AWS Recent Announcements RSS

Recent Announcements
The AWS Cloud platform expands daily. Learn about announcements, launches, news, innovation and more from Amazon Web Services.

Announcing sticky session routing for Amazon SageMaker Inference

Today, we are announcing the availability of sticky session routing on Amazon SageMaker Inference which helps customers improve the performance and user experience of their generative AI applications by leveraging their previously processed information. Amazon SageMaker makes it easier to deploy ML models including foundation models (FMs) to make inference requests at the best price performance for any use case.

By enabling sticky sessions, all requests for the same session will be routed to the same instance, allowing your ML application to reuse previously processed information to reduce latency and improve user experience. This is particularly valuable when customers want to use large data payloads or have the need for seamless interactive experiences. By leveraging their previous inference requests, customers can now take advantage of this feature to build innovative state-aware AI applications on SageMaker. To do this customers will have to create a session id with their first request and then use that session id to indicate that SageMaker should route all the subsequent requests to the same instance. Sessions can also be deleted when done to free up resources for new sessions.

This feature is available in all regions where SageMaker is available. You can learn more about deploying models on SageMaker here and more about this feature in our documentation.

AWS Network Firewall now supports AWS PrivateLink

AWS Network Firewall now supports AWS PrivateLink. Customers can now access and manage their Network Firewalls privately, without going through the public internet. AWS PrivateLink provides private connectivity between VPCs, AWS services, and on-premises applications, securely over the Amazon network. When AWS PrivateLink is used with AWS Network Firewall, all management and control traffic between clients and Network Firewall flows over a private network.

AWS Network Firewall is a managed firewall service that makes it easy to deploy essential network protections for all your Amazon VPCs. Customers can use AWS PrivateLink with Network Firewall in regions where AWS Network Firewall is available today, including the AWS GovCloud (US) Regions. For more information about the AWS Regions where AWS Network Firewall is available, see the AWS Region table.

To learn more about configuring AWS Network Firewall, please refer to the service documentation.

Amazon EC2 R6in and R6idn instances are now available in an additional region

Starting today, Amazon Elastic Compute Cloud (Amazon EC2) R6in and R6idn instances are available in AWS Region Asia Pacific (Sydney). These sixth-generation network optimized instances, powered by 3rd Generation Intel Xeon Scalable processors and built on the AWS Nitro System, deliver up to 200Gbps network bandwidth, 2x more network bandwidth, and up to 2x higher packet-processing performance over comparable fifth-generation instances. Customers can use R6in and R6idn instances to scale the performance and throughput of network-intensive workloads such as memory-intensive SQL and NoSQL databases, distributed web scale in-memory caches (Memcached, Redis), in-memory databases (SAP HANA), and real-time big data analytics (Apache Hadoop, Apache Spark).

R6in and R6idn instances are available in 10 different instance sizes including metal, with up to 128 vCPUs and 1024 GiB of memory. They deliver up to 100 Gbps of Amazon Elastic Block Store (EBS) bandwidth, and up to 400K IOPS. R6in and R6idn instances offer Elastic Fabric Adapter (EFA) networking support on 32xlarge and metal sizes. R6idn instances offer up to 7.6 TB of high-speed, low-latency instance storage.

With this regional expansion, R6in and R6idn instances are available in the following AWS Regions: US East (Ohio), US East (N. Virginia, Oregon), Europe (Ireland, Frankfurt, Stockholm), Asia Pacific (Singapore, Sydney, Tokyo), and AWS GovCloud (US-West). Customers can purchase the new instances through Savings Plans, Reserved, On-Demand, and Spot instances. To learn more, see R6in and R6idn instances page. To get started, see AWS Management Console, AWS Command Line Interface (AWS CLI), and AWS SDKs.

AWS Application Migration Service supports Trend Micro post-launch action

Starting today, AWS Application Migration Service (AWS MGN) provides an action for installing the Trend Micro Vision One Server & Workload Protection Agent on your migrated instances. For each migrated server, you can choose to automatically install the agent to support your security needs.

Application Migration Service minimizes time-intensive, error-prone manual processes by automating the conversion of your source servers to run natively on AWS. It also helps simplify modernization of your migrated applications by allowing you to select preconfigured and custom optimization options during migration.

This feature is now available in all of the Commercial regions where Application Migration Service is available. Access the AWS Regional Services List for the most up-to-date availability information.

To start using Application Migration Service for free, sign in through the AWS Management Console. For more information, visit the Application Migration Service product page.

For more information on Trend Micro and to create a trial account, visit the Trend Micro sign-up page.

AWS Elemental Media Services now support live AV1 encoding

The AV1 video codec is now supported in the AWS Elemental Media Services. You can use AV1 in AWS Elemental MediaLive, MediaPackage, MediaTailor, and MediaConvert to produce both live and on-demand streams with ad insertion.

This launch enables live streaming and packaging of AV1 encoded content, preparation of AV1 VOD assets and ads, and ad insertion into AV1 encoded videos. AV1 provides a lower bitrate with a similar viewing experience when compared to HEVC and AVC, reducing both the bandwidth required to deliver live events and CDN costs. AV1 can also deliver better video quality for viewing on mobile devices and in network constrained environments.

