruh atas arsitektur beban kerja, komponen, dan dependensi Anda untuk mengidentifikasi dan merencanakan mekanisme pemulihan otomatis. Kategorikan dependensi beban kerja Anda ke dalam dependensi mutlak dan relatif. Dependensi mutlak

adalah dependensi yang tanpanya beban kerja tidak dapat beroperasi dan tidak ada pengganti yang dapat disediakan. Dependensi relatif adalah dependensi yang biasanya digunakan oleh beban kerja tetapi dapat diganti dengan sistem atau proses pengganti sementara atau dapat ditangani dengan penurunan terkendali.

b. Tetapkan proses untuk mengidentifikasi dan memulihkan data yang hilang atau rusak.

c. Tentukan langkah-langkah untuk mengonfirmasi kondisi stabil dan pulih setelah tindakan pemulihan selesai.

d. Pertimbangkan tindakan apa pun yang diperlukan untuk membuat sistem yang pulih siap untuk layanan penuh, seperti pra-pemanasan dan pengisian cache.

e. Pikirkan berbagai masalah yang dapat muncul selama proses pemulihan dan cara mendeteksi dan memperbaikinya.

f. Pertimbangkan skenario ketika situs primer dan bidang kontrolnya tidak dapat diakses. Verifikasi bahwa tindakan pemulihan dapat dilakukan secara independen tanpa bergantung pada situs primer. Pertimbangkan solusi seperti Pengontrol Pemulihan Aplikasi (ARC) Amazon untuk mengalihkan lalu lintas tanpa perlu mengubah catatan DNS secara manual.

2. Kembangkan proses pemulihan otomatis

a. Terapkan deteksi kesalahan otomatis dan mekanisme failover untuk pemulihan tanpa intervensi manual. Buat dasbor misalnya dengan Amazon CloudWatch untuk melaporkan progres dan kondisi prosedur pemulihan otomatis. Sertakan prosedur untuk memvalidasi pemulihan yang berhasil. Sediakan mekanisme untuk membatalkan pemulihan yang sedang berlangsung.

b. Buat playbook sebagai proses alternatif untuk kesalahan yang tidak dapat dipulihkan secara otomatis, dan pastikan keselarasannya dengan rencana pemulihan bencana Anda.

c. Uji proses pemulihan sebagaimana dibahas di REL13-BP03.

3. Bersiap untuk pemulihan

a. Evaluasi keadaan situs pemulihan Anda dan lakukan deployment komponen penting ke situs tersebut sebelumnya. Untuk detail lebih lanjut, lihat REL13-BP04.

b. Tentukan peran, tanggung jawab, dan proses pengambilan keputusan yang jelas untuk operasi pemulihan, yang melibatkan pemangku kepentingan dan tim yang relevan di keseluruhan organisasi.

c. Tentukan kondisi untuk memulai proses pemulihan Anda.

d. Buat rencana untuk mengembalikan proses pemulihan dan kembali ke situs primer Anda jika diperlukan atau setelah dianggap aman.

Sumber daya

Praktik-praktik terbaik terkait:

- REL07-BP01 Menggunakan otomatisasi ketika mendapatkan atau menskalakan sumber daya
- REL11-BP01 Memantau semua komponen beban kerja untuk mendeteksi kegagalan
- REL13-BP02 Menggunakan strategi pemulihan untuk memenuhi sasaran pemulihan
- REL13-BP03 Menguji implementasi pemulihan bencana untuk memvalidasi implementasi
- REL13-BP04 Mengelola penyimpangan konfigurasi di lokasi atau Wilayah Pemulihan Bencana (DR)

Dokumen terkait:

