Laporan Resmi AWS

Pemulihan Bencana Beban Kerja di AWS: Pemulihan di Cloud

Pemulihan Bencana Beban Kerja di AWS: Pemulihan di Cloud: Laporan Resmi AWS

Table of Contents

Pemulihan Bencana Beban Kerja di AWS: Pemulihan di Cloud

Tanggal publikasi: 12 Februari 2021 (Revisi dokumen)

Abstrak

Pemulihan bencana adalah proses persiapan dan pemulihan dari bencana. Peristiwa yang mencegah beban kerja atau sistem memenuhi sasaran bisnis di lokasi utama deployment-nya akan dianggap sebagai bencana. Laporan ini menjelaskan praktik terbaik untuk merencanakan dan menguji pemulihan bencana untuk setiap beban kerja yang di-deploy ke AWS, serta menawarkan pendekatan berbeda untuk memitigasi risiko dan memenuhi Sasaran Waktu Pemulihan (RTO) dan Sasaran Titik Pemulihan (RPO) untuk beban kerja tersebut.

Pengantar

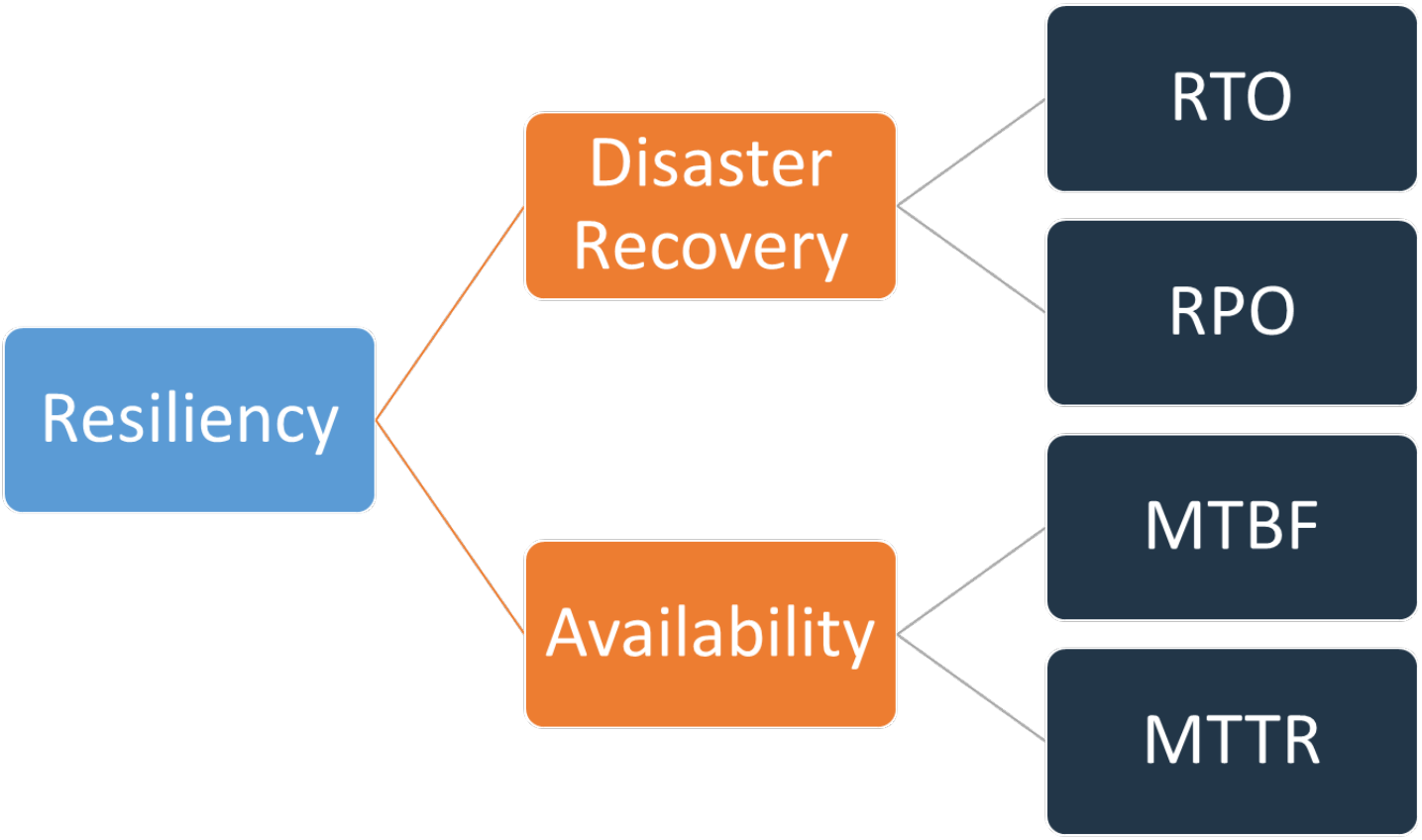
Beban kerja Anda harus melakukan fungsi yang dimaksudkan dengan benar dan konsisten. Untuk mencapai hal ini, Anda harus merancang ketahanan. Ketahanan adalah kemampuan beban kerja untuk pulih dari gangguan infrastruktur atau layanan, secara dinamis memperoleh sumber daya komputasi untuk memenuhi permintaan, dan mengurangi gangguan, seperti kesalahan konfigurasi atau masalah jaringan sementara.

Pemulihan bencana (DR) adalah bagian penting dari strategi ketahanan Anda dan berkaitan dengan bagaimana beban kerja Anda merespons ketika bencana terjadi (bencana adalah peristiwa yang menyebabkan dampak negatif serius pada bisnis Anda). Respons ini harus didasarkan pada sasaran bisnis organisasi Anda yang menentukan strategi beban kerja Anda untuk menghindari hilangnya data, yang dikenal sebagai Sasaran Titik Pemulihan (RPO), dan mengurangi waktu henti saat beban kerja Anda tidak tersedia untuk digunakan, yang dikenal sebagai Sasaran Waktu Pemulihan (RTO). Oleh karena itu, Anda harus menerapkan ketahanan dalam desain beban kerja Anda di cloud untuk memenuhi sasaran pemulihan Anda (RPO dan RTO) untuk peristiwa bencana satu kali tertentu. Pendekatan ini membantu organisasi Anda untuk mempertahankan kelangsungan bisnis sebagai bagian dari Rencana Kelangsungan Bisnis (BCP).

Laporan ini berfokus pada cara merencanakan, merancang, dan menerapkan arsitektur di AWS yang memenuhi sasaran pemulihan bencana untuk bisnis Anda. Informasi yang dibagikan di sini ditujukan untuk mereka yang memiliki peran teknologi, seperti CTO, arsitek, developer, dan anggota tim operasional.

Pemulihan bencana dan ketersediaan

Pemulihan bencana dapat dibandingkan dengan ketersediaan, yang merupakan komponen penting lain dari strategi ketahanan Anda. Berbeda dari pemulihan bencana yang mengukur sasaran untuk peristiwa satu kali, sasaran ketersediaan mengukur nilai rata-rata selama periode waktu tertentu.



Gambar 1 - Sasaran Ketahanan

Ketersediaan dihitung menggunakan Mean Time Between Failures (MTBF) dan Mean Time to Recover (MTTR):

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

Pendekatan ini sering disebut sebagai “sembilan”, dengan target ketersediaan 99,9% yang disebut sebagai “tiga sembilan”.

Untuk beban kerja Anda, mungkin lebih mudah menghitung permintaan yang berhasil dan gagal alih-alih menggunakan pendekatan berbasis waktu. Dalam hal ini, penghitungan berikut dapat digunakan:

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

Pemulihan bencana berfokus pada peristiwa bencana, sedangkan ketersediaan berfokus pada gangguan yang lebih umum dan berskala lebih kecil seperti kegagalan komponen, masalah jaringan, dan lonjakan beban. Sasaran pemulihan bencana adalah kelangsungan bisnis, sedangkan ketersediaan berkaitan dengan memaksimalkan waktu ketersediaan beban kerja untuk menjalankan fungsionalitas bisnis yang dimaksudkan. Keduanya harus menjadi bagian dari strategi ketahanan Anda.

Model Tanggung Jawab Bersama untuk Ketahanan

Ketahanan adalah tanggung jawab bersama antara AWS dan Anda sebagai pelanggan. Penting bagi Anda untuk memahami bagaimana pemulihan dan ketersediaan bencana, sebagai bagian dari ketahanan, beroperasi di bawah model bersama ini.

Tanggung jawab AWS “Ketahanan dari Cloud”

AWS bertanggung jawab atas ketahanan infrastruktur yang menjalankan semua layanan yang ditawarkan di dalam AWS Cloud. Infrastruktur ini terdiri dari perangkat keras, perangkat lunak, jaringan, dan fasilitas yang menjalankan layanan AWS Cloud. AWS menggunakan upaya yang wajar secara komersial untuk menyediakan layanan AWS Cloud ini, dengan memastikan ketersediaan layanan memenuhi atau melebihi Perjanjian Tingkat Layanan (SLA) AWS.

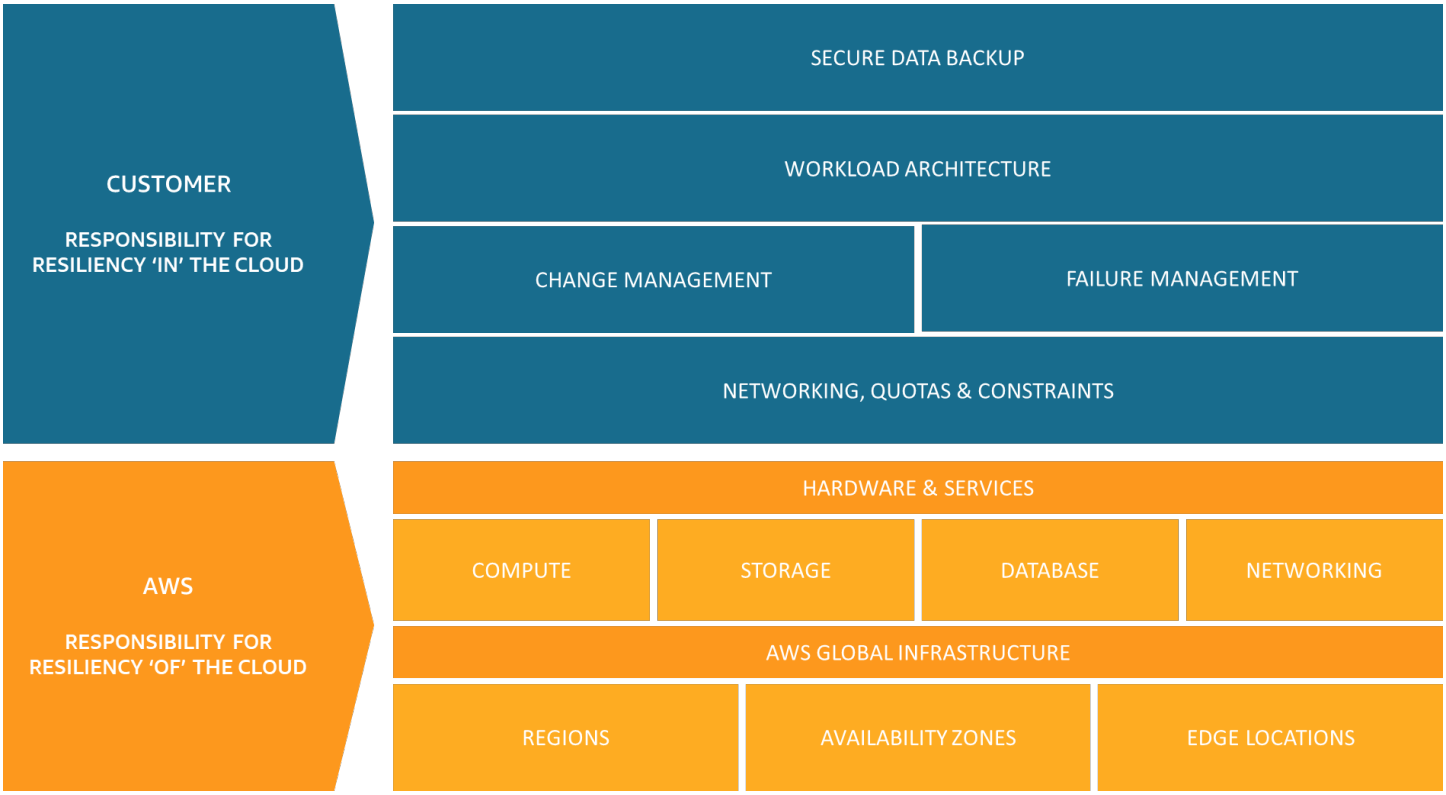
AWS Global Cloud Infrastructure dirancang untuk memungkinkan pelanggan membangun arsitektur beban kerja yang sangat tangguh. Setiap Wilayah AWS sepenuhnya terisolasi dan terdiri dari sejumlah Zona Ketersediaan, yang merupakan partisi infrastruktur yang terisolasi secara fisik. Zona Ketersediaan mengisolasi kesalahan yang dapat memengaruhi ketahanan beban kerja, sehingga tidak akan berdampak pada zona lain di Wilayahnya. Namun, semua zona di Wilayah AWS juga saling terhubung dengan jaringan bandwidth tinggi dan latensi rendah, melalui serat metro khusus yang redundan sepenuhnya, yang menyediakan jaringan dengan throughput tinggi dan latensi rendah di antara zona. Semua lalu lintas di antara zona dienkripsi. Performa jaringan cukup untuk melakukan replikasi sinkron di antara zona. Zona Ketersediaan menyederhanakan proses partisi aplikasi untuk ketersediaan tinggi.

Tanggung jawab pelanggan “Ketahanan di Cloud”

Tanggung jawab Anda akan ditentukan oleh layanan AWS Cloud yang Anda pilih. Hal ini menentukan jumlah pekerjaan konfigurasi yang harus Anda lakukan sebagai bagian dari tanggung jawab ketahanan Anda. Misalnya, layanan seperti Amazon Elastic Compute Cloud (Amazon EC2) mengharuskan pelanggan untuk melakukan semua tugas konfigurasi dan manajemen ketahanan yang diperlukan. Pelanggan yang men-deploy instans Amazon EC2 bertanggung jawab untuk men-deploy instans EC2 di sejumlah lokasi (seperti Zona Ketersediaan AWS), men-deploy pemulihan mandiri menggunakan layanan seperti AWS Auto Scaling, serta menggunakan praktik terbaik arsitektur beban kerja yang tangguh untuk aplikasi yang diinstal pada instans. Untuk layanan terkelola, seperti Amazon S3 dan Amazon DynamoDB, AWS mengoperasikan lapisan infrastruktur,

sistem operasi, dan platform, serta pelanggan mengakses titik akhir untuk menyimpan dan mengambil data. Anda bertanggung jawab untuk mengelola ketahanan data Anda termasuk strategi pencadangan, versioning, dan replikasi.

Men-deploy beban kerja Anda di sejumlah Zona Ketersediaan di Wilayah AWS merupakan bagian dari strategi ketersediaan tinggi yang dirancang untuk melindungi beban kerja dengan mengisolasi masalah ke satu Zona Ketersediaan, dan menggunakan redundansi Zona Ketersediaan lainnya untuk terus melayani permintaan. Arsitektur Multi-AZ juga merupakan bagian dari strategi DR yang dirancang untuk membuat beban kerja lebih terisolasi dan terlindungi dari masalah seperti pemadaman listrik, sambaran petir, tornado, gempa bumi, dan banyak lagi. Strategi DR juga dapat menggunakan sejumlah Wilayah AWS. Misalnya dalam konfigurasi aktif/pasif, layanan untuk beban kerja akan melakukan failover dari wilayah aktifnya ke wilayah DR jika Wilayah aktif tidak dapat lagi melayani permintaan.



Gambar 2 - Ketahanan adalah tanggung jawab bersama antara AWS dan pelanggan

Apa itu bencana?

Saat merencanakan pemulihan bencana, evaluasi rencana Anda untuk tiga kategori utama bencana ini:

- Bencana alam, seperti gempa bumi atau banjir
- Kegagalan teknis, seperti gangguan daya atau konektivitas jaringan
- Tindakan manusia, seperti kesalahan konfigurasi yang tidak disengaja atau akses atau modifikasi tidak sah/pihak luar

Masing-masing potensi bencana ini juga akan memiliki dampak geografis yang dapat bersifat lokal, regional, di seluruh negara, benua, atau global. Baik sifat bencana maupun dampak geografisnya sangat penting ketika mempertimbangkan strategi pemulihan bencana Anda. Misalnya, Anda dapat mengurangi dampak masalah banjir lokal yang menyebabkan gangguan pusat data dengan menggunakan strategi Multi-AZ, karena hal itu tidak akan memengaruhi lebih dari satu Zona Ketersediaan. Namun, serangan terhadap data produksi akan mengharuskan Anda untuk menjalankan strategi pemulihan bencana yang melakukan failover untuk membuat cadangan data di Wilayah AWS lain.

Ketersediaan tinggi bukanlah pemulihan bencana

Ketersediaan dan pemulihan bencana bergantung pada sejumlah praktik terbaik yang sama, seperti pemantauan kegagalan, deployment ke sejumlah lokasi, dan failover otomatis. Namun, Ketersediaan berfokus pada komponen beban kerja, sedangkan pemulihan bencana berfokus pada salinan diskret dari seluruh beban kerja. Pemulihan bencana memiliki sasaran yang berbeda dari Ketersediaan, yakni mengukur waktu pemulihan setelah peristiwa berskala lebih besar yang memenuhi syarat sebagai bencana. Anda harus terlebih dahulu memastikan beban kerja Anda memenuhi sasaran ketersediaan Anda, karena arsitektur yang berketersediaan tinggi akan memungkinkan Anda memenuhi kebutuhan pelanggan jika terjadi peristiwa yang berdampak pada ketersediaan. Strategi pemulihan bencana Anda memerlukan pendekatan yang berbeda dari ketersediaan, dengan fokus pada deployment sistem diskret ke sejumlah lokasi, sehingga Anda dapat melakukan failover seluruh beban kerja jika diperlukan.

Anda harus mempertimbangkan ketersediaan beban kerja Anda dalam perencanaan pemulihan bencana, karena akan memengaruhi pendekatan yang Anda ambil. Beban kerja yang berjalan pada instans Amazon EC2 tunggal dalam satu Zona Ketersediaan tidak memiliki ketersediaan tinggi. Jika masalah banjir lokal memengaruhi Zona Ketersediaan tersebut, skenario ini memerlukan failover ke AZ lain untuk memenuhi sasaran DR. Bandingkan skenario ini dengan beban kerja yang berketersediaan tinggi yang di-deploy dengan strategi multi-lokasi aktif/aktif saat beban kerja di-deploy di sejumlah Wilayah aktif, dan semua Wilayah ini melayani lalu lintas produksi. Dalam hal ini, bahkan dalam peristiwa yang tidak mungkin terjadi seperti bencana besar yang mengganggu seluruh Wilayah, strategi DR dapat dicapai dengan merutekan semua lalu lintas ke Wilayah yang tersisa.

Cara Anda melakukan pendekatan data juga berbeda antara ketersediaan dan pemulihan bencana. Pertimbangkan solusi penyimpanan yang terus mereplikasi ke situs lain untuk mencapai ketersediaan tinggi (seperti beban kerja multi-situs aktif/aktif). Jika satu file atau sejumlah file dihapus atau rusak pada perangkat penyimpanan utama, perubahan destruktif tersebut dapat direplikasi ke perangkat penyimpanan sekunder. Dalam skenario ini, meskipun ada ketersediaan tinggi, kemampuan failover jika terjadi penghapusan atau kerusakan data akan dirugikan. Sebagai gantinya, pencadangan point-in-time juga diperlukan sebagai bagian dari strategi DR.

Rencana Kelangsungan Bisnis (BCP)

Rencana pemulihan bencana Anda harus menjadi bagian dari rencana kelangsungan bisnis (BCP) organisasi Anda, bukan sebagai dokumen mandiri. Tidak ada gunanya mempertahankan target pemulihan bencana yang agresif untuk memulihkan beban kerja jika sasaran bisnis beban kerja tersebut tidak dapat dicapai karena dampak bencana pada elemen bisnis Anda di samping beban kerja Anda. Misalnya, gempa bumi dapat mencegah Anda mengangkut produk yang dibeli di aplikasi perdagangan elektronik Anda – bahkan jika DR yang efektif menjaga beban kerja Anda tetap berfungsi, BCP Anda perlu mengakomodasi kebutuhan pengangkutan. Strategi DR Anda harus didasarkan pada persyaratan bisnis, prioritas, dan konteks.

Analisis dampak bisnis dan penilaian risiko

Analisis dampak bisnis harus mengukur dampak bisnis dari gangguan beban kerja Anda. Ini harus mengidentifikasi dampak pada pelanggan internal dan eksternal yang tidak dapat menggunakan beban kerja Anda, dan efek yang ada pada bisnis Anda. Analisis ini akan membantu menentukan seberapa cepat beban kerja perlu tersedia dan berapa banyak kehilangan data yang dapat ditoleransi. Namun, penting untuk dicatat bahwa sasaran pemulihan tidak boleh dibuat secara terpisah; kemungkinan gangguan dan biaya pemulihan adalah faktor kunci yang membantu menentukan nilai bisnis dalam menyediakan pemulihan bencana untuk sebuah beban kerja.

Dampak bisnis mungkin akan tergantung waktu. Anda sebaiknya mempertimbangkan hal ini untuk perencanaan pemulihan bencana Anda. Misalnya, gangguan pada sistem penggajian Anda cenderung memiliki dampak yang sangat tinggi terhadap bisnis jika terjadi sebelum jadwal penerimaan gaji karyawan, tetapi mungkin memiliki dampak rendah setelah semua karyawan menerima gaji mereka.

Penilaian risiko terhadap jenis bencana dan dampak geografis bersama dengan gambaran umum pelaksanaan teknis beban kerja Anda akan menentukan kemungkinan gangguan yang terjadi untuk setiap jenis bencana.

Untuk beban kerja yang sangat penting, Anda dapat mempertimbangkan ketersediaan tinggi yang ada di sejumlah Wilayah dengan pencadangan berkelanjutan untuk meminimalkan dampak bisnis. Untuk beban kerja yang kurang penting, tidak menjalankan pemulihan bencana sama sekali mungkin merupakan strategi yang valid. Untuk beberapa skenario bencana, juga valid untuk tidak memiliki strategi pemulihan bencana yang berlaku sebagai keputusan yang matang berdasarkan probabilitas kemunculan bencana yang rendah. Ingatlah bahwa Zona Ketersediaan dalam Wilayah AWS sudah