- Blog Arsitektur AWS: Seri Pemulihan Bencana
- Pemulihan Bencana Beban Kerja di AWS: Pemulihan di Cloud (Laporan Resmi AWS)
- Mengorkestrasi Otomatisasi Pemulihan Bencana menggunakan ARC Amazon Route 53 dan AWS Step Functions
- Membuat runbook Otomatisasi AWS Systems Manager menggunakan AWS CDK
- AWS Marketplace: Produk yang Dapat Digunakan untuk Pemulihan Bencana
- AWS Systems Manager Automation
- AWS Elastic Disaster Recovery
- Menggunakan Elastic Disaster Recovery untuk Failover dan Failback
- Sumber Daya AWS Elastic Disaster Recovery
- Partner APN: Partner yang Dapat Membantu Anda Melakukan Pemulihan Bencana

Video terkait:

- AWS re:Invent 2018: Pola Arsitektur untuk Aplikasi Aktif-Aktif Multi-Wilayah (ARC209-R2)
- AWS re:Invent 2022: AWS On Air ft. AWS Failback untuk AWS Elastic Disaster Recovery

Kesimpulan

Kami harap laporan resmi ini dapat mengasah cara berpikir Anda, menawarkan ide baru, atau menimbulkan serangkaian pertanyaan yang mengarahkan pada argumen logis untuk Anda, baik Anda belum familier dengan topik ketersediaan dan keandalan, maupun sudah berpengalaman dan sedang mencari wawasan untuk memaksimalkan ketersediaan beban kerja misi kritis Anda. Kami harap ini akan memberikan Anda pemahaman yang lebih mendalam tentang tingkat ketersediaan yang tepat berdasarkan kebutuhan bisnis, dan cara mendesain serta mencapai keandalan. Kami mendorong Anda untuk memanfaatkan rekomendasi desain, operasional, dan berorientasi pemulihan yang ditawarkan di sini serta pengetahuan dan pengalaman Arsitek Solusi kami AWS . Kami ingin mendengar dari Anda—terutama tentang kisah sukses Anda yang mencapai tingkat ketersediaan yang tinggi. AWS Hubungi tim akun Anda atau gunakan Hubungi KAMI di situs web kami.

Kontributor

Para kontributor untuk dokumen ini antara lain:

- Michael Fischer, Arsitek Solusi Utama, Amazon Web Services
- Seth Eliot, Advokat Pengembang Utama, Amazon Web Services
- Mahanth Jayadeva, Solutions Architect – Well-Architected, Amazon Web Services
- Amulya Sharma, Arsitek Solusi Utama, Amazon Web Services
- Jason DiDomenico, Senior Solutions Architect – Cloud Foundations, Amazon Web Services
- Marcin Bednarz, Arsitek Solusi Utama, Amazon Web Services
- Tyler Applebaum, Senior Solutions Architect, Amazon Web Services
- Rodney Lester, Arsitek Solusi Utama, Amazon Web Services
- Joe Chapman, Arsitek Solusi Senior, Amazon Web Services
- Adrian Hornsby, Rekayasawan Pengembangan Sistem Utama, Amazon Web Services
- Kevin Miller, Wakil Presiden — S3, Amazon Web Services
- Shannon Richards, Manajer Program Teknis Utama, Amazon Web Services
- Laurent Domb, Kepala Ahli Teknologi - Fed Fin, Amazon Web Services
- Kevin Schwarz, Senior Solutions Architect, Amazon Web Services
- Rob Martell, Arsitek Ketahanan Cloud Utama, Amazon Web Services
- Priyam Reddy, Senior Solutions Architect Manager DR, Amazon Web Services
- Jeff Ferris, Ahli Teknologi Utama, Amazon Web Services
- Matias Battaglia, Senior Solutions Architect, Amazon Web Services

Sumber bacaan lebih lanjut

Untuk mendapatkan informasi tambahan, buka:

- AWS Kerangka Well-Architected
- Pusat Arsitektur AWS

Revisi dokumen

Untuk mengetahui jika ada perubahan pada laporan resmi ini, Anda dapat berlangganan umpan RSS.