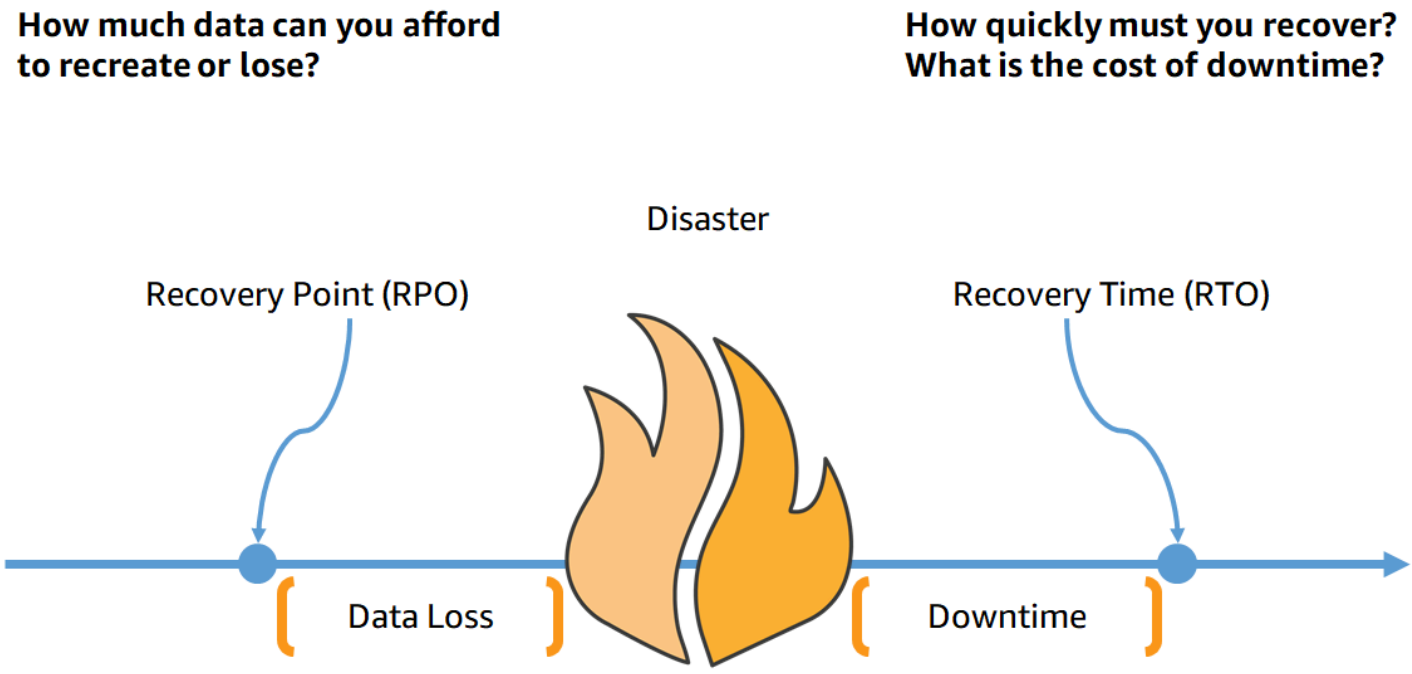
dirancang dengan jarak yang tepat antara satu sama lain dan perencanaan lokasi yang cermat, sehingga bencana yang paling umum hanya memengaruhi satu zona dan bukan yang lain. Oleh karena itu, arsitektur Multi-AZ dalam Wilayah AWS mungkin sudah memenuhi kebutuhan mitigasi risiko Anda.

Biaya opsi pemulihan bencana harus dievaluasi untuk memastikan bahwa strategi pemulihan bencana memberikan tingkat nilai bisnis yang benar dengan mempertimbangkan dampak dan risikonya terhadap bisnis.

Dengan semua informasi ini, Anda dapat mendokumentasikan ancaman, risiko, dampak, dan biaya dari skenario bencana yang berbeda-beda serta opsi pemulihan terkait. Informasi ini harus digunakan untuk menentukan sasaran pemulihan Anda untuk setiap beban kerja Anda.

Sasaran pemulihan (RTO dan RPO)

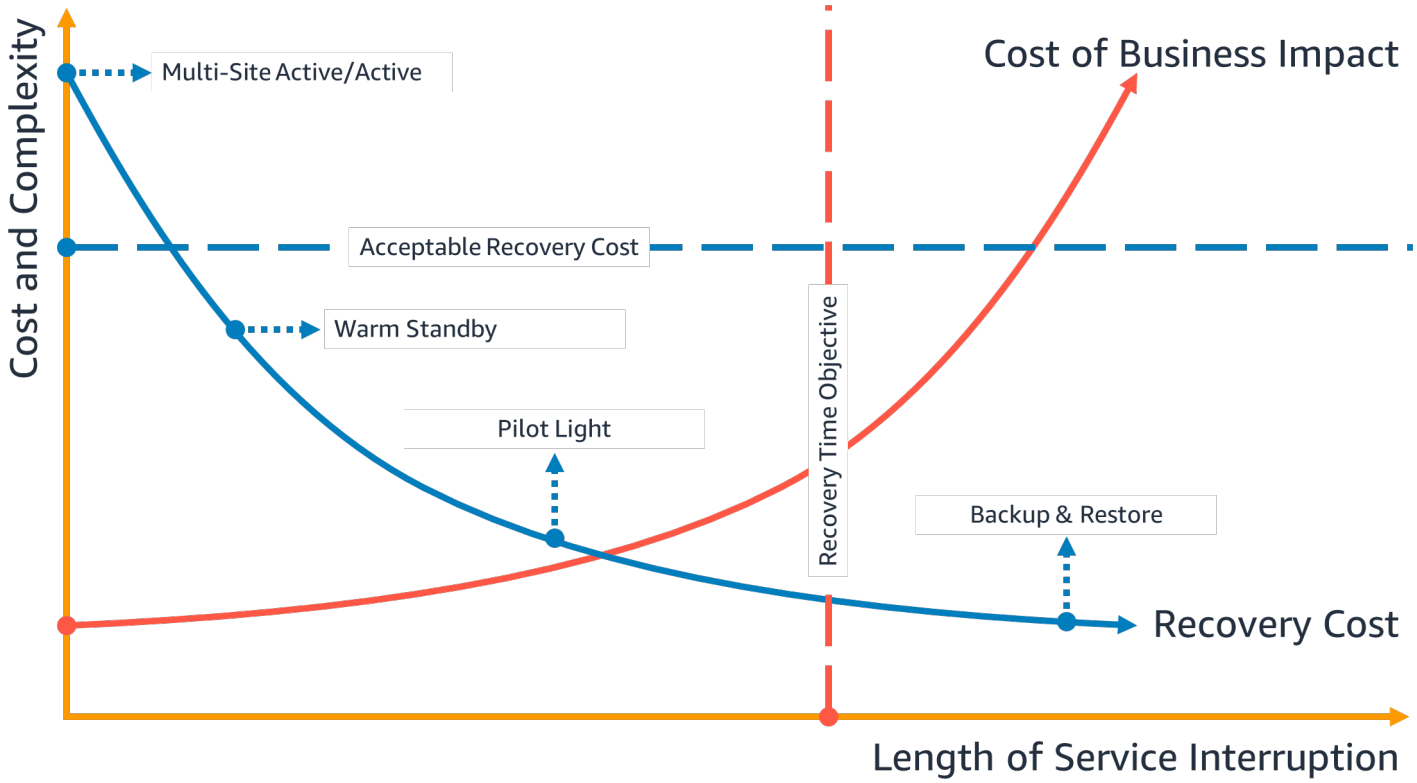
Saat membuat strategi Pemulihan Bencana (DR), organisasi paling sering merencanakan untuk Sasaran Waktu Pemulihan (RTO) dan Sasaran Titik Pemulihan (RPO).



Gambar 3 - Sasaran pemulihan

Sasaran Waktu Pemulihan (RTO) adalah penundaan maksimum yang dapat diterima antara gangguan layanan dan pemulihan layanan. Sasaran ini menentukan periode yang dianggap dapat diterima ketika layanan tidak tersedia, dan ini ditentukan oleh organisasi.

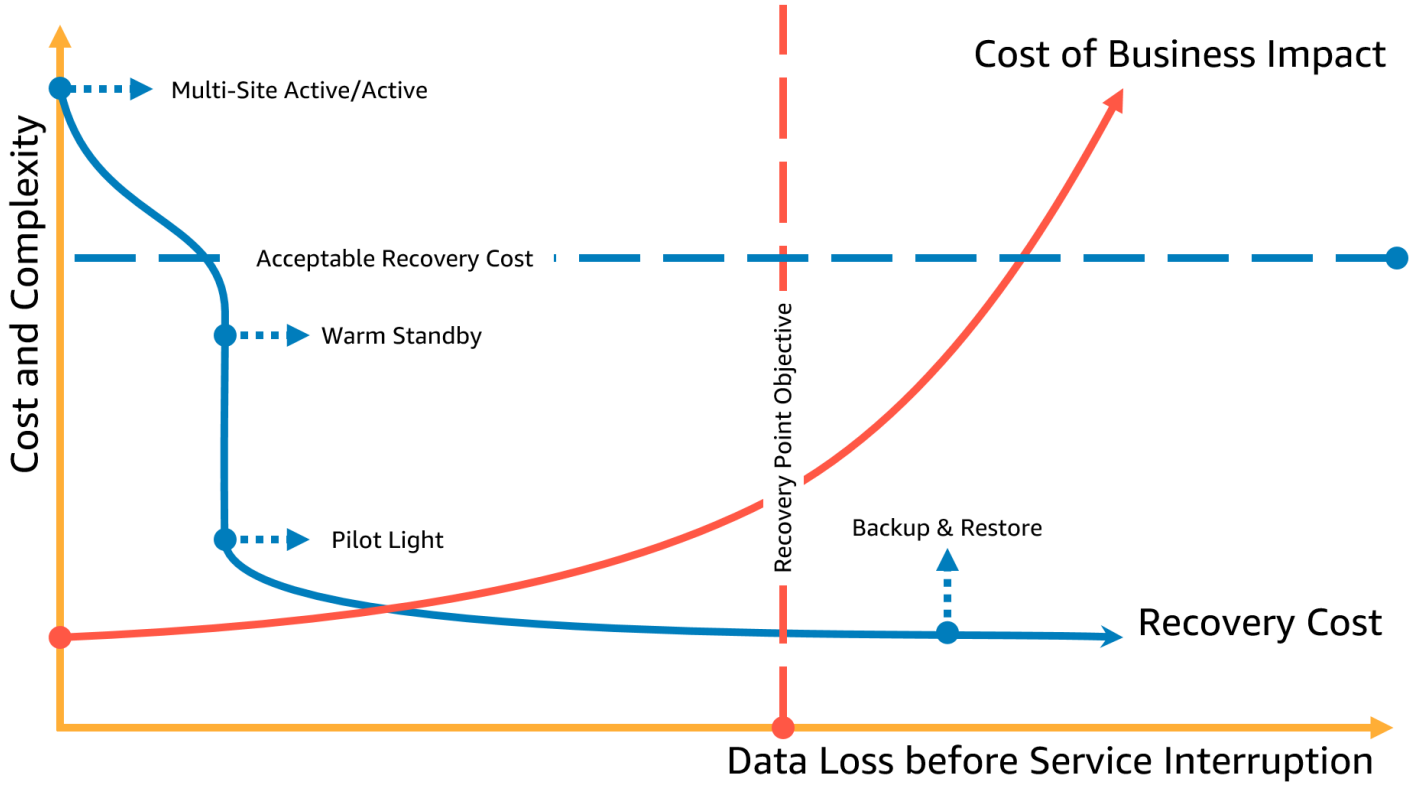
Ada empat strategi DR yang dibahas dalam laporan ini: pencadangan dan pemulihan, pilot light, warm standby, dan multi-lokasi aktif/aktif (lihat Opsi Pemulihan Bencana di Cloud). Dalam diagram berikut, bisnis telah menentukan RTO maksimum yang diizinkan serta batas jumlah pembelanjaan yang dapat mereka habiskan untuk strategi pemulihan layanan mereka. Mengingat sasaran bisnis ini, strategi DR Pilot Light atau Warm Standby akan memenuhi RTO dan kriteria biaya tersebut.



Gambar 4 - Sasaran waktu pemulihan

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

Dalam diagram berikut, bisnis telah menentukan RPO maksimum yang diizinkan serta batas jumlah pembelanjaan yang dapat mereka habiskan untuk strategi pemulihan data mereka. Dari empat strategi DR tersebut, baik strategi DR Pilot Light atau Warm Standby memenuhi kriteria RPO dan biaya.



Gambar 5 - Sasaran titik pemulihan

> **Note**
>
> Jika biaya pemulihan lebih tinggi daripada biaya kegagalan atau kerugian, opsi pemulihan tidak boleh diterapkan kecuali jika ada alasan tambahan, misalnya persyaratan peraturan.

Pemulihan bencana akan berbeda di cloud

Strategi pemulihan bencana berkembang seiring dengan inovasi teknis. Rencana pemulihan bencana di on-premise mungkin mengharuskan pengangkutan media pita secara fisik atau replikasi data ke lokasi lain. Organisasi Anda perlu mengevaluasi kembali dampak bisnis, risiko, dan biaya strategi pemulihan bencana sebelumnya untuk memenuhi sasaran DR di AWS. Pemulihan bencana di AWS Cloud mencakup keuntungan berikut dibandingkan lingkungan tradisional:

- Memulihkan dengan cepat dari bencana dengan kompleksitas yang rendah
- Pengujian yang sederhana dan berulang memungkinkan Anda untuk menguji secara lebih mudah dan lebih sering
- Overhead manajemen yang lebih rendah mengurangi beban operasional
- Peluang otomatisasi akan mengurangi kemungkinan kesalahan dan mempercepat waktu pemulihan

AWS memungkinkan Anda untuk meninggalkan biaya modal tetap dari pusat data cadangan fisik dan beralih ke biaya operasi variabel dari lingkungan yang dapat disesuaikan ukurannya di cloud, sehingga mengurangi biaya secara signifikan.