Perubahan	Deskripsi	Tanggal
Panduan praktik terbaik yang sudah diperbarui	Praktik terbaik diperbarui dengan panduan baru di area berikut: REL 1, REL 2, REL 4, REL 6, REL 7, REL 8, REL 10, REL 12, dan REL 13. Panduan telah diperluas dan diklarifi kasi di seluruh pilar. Panduan REL10-BP02 dan REL12-BP03 telah digabungkan ke dalam praktik terbaik lainnya. Sumber daya di seluruh pilar telah diperbarui.	6 November 2024
Panduan praktik terbaik yang sudah diperbarui	Pembaruan kecil untuk praktik-praktik terbaik di REL 2, 4, 5, 6, 7, dan 8.	27 Juni 2024
Panduan praktik terbaik yang sudah diperbarui	Praktik terbaik diperbaru i dengan panduan baru di bidang-bidang berikut ini: Merancang interaksi dalam sebuah sistem terdistri busi untuk mencegah kegagalan, Merancang interaksi dalam sebuah sistem terdistribusi untuk mengurangi atau menahan kegagalan, Memantau sumber daya beban kerja, Merancang beban kerja Anda	6 Desember 2023

	untuk beradaptasi dengan perubahan permintaan, Menerapkan perubahan, dan Menguji keandalan.	
Panduan praktik terbaik yang sudah diperbarui	Praktik terbaik diperbarui dengan panduan baru di bidang-bidang berikut ini: Memantau sumber daya beban kerja dan Merancang beban kerja Anda untuk menahan kegagalan komponen.	3 Oktober 2023
Panduan praktik terbaik yang sudah diperbarui	Praktik terbaik diperbarui dengan panduan baru di bidang-bidang berikut ini: Merancang arsitektur layanan beban kerja Anda, Merancang interaksi dalam sebuah sistem terdistribusi untuk mengurangi atau menahan kegagalan, dan Memantau sumber daya beban kerja.	13 Juli 2023
Pembaruan kecil	Bahasa non-inklusif dihilangkan.	13 April 2023
Pembaruan untuk Kerangka Kerja baru	Praktik terbaik diperbarui dengan panduan preskriptif dan praktik terbaik baru ditambahkan.	10 April 2023
Laporan resmi diperbarui	Praktik terbaik sudah diperbarui dengan panduan implementasi yang baru.	15 Desember 2022

Pembaruan kecil	Koreksi angka dan sejumlah perubahan kecil lainnya.	17 November 2022
Laporan resmi diperbarui	Praktik terbaik diperluas dan rencana pengembangan sudah ditambahkan.	20 Oktober 2022
Laporan resmi diperbarui	Menambahkan dua praktik terbaik baru ke Pilar Keandalan di beberapa bagian Menggunakan Isolasi Kesalahan untuk Melindung i Beban Kerja Anda dan Merancang Beban Kerja Anda untuk Menahan Kegagalan Komponen.	5 Mei 2022
Laporan resmi diperbarui	Perbarui panduan Pemulihan Bencana untuk mencakup Pengontrol Pemulihan Aplikasi Route 53. Tambahkan referensi ke DevOps Guru. Perbarui beberapa tautan Sumber daya, dan perubahan -perubahan editorial kecil lainnya.	26 Oktober 2021
Pembaruan kecil	Informasi tentang AWS Fault Injection Service (AWS FIS) ditambahkan.	15 Maret 2021
Pembaruan kecil	Pembaruan teks kecil.	4 Januari 2021

Laporan resmi diperbarui	Lampiran A diperbarui untuk memperbarui Tujuan Desain Ketersediaan untuk Amazon SQS, Amazon SNS, dan Amazon MQ; Susun ulang baris dalam tabel untuk pencarian yang lebih mudah; Tingkatkan penjelasan tentang perbedaan antara ketersedi aan dan pemulihan bencana dan bagaimana keduanya berkontribusi pada keandalan ; Perluas cakupan arsitektur multi-wilayah (untuk ketersedi aan) dan strategi multi-wilayah (untuk pemulihan bencana); Perbarui buku referensi ke versi terbaru; Perluas penghitungan ketersediaan untuk menyertakan penghitun gan berbasis permintaan, dan penghitungan pintasan; Tingkatkan deskripsi untuk Game Days	7 Desember 2020
Pembaruan kecil	Lampiran A diperbarui untuk memperbarui Tujuan Desain Ketersediaan untuk AWS Lambda	27 Oktober 2020
Pembaruan kecil	Lampiran A diperbarui untuk menambahkan Tujuan Desain Ketersediaan untuk AWS Global Accelerator	24 Juli 2020

Pembaruan untuk Kerangka Kerja baru	Pembaruan substansial dan konten baru/yang direvisi, termasuk: Penambaha n bagian praktik terbaik “Arsitektur Beban Kerja”, pengorganisasian ulang praktik terbaik ke dalam bagian Manajemen Perubahan dan Manajemen Kegagalan , pembaruan Sumber Daya, pembaruan untuk menyertak an sumber daya dan layanan AWS terbaru seperti AWS Global Accelerator, AWS Service Quotas, AWS Transit Gateway, penambahan/ pembaruan definisi untuk Keandalan, Ketersediaan, Ketahanan, laporan resmi yang lebih selaras dengan AWS Well-Architected Tool (pertanyaan dan praktik terbaik) yang digunakan untuk Peninjauan Well-Architected, menyusun ulang prinsip-p rinsip desain, memindahkan Melakukan pemulihan dari kegagalan secara otomatis sebelum Mencoba prosedur pemulihan, pembaruan diagram dan format untuk persamaan, penghapus an bagian Layanan Kunci and menggantinya dengan mengintegrasikan referensi -referensi terhadap layanan	8 Juli 2020

	AWS kunci ke dalam praktik terbaik.	
Pembaruan kecil	Perbaikan tautan yang bermasalah	1 Oktober 2019
Laporan resmi diperbarui	Lampiran A diperbarui	1 April 2019
Laporan resmi diperbarui	Penambahan rekomendasi jaringan AWS Direct Connect spesifik dan tujuan-tujuan desain layanan tambahan	1 September 2018
Laporan resmi diperbarui	Penambahan bagian Prinsip Desain dan Manajemen Batas. Pembaruan tautan, penghapusan ambiguita s terminologi upstream/ downstream, dan penambaha n referensi eksplisit ke topik Pilar Keandalan di dalam skenario ketersediaan.	1 Juni 2018
Laporan resmi diperbarui	Solusi Lintas Wilayah DynamoDB diganti menjadi Tabel Global DynamoDB. Tujuan desain layanan ditambahkan	1 Maret 2018
Pembaruan kecil	Koreksi kecil pada perhitungan ketersediaan agar menyertak an ketersediaan aplikasi	1 Desember 2017

Laporan resmi diperbarui	Pembaruan untuk memberikan panduan mengenai desain dengan ketersediaan tinggi, termasuk konsep, praktik terbaik, dan implementasi contoh.	1 November 2017
Publikasi awal	Pilar Keandalan - Kerangka Kerja AWS Well-Architected diterbitkan.	1 November 2016

Pemberitahuan

Pelanggan bertanggung jawab untuk membuat penilaian independen mereka sendiri atas informasi dalam dokumen ini. Dokumen ini: (a) hanya untuk tujuan informasi, (b) mewakili penawaran dan praktik AWS produk saat ini, yang dapat berubah tanpa pemberitahuan, dan (c) tidak membuat komitmen atau jaminan apa pun dari AWS dan afiliasinya, pemasok, atau pemberi lisensinya. AWS produk atau layanan disediakan “sebagaimana adanya” tanpa jaminan, representasi, atau kondisi apa pun, baik tersurat maupun tersirat. Tanggung jawab dan kewajiban AWS kepada pelanggannya dikendalikan oleh AWS perjanjian, dan dokumen ini bukan bagian dari, juga tidak mengubah, perjanjian apa pun antara AWS dan pelanggannya.

AWS Glosarium

Untuk AWS terminologi terbaru, lihat AWS glosarium di Referensi.Glosarium AWS