Untuk banyak organisasi, pemulihan bencana on-premise didasarkan pada risiko gangguan pada sebuah beban kerja atau sejumlah beban kerja di pusat data dan pemulihan data yang dicadangkan atau direplikasi ke pusat data sekunder. Ketika organisasi men-deploy beban kerja di AWS, mereka dapat men-deploy beban kerja yang dirancang dengan baik dan mengandalkan desain AWS Global Cloud Infrastructure untuk membantu mengurangi dampak gangguan tersebut. Lihat laporan resmi AWS Well-Architected Framework - Pilar Keandalan untuk informasi lebih lanjut tentang praktik terbaik arsitektur untuk merancang dan mengoperasikan beban kerja yang andal, aman, efisien, dan hemat biaya di cloud.

Jika beban kerja Anda ada di AWS, Anda tidak perlu khawatir dengan konektivitas pusat data (kecuali kemampuan Anda untuk mengaksesnya), daya listrik, AC, pemadaman kebakaran, dan perangkat keras. Semua hal ini dikelola untuk Anda dan Anda memiliki akses ke sejumlah zona ketersediaan yang terisolasi dari kesalahan (masing-masing terdiri dari satu atau beberapa pusat data diskret).

Wilayah AWS Tunggal

Untuk peristiwa bencana berdasarkan gangguan atau kehilangan satu pusat data fisik, menerapkan beban kerja yang berketersediaan tinggi di sejumlah zona ketersediaan dalam satu Wilayah AWS akan membantu mengurangi efek bencana alam dan teknis serta mengurangi risiko terhadap ancaman manusia seperti kesalahan atau aktivitas tidak sah yang dapat mengakibatkan kehilangan data. Setiap Wilayah AWS terdiri dari sejumlah Zona Ketersediaan, yang masing-masing terisolasi dari gangguan di zona lain. Selain itu, setiap Zona Ketersediaan terdiri dari sejumlah pusat data fisik. Untuk lebih mengisolasi masalah yang berdampak besar dan mencapai ketersediaan tinggi, Anda dapat mempartisi beban kerja di sejumlah zona di Wilayah yang sama. Zona Ketersediaan dirancang untuk redundansi fisik dan menyediakan ketahanan, yang memungkinkan performa tanpa gangguan, bahkan pada saat listrik padam, Internet mati, banjir, dan bencana alam lainnya. Lihat AWS Global Cloud Infrastructure untuk mengetahui cara AWS melakukannya.

Dengan men-deploy di sejumlah Zona Ketersediaan dalam satu Wilayah AWS, beban kerja Anda lebih terlindungi dari kegagalan satu (atau bahkan banyak) pusat data. Untuk jaminan tambahan dengan deployment Wilayah Tunggal, Anda dapat mencadangkan data dan konfigurasi (termasuk definisi infrastruktur) ke Wilayah lain. Strategi ini mengurangi cakupan rencana pemulihan bencana Anda agar hanya menyertakan pencadangan dan pemulihan data. Memanfaatkan ketahanan multi-wilayah dengan mencadangkan ke Wilayah AWS lain relatif sederhana dan murah dibandingkan dengan opsi Multi-Wilayah lainnya yang dijelaskan di bagian berikut. Misalnya, mencadangkan ke Amazon Simple Storage Service (Amazon S3) memberi Anda akses ke pengambilan data secara langsung. Namun, jika strategi DR untuk sebagian data Anda memiliki persyaratan yang tidak terlalu ketat dalam hal waktu pengambilan data (dari hitungan menit ke jam), maka menggunakan Amazon S3 Glacier atau Amazon S3 Glacier Deep Archive akan secara signifikan mengurangi biaya strategi pencadangan dan pemulihan Anda.

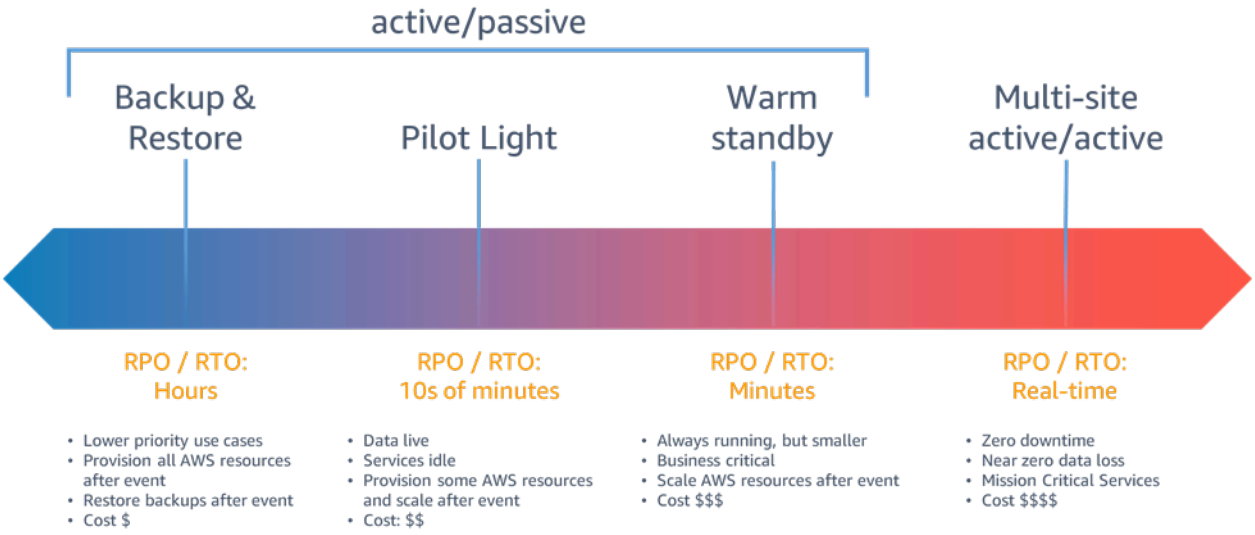
Beberapa beban kerja mungkin memiliki persyaratan peraturan residensi data. Jika ini berlaku untuk beban kerja Anda di daerah yang saat ini hanya memiliki satu Wilayah AWS, maka selain merancang beban kerja Multi-AZ untuk ketersediaan tinggi seperti yang dibahas di atas, Anda juga dapat menggunakan AZ dalam Wilayah tersebut sebagai lokasi diskret, yang dapat membantu mengatasi persyaratan residensi data yang berlaku untuk beban kerja Anda dalam Wilayah tersebut. Strategi DR yang dijelaskan di bagian berikut menggunakan beberapa Wilayah AWS, tetapi juga dapat diimplementasikan menggunakan Zona Ketersediaan, bukan Wilayah.

Beberapa Wilayah AWS

Untuk peristiwa bencana yang dapat menimbulkan risiko kehilangan sejumlah pusat data jarak yang jauh dari satu sama lain, Anda harus mempertimbangkan opsi pemulihan bencana untuk mengurangi dampak bencana alam dan teknis yang memengaruhi seluruh Wilayah dalam AWS. Semua opsi yang dijelaskan dalam bagian berikut dapat diimplementasikan sebagai arsitektur Multi-wilayah untuk melindungi terhadap bencana tersebut.

Opsi pemulihan bencana di cloud

Strategi pemulihan bencana yang tersedia bagi Anda dalam AWS dapat dikategorikan secara luas menjadi empat pendekatan, mulai dari biaya rendah dan kompleksitas rendah dalam membuat cadangan hingga strategi yang lebih kompleks menggunakan sejumlah Wilayah aktif. Penting untuk secara rutin menguji strategi pemulihan bencana Anda sehingga Anda memiliki keyakinan dalam menjalankannya, saat strategi ini diperlukan.



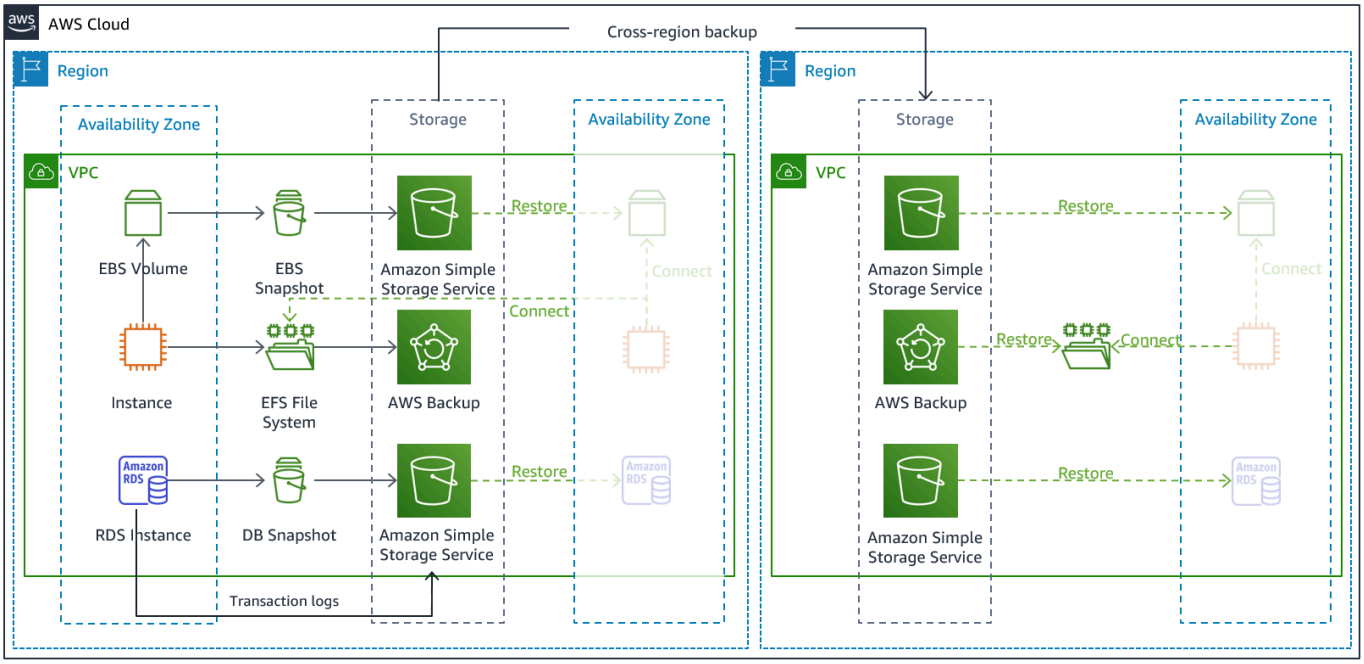
Gambar 6 - Strategi pemulihan bencana

Untuk peristiwa bencana berdasarkan gangguan atau kehilangan satu pusat data fisik untuk beban kerja yang dirancang dengan baik dan berketersediaan tinggi, Anda mungkin hanya memerlukan pendekatan pencadangan dan pemulihan untuk pemulihan bencana. Jika definisi bencana Anda tidak hanya berupa gangguan atau kehilangan pusat data fisik di suatu Wilayah atau jika Anda tunduk pada persyaratan peraturan yang mewajibkannya, maka Anda harus mempertimbangkan Pilot Light, Warm Standby, atau Multi-Lokasi Aktif/Aktif.

Pencadangan dan pemulihan

Pencadangan dan pemulihan adalah pendekatan yang cocok untuk mengurangi kehilangan atau kerusakan data. Pendekatan ini juga dapat digunakan untuk mengurangi bencana regional dengan mereplikasi data ke Wilayah AWS lainnya, atau untuk memitigasi kurangnya redundansi untuk beban kerja yang di-deploy ke Zona Ketersediaan tunggal. Selain data, Anda harus men-deploy ulang infrastruktur, konfigurasi, dan kode aplikasi di Wilayah pemulihan. Untuk memungkinkan infrastruktur di-deploy kembali dengan cepat tanpa kesalahan, Anda harus selalu men-deploy

infrastruktur sebagai kode (IAC) menggunakan layanan seperti AWS CloudFormation atau AWS Cloud Development Kit (AWS CDK). Tanpa IAC, mungkin akan rumit untuk memulihkan beban kerja di Wilayah pemulihan, yang akan menyebabkan penambahan waktu pemulihan dan mungkin melebihi RTO Anda. Selain data pengguna, pastikan juga untuk mencadangkan kode dan konfigurasi, termasuk Amazon Machine Image (AMI) yang Anda gunakan untuk membuat instans Amazon EC2. Anda dapat menggunakan AWS CodePipeline untuk mengotomatisasi deployment ulang kode aplikasi dan konfigurasi.



Gambar 7 - Arsitektur pencadangan dan pemulihan

Layanan AWS

Data beban kerja Anda akan memerlukan strategi pencadangan yang berjalan secara berkala atau terus-menerus. Seberapa sering Anda menjalankan cadangan Anda akan menentukan titik pemulihan yang dapat dicapai (yang harus selaras untuk memenuhi RPO Anda). Cadangannya juga harus menawarkan cara untuk dipulihkan ke titik waktu saat cadangan tersebut dibuat. Cadangan dengan pemulihan titik waktu (PITR) tersedia melalui layanan dan sumber daya berikut:

- Snapshot Amazon Elastic Block Store (Amazon EBS)
- Cadangan Amazon DynamoDB
- Snapshot Amazon RDS
- Snapshot Amazon Aurora DB

- Cadangan Amazon EFS (saat menggunakan AWS Backup)
- Snapshot Amazon Redshift
- Snapshot Amazon Neptunus

Untuk Amazon Simple Storage Service (Amazon S3), Anda dapat menggunakan Amazon S3 Cross-Region Replication (CRR) untuk secara asinkron menyalin objek ke bucket S3 di wilayah DR secara terus-menerus, sambil memberikan versioning untuk objek yang disimpan sehingga Anda dapat memilih titik pemulihan Anda. Replikasi data yang berkelanjutan memiliki keuntungan waktu yang paling singkat (mendekati nol) untuk mencadangkan data Anda, tetapi mungkin tidak melindungi dari peristiwa bencana seperti kerusakan data atau serangan jahat (seperti penghapusan data yang tidak sah) serta pencadangan titik waktu (point-in-time). Replikasi berkelanjutan dibahas dalam bagian Layanan AWS untuk Pilot Light.

AWS Backup menyediakan lokasi terpusat untuk mengonfigurasi, menjadwalkan, dan memantau kemampuan pencadangan AWS untuk layanan dan sumber daya berikut:

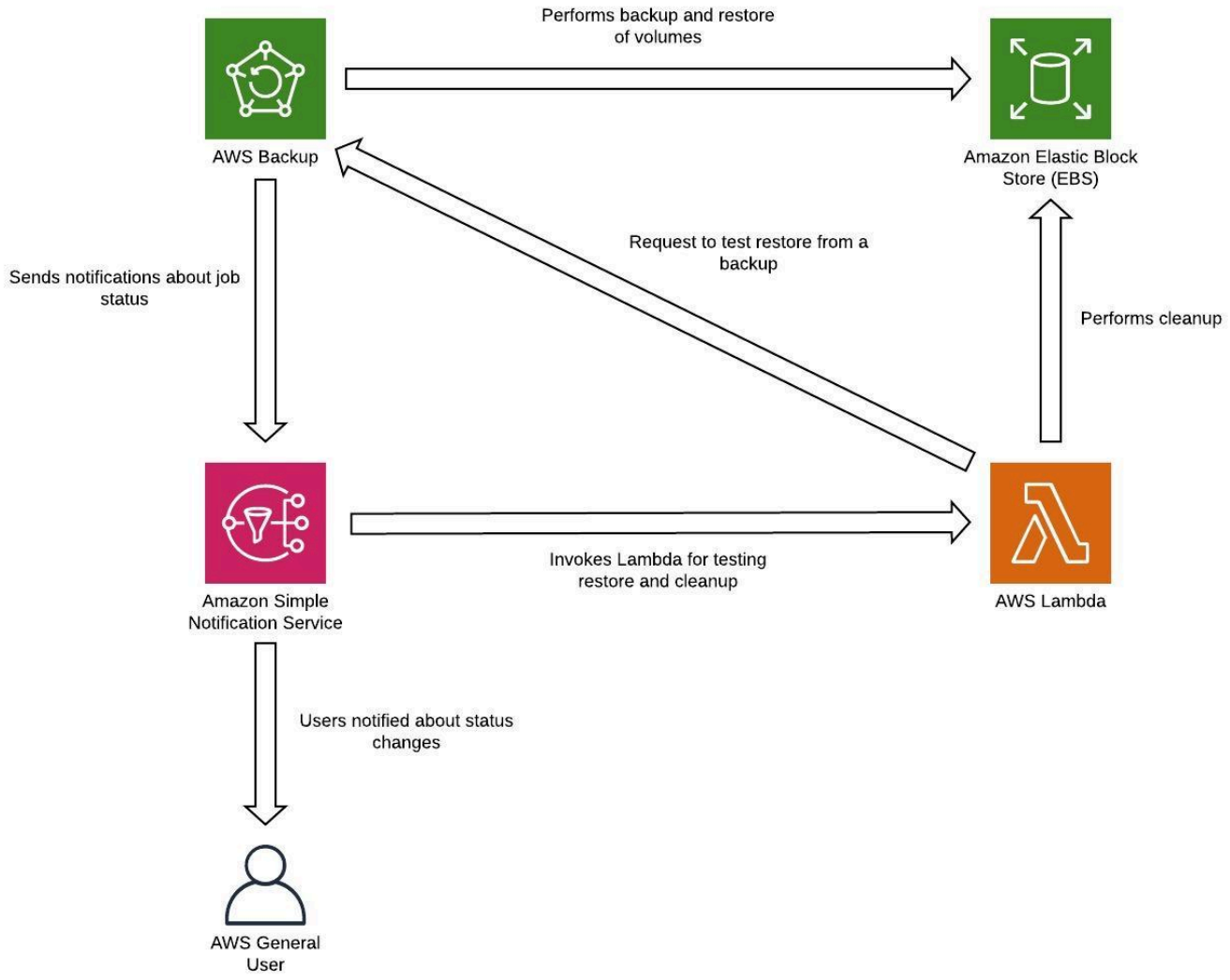
- Volume Amazon Elastic Block Store (Amazon EBS)
- Instans Amazon EC2
- Basis data Amazon Relational Database Service (Amazon RDS) (termasuk basis data Amazon Aurora)
- Tabel Amazon DynamoDB
- Sistem file Amazon Elastic File System (Amazon EFS)
- Volume AWS Storage Gateway
- Amazon FSx for Windows File Server dan Amazon FSx for Lustre

AWS Backup mendukung penyalinan cadangan di seluruh Wilayah, misalnya ke Wilayah pemulihan bencana.

Sebagai strategi pemulihan bencana tambahan untuk data Amazon S3 Anda, aktifkan versioning objek S3. Versioning objek melindungi data Anda di S3 dari konsekuensi tindakan penghapusan atau modifikasi dengan mempertahankan versi asli sebelum tindakan tersebut. Versioning objek dapat menjadi mitigasi yang berguna untuk bencana jenis kesalahan manusia. Jika Anda menggunakan replikasi S3 untuk mencadangkan data ke wilayah DR Anda, maka, secara default, ketika objek dihapus dalam bucket sumber, Amazon S3 menambahkan delete marker di bucket sumber saja. Pendekatan ini melindungi data di Wilayah DR dari penghapusan berniat jahat di wilayah sumber.

Selain data, Anda juga harus mencadangkan konfigurasi dan infrastruktur yang diperlukan untuk men-deploy kembali beban kerja Anda dan memenuhi Sasaran Waktu Pemulihan (RTO). AWS CloudFormation menyediakan Infrastruktur sebagai kode (IaC), dan memungkinkan Anda menentukan semua sumber daya AWS dalam beban kerja sehingga Anda dapat men-deploy dan men-deploy kembali secara andal ke sejumlah akun AWS dan Wilayah AWS. Anda dapat mencadangkan instans Amazon EC2 yang digunakan oleh beban kerja Anda sebagai Amazon Machine Image (AMI). AMI dibuat dari snapshot volume root instans Anda dan volume EBS lainnya yang dilampirkan pada instans Anda. Anda dapat menggunakan AMI ini untuk meluncurkan versi instans EC2 yang dipulihkan. AMI dapat disalin di dalam atau di antara Wilayah. Atau, Anda dapat menggunakan AWS Backup untuk menyalin cadangan di antara akun dan ke Wilayah AWS lainnya. Kemampuan pencadangan lintas-akun membantu melindungi dari peristiwa bencana yang mencakup ancaman orang dalam atau kebocoran keamanan akun. AWS Backup juga menambahkan kemampuan tambahan untuk cadangan EC2—selain volume EBS individual untuk instans, AWS Backup juga menyimpan dan melacak metadata berikut: jenis instans, Virtual Private Cloud (VPC) yang dikonfigurasi, grup keamanan, IAM role, konfigurasi pemantauan, dan tag. Namun, metadata tambahan ini hanya digunakan saat memulihkan cadangan EC2 ke Wilayah AWS yang sama.

Setiap data yang tersimpan di Wilayah pemulihan bencana sebagai cadangan harus dipulihkan pada saat failover. AWS Backup menawarkan kemampuan pemulihan, tetapi saat ini tidak memungkinkan pemulihan terjadwal atau otomatis. Anda dapat menerapkan pemulihan otomatis ke wilayah DR menggunakan SDK AWS untuk memanggil API AWS Backup. Anda dapat menyiapkannya sebagai tugas yang berulang secara rutin atau memicu pemulihan setiap kali pencadangan selesai. Gambar berikut menunjukkan contoh pemulihan otomatis menggunakan Amazon Simple Notification Service (Amazon SNS) dan AWS Lambda.



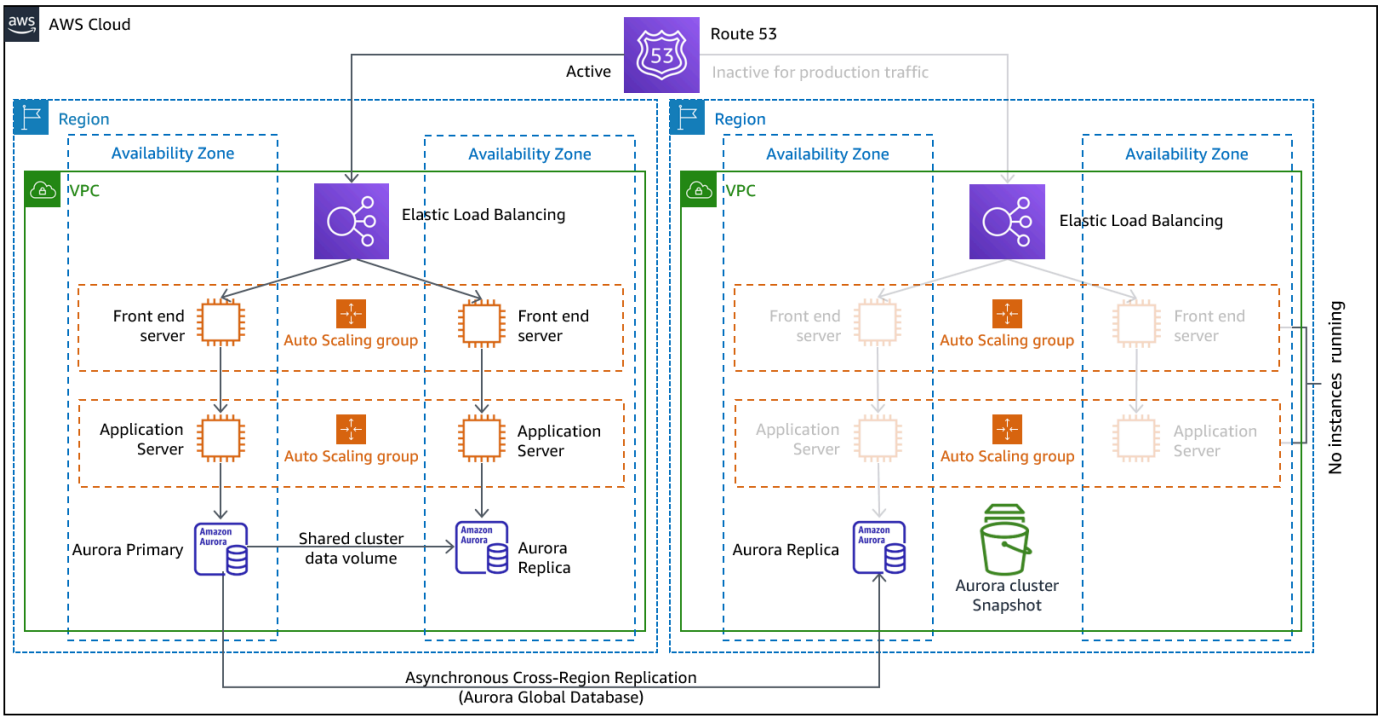
Gambar 8 - Memulihkan dan menguji cadangan

> **Note**
>
> Strategi pencadangan Anda harus menyertakan pengujian cadangan Anda. Lihat bagian Menguji Pemulihan Bencana untuk informasi lebih lanjut. Lihat AWS Well-Architected Lab: Menguji Pencadangan dan Pemulihan Data untuk demonstrasi implementasi langsung.

Pilot light

Dengan pendekatan pilot light, Anda mereplikasi data Anda dari satu Wilayah ke Wilayah lainnya dan menyediakan salinan infrastruktur beban kerja inti Anda. Sumber daya yang diperlukan untuk

mendukung replikasi data dan cadangan, seperti basis data dan penyimpanan objek, selalu aktif. Elemen lain, seperti server aplikasi, dimuat dengan kode aplikasi dan konfigurasi, tetapi dimatikan dan hanya digunakan selama pengujian atau ketika failover pemulihan bencana dipanggil. Berbeda dengan pendekatan pencadangan dan pemulihan, infrastruktur inti Anda selalu tersedia dan Anda selalu memiliki opsi untuk menyediakan lingkungan produksi skala penuh dengan mengaktifkan dan menskalakan keluar server aplikasi Anda.



Gambar 9 - Arsitektur pilot light

Pendekatan pilot light meminimalkan biaya pemulihan bencana yang berkelanjutan dengan meminimalkan sumber daya aktif, dan menyederhanakan pemulihan pada saat bencana karena semua persyaratan infrastruktur inti sudah terpenuhi. Opsi pemulihan ini mengharuskan Anda untuk mengubah pendekatan deployment Anda. Anda perlu membuat perubahan infrastruktur inti ke setiap Wilayah dan men-deploy perubahan beban kerja (konfigurasi, kode) secara bersamaan ke setiap Wilayah. Langkah ini dapat disederhanakan dengan mengotomatisasi deployment Anda dan menggunakan infrastruktur sebagai kode (IAC) untuk men-deploy infrastruktur di sejumlah akun dan Wilayah (deployment infrastruktur penuh ke Wilayah utama dan deployment infrastruktur yang berskala rendah/dimatikan ke wilayah DR). Sebaiknya Anda menggunakan akun yang berbeda per Wilayah untuk memberikan tingkat isolasi sumber daya dan keamanan tertinggi (jika kredensial yang mengalami kebocoran keamanan juga merupakan bagian dari rencana pemulihan bencana Anda).

Dengan pendekatan ini, Anda juga harus mengurangi dampak bencana data. Replikasi data berkelanjutan melindungi Anda dari sebagian jenis bencana, tetapi mungkin tidak melindungi Anda dari kerusakan atau pemusnahan data kecuali jika strategi Anda juga mencakup versioning data yang disimpan atau opsi untuk pemulihan titik waktu (PITR). Anda dapat mencadangkan data yang direplikasi di Wilayah bencana untuk membuat cadangan point-in-time di Wilayah yang sama.

Layanan AWS

Selain menggunakan layanan AWS yang tercakup dalam bagian Pencadangan dan Pemulihan untuk membuat cadangan point-in-time, pertimbangkan juga layanan berikut untuk strategi pilot light Anda.

Untuk pilot light, replikasi data terus-menerus ke basis data dan penyimpanan data aktif di wilayah DR adalah pendekatan terbaik untuk RPO rendah (jika digunakan di samping pencadangan point-in-time yang dibahas sebelumnya). AWS menyediakan replikasi data asinkron lintas wilayah yang berkelanjutan untuk data menggunakan layanan dan sumber daya berikut:

- Replikasi Amazon Simple Storage Service (Amazon S3)
- Replika baca Amazon RDS
- Basis data global Amazon Aurora
- Tabel global Amazon DynamoDB

Dengan replikasi berkelanjutan, versi data Anda segera tersedia di Wilayah DR Anda. Waktu replikasi aktual dapat dipantau menggunakan fitur layanan seperti S3 Replication Time Control (S3 RTC) untuk objek S3 dan fitur manajemen basis data global Amazon Aurora.

Ketika melakukan failover untuk menjalankan beban kerja baca/tulis Anda dari Wilayah pemulihan bencana, Anda harus mempromosikan replika baca RDS untuk menjadi instans utama. Untuk instans DB selain Aurora, prosesnya membutuhkan waktu beberapa menit untuk selesai, termasuk rebooting. Untuk Cross-Region Replication (CRR) dan failover dengan RDS, penggunaan basis data global Amazon Aurora memberikan beberapa keuntungan. Basis data global menggunakan infrastruktur khusus yang membuat basis data Anda sepenuhnya tersedia untuk melayani aplikasi Anda, dan dapat mereplikasi ke Wilayah sekunder dengan latensi standar di bawah satu detik (dan dalam Wilayah AWS, jauh lebih rendah dari 100 milidetik). Dengan basis data global Amazon Aurora, jika Wilayah utama Anda mengalami penurunan atau gangguan performa, Anda dapat mempromosikan salah satu wilayah sekunder untuk mengambil tanggung jawab baca/tulis dalam waktu kurang dari 1 menit bahkan jika terjadi gangguan regional yang menyeluruh. Promosi ini bisa otomatis dan tidak ada reboot.

Versi yang berskala rendah dari infrastruktur beban kerja inti Anda dengan sumber daya yang lebih sedikit atau lebih kecil harus di-deploy di Wilayah DR Anda. Dengan menggunakan AWS CloudFormation, Anda dapat menentukan infrastruktur Anda dan men-deploy-nya secara konsisten di seluruh akun AWS dan di seluruh Wilayah AWS. AWS CloudFormation menggunakan parameter semu yang telah ditetapkan untuk mengidentifikasi akun AWS dan Wilayah AWS tempat layanan ini digunakan. Oleh karena itu, Anda dapat menerapkan logika kondisi di templat CloudFormation Anda untuk men-deploy hanya versi berskala rendah dari infrastruktur Anda di Wilayah DR. Untuk deployment instans EC2, Amazon Machine Image (AMI) menyediakan informasi seperti konfigurasi perangkat keras dan perangkat lunak yang diinstal. Anda dapat menerapkan alur Image Builder yang membuat AMI yang Anda butuhkan dan menyalinnya ke Wilayah utama dan cadangan Anda. Ini membantu memastikan bahwa golden AMI ini memiliki semua yang Anda butuhkan untuk men-deploy ulang atau menskalakan keluar beban kerja Anda di wilayah baru, jika terjadi peristiwa bencana. Instans Amazon EC2 digunakan dalam konfigurasi berskala rendah (lebih sedikit instans daripada di Wilayah utama Anda). Anda dapat menggunakan hibernasi untuk mengalihkan instans EC2 ke dalam status berhenti yang tidak mengharuskan Anda membayar biaya EC2, Anda hanya membayar penyimpanan yang digunakan. Untuk memulai instans EC2, Anda dapat membuat skrip menggunakan AWS Command Line Interface (CLI) atau SDK AWS. Untuk menskalakan keluar infrastruktur untuk mendukung lalu lintas produksi, lihat AWS Auto Scaling di bagian Warm Standby.

Untuk konfigurasi aktif/siaga seperti pilot light, semua lalu lintas awalnya pergi ke Wilayah utama dan beralih ke Wilayah pemulihan bencana jika Wilayah utama tidak lagi tersedia. Ada dua opsi manajemen lalu lintas yang perlu dipertimbangkan dengan menggunakan layanan AWS. Opsi pertama adalah menggunakan Amazon Route 53. Menggunakan Amazon Route 53, Anda dapat mengaitkan sejumlah titik akhir IP di satu atau beberapa Wilayah AWS dengan nama domain Route 53. Kemudian, Anda dapat mengarahkan lalu lintas ke titik akhir yang sesuai dengan nama domain tersebut. Pemeriksaan kondisi Amazon Route 53 memantau titik akhir ini. Dengan menggunakan pemeriksaan kondisi ini, Anda dapat mengonfigurasi failover DNS untuk memastikan lalu lintas dikirim ke titik akhir yang berkondisi baik.

Pilihan kedua adalah dengan menggunakan AWS Global Accelerator. Dengan menggunakan AnyCast IP, Anda dapat mengaitkan sejumlah titik akhir dalam satu atau beberapa Wilayah AWS dengan alamat IP statis yang sama. AWS Global Accelerator kemudian merutekan lalu lintas ke titik akhir yang sesuai yang terkait dengan alamat itu. Pemeriksaan kondisi Accelerator Global memantau titik akhir. Dengan menggunakan pemeriksaan kondisi ini, AWS Global Accelerator secara otomatis memeriksa kondisi aplikasi Anda dan mengarahkan lalu lintas pengguna hanya ke titik akhir aplikasi yang berkondisi baik. Global Accelerator menawarkan latensi yang lebih rendah ke titik akhir aplikasi karena memanfaatkan jaringan edge AWS yang luas untuk menempatkan lalu lintas di tulang

punggung jaringan AWS sesegera mungkin. Global Accelerator juga menghindari masalah caching yang dapat terjadi dengan sistem DNS (seperti Route 53).

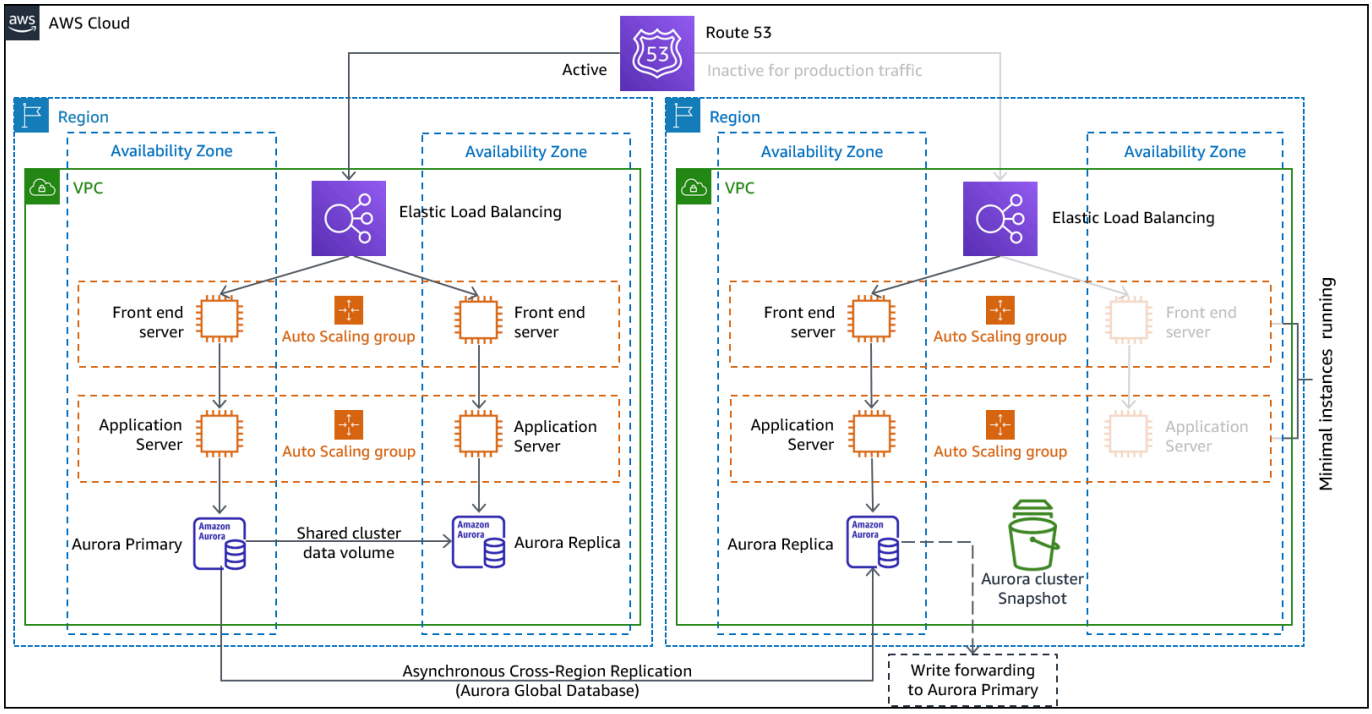
CloudEndure Disaster Recovery

CloudEndure Disaster Recovery, yang tersedia dari AWS Marketplace, terus mereplikasi aplikasi yang di-host server dan basis data yang di-host server dari sumber mana pun ke AWS menggunakan replikasi tingkat blok terhadap server yang mendasarinya. CloudEndure Disaster Recovery memungkinkan Anda menggunakan AWS Cloud sebagai Wilayah pemulihan bencana untuk beban kerja on-premise dan lingkungannya. Ini juga dapat digunakan untuk pemulihan bencana beban kerja yang di-host AWS jika hanya terdiri dari aplikasi dan basis data yang di-host di EC2 (yaitu bukan RDS). CloudEndure Disaster Recovery menggunakan strategi Pilot Light, dengan memelihara salinan data dan sumber daya yang dimatikan di Amazon Virtual Private Cloud (Amazon VPC) yang digunakan sebagai area persiapan. Ketika peristiwa failover dipicu, sumber daya yang dipersiapkan akan digunakan untuk secara otomatis membuat deployment kapasitas penuh di Amazon VPC target yang digunakan sebagai lokasi pemulihan.

Gambar 10 - CloudEndure Disaster Recovery

Warm standby

Pendekatan warm standby berfokus untuk memastikan bahwa ada salinan lingkungan produksi Anda yang berskala rendah, namun berfungsi penuh, di Wilayah lain. Pendekatan ini memperluas konsep pilot light dan mengurangi waktu pemulihan karena beban kerja Anda selalu aktif di Wilayah lain. Pendekatan ini juga memungkinkan Anda untuk lebih mudah melakukan pengujian atau menerapkan pengujian berkelanjutan untuk meningkatkan kepercayaan pada kemampuan Anda untuk pulih dari bencana.



Gambar 11 - Arsitektur warm standby

Catatan: Perbedaan antara pilot light dan warm standby terkadang sulit dimengerti. Keduanya mencakup sebuah lingkungan di Wilayah DR Anda dengan salinan aset Wilayah utama Anda. Perbedaannya adalah bahwa pilot light tidak dapat memproses permintaan tanpa tindakan tambahan yang diambil terlebih dahulu, sedangkan warm standby dapat menangani lalu lintas (pada tingkat kapasitas yang berkurang) dengan segera. Pendekatan pilot light mengharuskan Anda untuk "menghidupkan" server, mungkin men-deploy infrastruktur tambahan (non-inti), dan menaikkan skala, sedangkan Warm Standby hanya mengharuskan Anda untuk menaikkan skala (semuanya sudah di-deploy dan berjalan). Gunakan kebutuhan RTO dan RPO Anda untuk membantu Anda memilih di antara pendekatan ini.

Layanan AWS

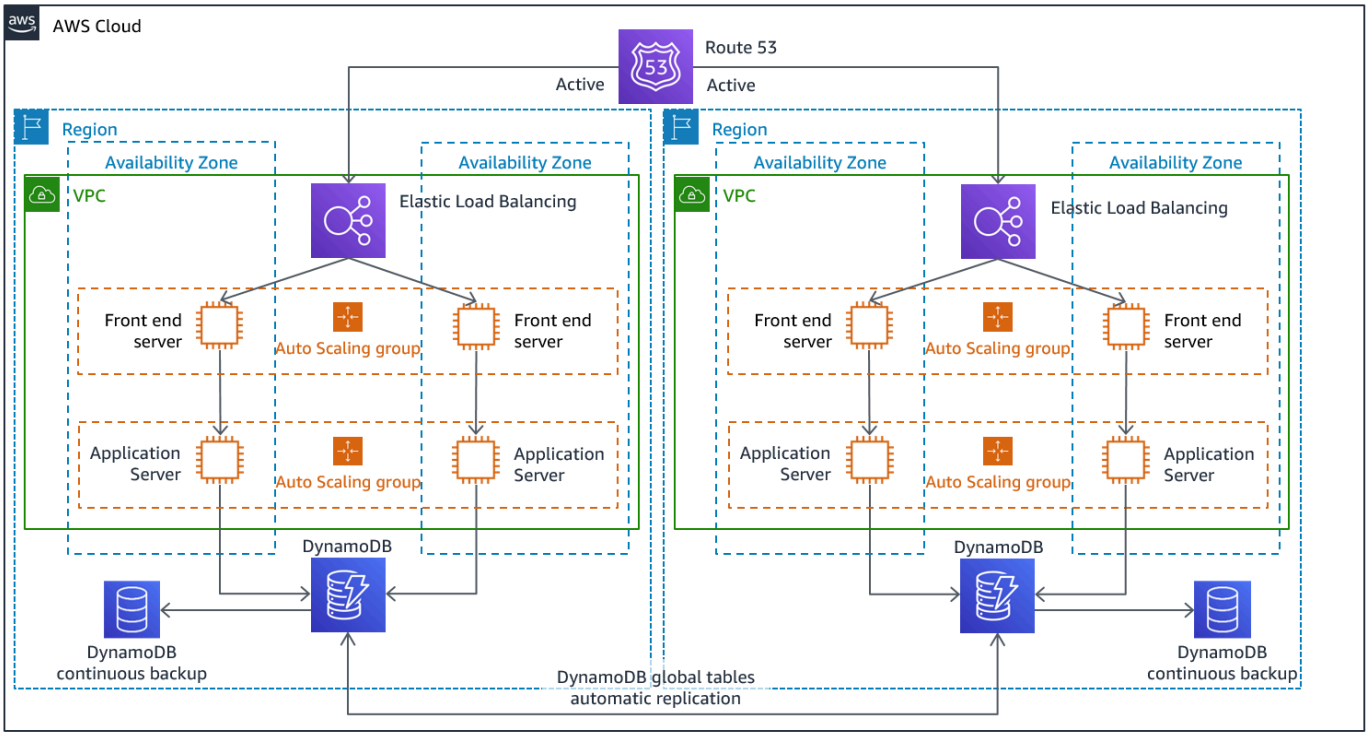
Semua layanan AWS yang tercakup dalam pencadangan dan pemulihan dan pilot light juga digunakan dalam warm standby untuk pencadangan data, replikasi data, perutean lalu lintas aktif/siaga, dan deployment infrastruktur termasuk instans EC2.

AWS Auto Scaling digunakan untuk menskalakan sumber daya termasuk instans Amazon EC2, tugas Amazon ECS, throughput Amazon DynamoDB, dan replika Amazon Aurora dalam Wilayah AWS. Amazon EC2 Auto Scaling menskalakan deployment instans EC2 di seluruh Zona Ketersediaan

dalam Wilayah AWS, yang memberikan ketahanan dalam Wilayah tersebut. Gunakan Auto Scaling untuk menskalakan keluar Wilayah DR Anda untuk kemampuan produksi penuh, sebagai bagian dari strategi pilot light atau warm standby. Misalnya, untuk EC2, tingkatkan pengaturan kapasitas yang diinginkan pada grup Auto Scaling. Anda dapat menyesuaikan pengaturan ini secara manual melalui AWS Management Console, secara otomatis melalui SDK AWS, atau dengan men-deploy ulang templat AWS CloudFormation Anda menggunakan nilai kapasitas baru yang diinginkan. Anda dapat menggunakan parameter AWS CloudFormation untuk mempermudah deployment ulang templat CloudFormation. Pastikan kuota layanan di Wilayah DR Anda ditetapkan cukup tinggi sehingga tidak membatasi Anda untuk menskalakan ke atas hingga kapasitas produksi.

Multi-lokasi aktif/aktif

Anda dapat menjalankan beban kerja Anda secara bersamaan di sejumlah Wilayah sebagai bagian dari strategi multi-lokasi aktif/aktif atau hot standby aktif/pasif. Multi-lokasi aktif/aktif melayani lalu lintas dari semua wilayah tempat strategi ini di-deploy, sedangkan hot standby melayani lalu lintas hanya dari satu wilayah, dan Wilayah lainnya hanya digunakan untuk pemulihan bencana. Dengan pendekatan multi-lokasi aktif/aktif, pengguna dapat mengakses beban kerja Anda di Wilayah mana pun tempat strategi ini di-deploy. Pendekatan ini adalah pendekatan yang paling kompleks dan mahal untuk pemulihan bencana, tetapi dapat mengurangi waktu pemulihan Anda mendekati nol untuk sebagian besar bencana dengan pilihan dan implementasi teknologi yang benar (namun kerusakan data mungkin akan memerlukan cadangan, yang biasanya menghasilkan titik pemulihan non-nol). Hot standby menggunakan konfigurasi aktif/pasif yang hanya mengarahkan pengguna ke satu wilayah, dan wilayah DR tidak mengambil lalu lintas. Sebagian besar pelanggan menemukan bahwa jika mereka akan menyiapkan lingkungan penuh di Wilayah kedua, masuk akal untuk menggunakannya dengan konfigurasi aktif/aktif. Atau, jika Anda tidak ingin menggunakan kedua Wilayah untuk menangani lalu lintas pengguna, maka Warm Standby menawarkan pendekatan yang lebih ekonomis dan secara operasional tidak kompleks.



Gambar 12 - Arsitektur multi-lokasi aktif/aktif (ubah satu jalur Aktif menjadi Tidak Aktif untuk hot standby)

Dengan multi-lokasi aktif/aktif, karena beban kerja berjalan di lebih dari satu Wilayah, tidak akan ada failover yang diperlukan dalam skenario ini. Pengujian pemulihan bencana dalam hal ini akan berfokus pada bagaimana beban kerja bereaksi terhadap kehilangan suatu Wilayah: Apakah lalu lintas dialihkan jauh dari Wilayah yang gagal? Dapatkah Wilayah lain menangani semua lalu lintas? Pengujian untuk bencana data juga diperlukan. Pencadangan dan pemulihan masih diperlukan dan harus diuji secara teratur. Perlu juga dicatat bahwa waktu pemulihan untuk bencana data yang mencakup kerusakan, penghapusan, atau obfuskasi data akan selalu lebih besar dari nol dan titik pemulihan akan selalu berada di titik tertentu sebelum bencana ditemukan. Jika kompleksitas dan biaya tambahan dari pendekatan multi-lokasi aktif/aktif (atau hot standby) diperlukan untuk mempertahankan waktu pemulihan hampir nol, maka upaya tambahan harus dilakukan untuk menjaga keamanan dan mencegah kesalahan manusia agar dapat mengurangi bencana manusia.

Layanan AWS

Semua layanan AWS yang tercakup dalam pencadangan dan pemulihan, pilot light, dan warm standby juga digunakan di sini untuk pencadangan data point-in-time, replikasi data, perutean lalu lintas aktif/aktif, dan deployment dan penskalaan infrastruktur termasuk instans EC2.

Untuk skenario aktif/pasif yang dibahas sebelumnya (Pilot Light dan Warm Standby), Amazon Route 53 dan AWS Global Accelerator dapat digunakan untuk merutekan lalu lintas jaringan ke wilayah aktif. Untuk strategi aktif/aktif di sini, kedua layanan ini juga memungkinkan definisi kebijakan yang menentukan pengguna mana yang dialihkan ke titik akhir regional aktif tertentu. Dengan AWS Global Accelerator, Anda menetapkan dial lalu lintas untuk mengontrol persentase lalu lintas yang diarahkan ke setiap titik akhir aplikasi. Amazon Route 53 mendukung pendekatan persentase ini, dan juga sejumlah kebijakan lain yang tersedia termasuk yang berbasis geoproksimitas dan latensi. Global Accelerator secara otomatis memanfaatkan jaringan server edge AWS yang luas, untuk melakukan onboarding lalu lintas ke tulang punggung jaringan AWS sesegera mungkin, sehingga menghasilkan latensi permintaan yang lebih rendah.

Replikasi data dengan strategi ini memungkinkan RPO mendekati nol. Layanan AWS, seperti basis data global Aurora , menggunakan infrastruktur khusus yang membuat basis data Anda tersedia sepenuhnya untuk melayani aplikasi, dan dapat mereplikasi ke satu wilayah sekunder dengan latensi standar di bawah satu detik. Dengan strategi aktif/pasif, penulisan hanya terjadi pada Wilayah utama. Perbedaan dengan aktif/aktif adalah merancang bagaimana menangani penulisan untuk setiap Wilayah aktif. Biasanya pembacaan pengguna dirancang agar disajikan dari Wilayah yang terdekat, yang dikenal sebagai read local. Dengan penulisan, Anda memiliki beberapa opsi:

- Strategi write global akan merutekan semua penulisan ke satu Wilayah. Jika Wilayah itu gagal, Wilayah lain akan dipromosikan untuk menerima penulisan. Basis data global Aurora sangat cocok untuk write global, karena mendukung sinkronisasi dengan replika baca di seluruh Wilayah, dan Anda dapat mempromosikan salah satu Wilayah sekunder untuk mengambil tanggung jawab baca/tulis dalam waktu kurang dari 1 menit.
- Strategi write local merutekan penulisan ke Wilayah terdekat (seperti pembacaan). Tabel global Amazon DynamoDB memungkinkan strategi seperti itu, sehingga memungkinkan pembacaan dan penulisan dari setiap wilayah, tempat tabel global Anda di-deploy. Tabel global Amazon DynamoDB menggunakan rekonsiliasi penulis terakhir diprioritaskan jika ada pembaruan yang bersamaan.
- Strategi write partitioned menetapkan penulisan ke Wilayah tertentu berdasarkan kunci partisi (seperti ID pengguna) untuk menghindari konflik penulisan. Replikasi Amazon S3 yang dikonfigurasi secara dua arah dapat digunakan untuk kasus ini, dan saat ini mendukung replikasi antara dua Wilayah. Saat menerapkan pendekatan ini, pastikan untuk mengaktifkan sinkronisasi modifikasi replika pada kedua bucket A dan B untuk mereplikasi perubahan metadata seperti daftar kontrol akses (ACL) objek, tag objek, atau kunci objek pada objek yang direplikasi. Anda juga dapat mengonfigurasi apakah akan mereplikasi delete marker di antara bucket di Wilayah aktif

Anda. Selain replikasi, strategi Anda juga harus menyertakan pencadangan point-in-time untuk melindungi terhadap peristiwa kerusakan atau pemusnahan data.

AWS CloudFormation adalah alat yang ampuh untuk menegakkan infrastruktur yang di-deploy secara konsisten di antara akun AWS di sejumlah Wilayah AWS. AWS CloudFormation StackSets memperluas fungsionalitas ini dengan memungkinkan Anda membuat, memperbarui, atau menghapus tumpukan CloudFormation di sejumlah akun dan Wilayah dengan satu operasi. Meskipun AWS CloudFormation menggunakan YAML atau JSON untuk mendefinisikan Infrastruktur sebagai Kode (IaC), AWS Cloud Development Kit (AWS CDK) memungkinkan Anda untuk mendefinisikan Infrastruktur sebagai Kode (IaC) menggunakan bahasa pemrograman yang sudah dikenal. Kode Anda dikonversi ke CloudFormation yang kemudian digunakan untuk men-deploy sumber daya di AWS.

Deteksi

Penting untuk mengetahui sesegera mungkin bahwa beban kerja Anda tidak memberikan hasil bisnis yang seharusnya diberikan. Dengan cara ini, Anda dapat dengan cepat menyatakan sebuah bencana dan melakukan pemulihan dari sebuah insiden. Untuk sasaran pemulihan agresif, waktu respons ini ditambah dengan informasi yang tepat sangat penting dalam memenuhi sasaran pemulihan. Jika sasaran titik pemulihan Anda adalah satu jam, maka Anda perlu mendeteksi insiden tersebut, memberi tahu personel yang sesuai, menjalankan proses eskalasi Anda, mengevaluasi informasi (jika ada) terkait waktu yang diharapkan untuk pemulihan (tanpa menjalankan rencana DR), menyatakan sebuah bencana, dan melakukan pemulihan dalam waktu satu jam.

> **Note**
>
> Jika para pemangku kepentingan memutuskan untuk tidak menjalankan DR meskipun RTO akan berisiko tidak tercapai, maka evaluasi kembali rencana dan sasaran DR. Keputusan untuk tidak menjalankan rencana DR mungkin karena rencananya tidak memadai atau ada kurangnya keyakinan dalam pelaksanaan.

Penting untuk memperhitungkan deteksi insiden, pemberitahuan, eskalasi, penemuan bencana, dan pernyataan bencana dalam perencanaan dan sasaran Anda untuk memberikan sasaran realistis yang dapat dicapai dan memberikan nilai bisnis.

AWS menerbitkan informasi terbaru tentang ketersediaan layanan di Service Health Dashboard. Periksa dasbor ini kapan saja untuk mendapatkan informasi status saat ini, atau berlangganan umpan RSS untuk diberi tahu tentang gangguan pada setiap layanan. Jika Anda mengalami masalah operasional waktu nyata dengan salah satu layanan kami yang tidak ditampilkan di Service Health Dashboard, Anda dapat membuat Permintaan Dukungan.

AWS Health Dashboard menyediakan informasi tentang peristiwa AWS Health yang dapat memengaruhi akun Anda. Informasi ini disajikan dalam dua cara: dasbor yang menunjukkan peristiwa terbaru dan mendatang yang disusun berdasarkan kategori, dan log peristiwa lengkap yang menunjukkan semua peristiwa dari 90 hari terakhir.

Untuk persyaratan RTO yang paling ketat, Anda dapat menerapkan failover otomatis berdasarkan pemeriksaan kondisi. Rancang pemeriksaan kondisi yang merepresentasikan pengalaman pengguna dan berdasarkan Indikator Performa Utama. Pemeriksaan kondisi yang mendetail akan menguji

fungsionalitas utama beban kerja Anda dan akan lebih efektif daripada pemeriksaan metrik pengoperasian yang umum. Gunakan pemeriksaan kondisi yang mendetail berdasarkan sejumlah sinyal. Berhati-hatilah dengan pendekatan ini agar Anda tidak memicu alarm palsu karena melakukan failover yang sebenarnya tidak diperlukan dapat dengan sendirinya menimbulkan risiko ketersediaan.

Menguji pemulihan bencana

Uji implementasi pemulihan bencana untuk memvalidasi implementasi dan uji failover secara teratur ke Wilayah DR beban kerja Anda untuk memastikan RTO dan RPO terpenuhi.

Pola yang harus dihindari adalah mengembangkan jalur pemulihan yang jarang dieksekusi. Misalnya, Anda mungkin memiliki penyimpanan data sekunder yang digunakan untuk kueri hanya-baca. Ketika Anda menulis ke sebuah penyimpanan data dan penyimpanan data utamanya gagal, Anda sebaiknya melakukan failover ke penyimpanan data sekunder. Jika Anda tidak sering menguji failover ini, Anda mungkin menemukan bahwa asumsi Anda tentang kemampuan penyimpanan data sekunder tidak benar. Kapasitas sekunder, yang mungkin sudah cukup ketika terakhir diuji, mungkin tidak lagi dapat menoleransi beban di bawah skenario ini, atau kuota layanan di wilayah sekunder mungkin tidak cukup.

Pengalaman kami telah menunjukkan bahwa satu-satunya pemulihan kesalahan yang efektif adalah jalur yang sering Anda uji. Itulah mengapa memiliki sejumlah kecil jalur pemulihan adalah pilihan terbaik.

Anda dapat menetapkan pola pemulihan dan secara teratur mengujinya. Jika Anda memiliki jalur pemulihan yang kompleks atau kritis, Anda masih perlu secara teratur menjalankan failover dalam produksi untuk memvalidasi bahwa jalur pemulihan berfungsi.

Kelola pergeseran konfigurasi di Wilayah DR. Pastikan infrastruktur, data, dan konfigurasi Anda sudah sesuai dengan kebutuhan di Wilayah DR. Misalnya, periksa apakah AMI dan kuota layanan sudah terbaru.

Anda dapat memanfaatkan AWS Config untuk terus memantau dan mencatat konfigurasi sumber daya AWS Anda. AWS Config dapat mendeteksi pergeseran dan memicu AWS Systems Manager Automation untuk memperbaiki pergeseran dan mengaktifkan alarm. AWS CloudFormation juga dapat mendeteksi pergeseran dalam tumpukan yang telah Anda deploy.

Kesimpulan

Pelanggan bertanggung jawab atas ketersediaan aplikasi mereka di cloud. Penting untuk mendefinisikan apa yang dianggap sebagai bencana dan memiliki rencana pemulihan bencana yang mencerminkan definisi ini dan dampaknya terhadap hasil bisnis. Buat Sasaran Waktu Pemulihan (RTO) dan Sasaran Titik Pemulihan (RPO) berdasarkan analisis dampak dan penilaian risiko dan kemudian memilih arsitektur yang sesuai untuk mengurangi efek dari bencana. Pastikan deteksi bencana dapat dilakukan dan tepat waktu — sangat penting untuk mengetahui kapan sasaran berisiko gagal tercapai. Pastikan Anda memiliki rencana dan memvalidasi rencana dengan pengujian. Rencana pemulihan bencana yang belum divalidasi akan berisiko tidak dilaksanakan karena kurangnya keyakinan atau gagal memenuhi sasaran pemulihan bencana.

Kontributor

Kontributor dokumen ini meliputi:

- Alex Livingstone, Kepala Praktik Operasi Cloud (Practice Lead Cloud Operations), AWS Enterprise Support
- Seth Eliot, Arsitek Solusi Utama untuk Keandalan (Principal Reliability Solutions Architect), Amazon Web Services

Bacaan lebih lanjut

Untuk informasi tambahan, lihat:

- Pilar Keandalan, AWS Well-Architected Framework
- Daftar Periksa Rencana Pemulihan Bencana
- Menerapkan Pemeriksaan Kondisi
- Penerapan Solusi AWS: Multi-Region Application Architecture
- AWS re:Invent 2018: Pola Arsitektur untuk Aplikasi Aktif-Aktif Multi-Wilayah (ARC209-R2)

Riwayat dokumen

Perubahan	Deskripsi	Tanggal
Publikasi awal	Publikasi pertama.	12 Februari 2021

Untuk menerima pemberitahuan tentang pembaruan laporan resmi ini, berlangganan umpan RSS.

Pemberitahuan

Pelanggan bertanggung jawab untuk membuat penilaian independen mereka sendiri atas informasi dalam dokumen ini. Dokumen ini: (a) hanya disediakan sebagai informasi, (b) berisi penawaran produk dan praktik AWS saat ini, yang dapat berubah tanpa pemberitahuan, dan (c) tidak menjadi komitmen atau jaminan apa pun dari AWS dan afiliasi, pemasok, atau pemberi lisensinya. Produk atau layanan AWS disediakan “sebagaimana adanya” tanpa jaminan, representasi, atau ketentuan apa pun, baik tersurat maupun tersirat. Tanggung jawab dan kewajiban AWS kepada pelanggannya dikendalikan oleh perjanjian AWS, dan dokumen ini bukan bagian dari, juga tidak mengubah, perjanjian apa pun antara AWS dan pelanggannya.