AWS Media Services enable you to transport, prepare, process, and deliver live and on-demand content in the cloud. These managed services let you build and adapt video workflows quickly, eliminate capacity planning, easily scale with growth, and benefit from pay-as-you-go pricing. Connect with other AWS services and third-party applications for live and on demand video streaming, media storage, machine learning, content protection, advertising and monetization, and more.

To learn how the AV1 codec can reduce bandwidth and improve the viewing experience, read the blog post. For more information about live AV1 pricing, please review the AWS Elemental MediaLive pricing page. To learn more about the services, please visit the AWS Media Services page.

Announcing AWS Elemental MediaLive Anywhere for live video encoding on your own hardware

Today, AWS announces the general availability of AWS Elemental MediaLive Anywhere, which allows you to run live video transcoding on your on-premises hardware. MediaLive Anywhere brings the cloud control and pay-as-you-go pricing of AWS Elemental MediaLive to compute resources you manage. With MediaLive Anywhere, you can take advantage of MediaLive's centralized configuration, control, and monitoring capabilities while processing live video on premises close to video sources and outputs.

With MediaLive Anywhere, you deploy the same broadcast-grade video encoding engine used in AWS Elemental MediaLive on your hardware. MediaLive Anywhere supports a wide range of hardware configurations and can ingest video from multicast, SDI, and standard internet-based transport protocols. You get a consistent set of APIs, channel profiles, logs, and monitoring metrics across your cloud and on-premises live video workflows.

AWS Elemental MediaLive Anywhere is available in all AWS Regions where AWS Elemental MediaLive is available.

To learn more, visit the AWS Elemental MediaLive Anywhere page.

Amazon ECR announces support for dual-layer server-side encryption in the AWS GovCloud (US) Regions

Amazon Elastic Container Registry (ECR) now supports dual-layer server-side encryption in the AWS GovCloud (US) Regions. This capability allows you to apply two independent layers of server-side encryption to images stored in Amazon ECR. Dual-layer server-side encryption with keys stored in AWS Key Management Service (DSSE-KMS) enables you to meet stronger compliance and regulatory requirements of applying multiple layers of encryption to your container images.

ECR supports server-side encryption of ECR images using either Amazon S3-managed encryption keys or keys stored in Amazon Key Management Service (KMS). This often meets your security requirements as it protects data at rest, however, if you operate in highly regulated environments that require rigorous security standards, you may require a second layer of encryption for your images. Now with DSSE-KMS, you can easily apply two layers of encryption and control the keys used for both layers. Once this feature is enabled, ECR automatically encrypts your images twice when pushed and decrypts twice when pulled using your encryption keys managed by Amazon Key Management Service (KMS). AWS KMS is a simple to use key management service that makes it easy for you to create, manage, and control keys by setting permissions per key and specifying key rotation schedules.

DSSE-KMS with ECR is available for use in the AWS GovCloud (US) Regions at an additional cost. For pricing information, visit the Amazon ECR pricing page. To learn more about all available encryption options on Amazon ECR and get started with this feature, visit our user guide.

AWS IAM Identity Center now supports language and visual mode preferences in the AWS access portal

The AWS access portal provides AWS IAM Identity Center users with single sign-on access to all their assigned AWS applications and AWS accounts. Today, AWS IAM Identity Center added support for user preferences on language and visual mode in the AWS access portal.

When customers need to access AWS applications and resources through the AWS access portal, having the ability to work with their preferred language and visual mode improves their efficiency and comfort. AWS access portal now supports 12 different language options, allowing users to work in their most natural and comfortable language. Users can now switch the visual mode of their AWS access portal to dark mode, helping reduce eye strain and improving readability in bright environments. AWS access portal inherits language and visual mode preferences from browser settings by default and allow users to customize further if needed.

The language and visual mode preferences for AWS access portal are available in all AWS regions where AWS IAM Identity Center is available.

IAM Identity Center helps customers connect or create their workforce identities, and manage their access to multiple AWS applications and AWS accounts. IAM Identity Center is available to customers at no additional cost. To learn more about IAM Identity Center, visit the product detail page. To get started with using the AWS access portal, please refer to the user guide.

Amazon EventBridge Pipes now supports customer managed KMS keys

Amazon EventBridge Pipes now supports AWS Key Management Service (KMS) customer managed keys, allowing you to encrypt Pipes filter patterns, enrichment parameters, and target parameters with your own keys instead of default AWS owned keys. Using keys that you create, own, and manage can satisfy your organization’s security and governance requirements.

Amazon EventBridge lets you use events to connect application components, making it easier to build scalable event-driven applications. EventBridge Pipes provides a simple, consistent, and cost-effective way to create point-to-point integrations between event producers and consumers. Pipes enables you to send data from one of 6 different event sources to any of the 20+ targets supported by the EventBridge Event Bus, including HTTPS endpoints through EventBridge API Destinations and event buses themselves. With support for Customer Managed Keys, you have more fine-grained security control over your Pipe’s configuration data to more easily meet your organization's regulatory and compliance requirements. You can also audit and track usage of your encryption keys with AWS CloudTrail.

Customer managed key support for EventBridge Pipes is available in all AWS Regions where EventBridge Pipes is available.

To get started, follow the directions provided in the EventBridge Pipes documentation. To learn more about customer managed keys, visit the AWS Key Management Service documentation.

Amazon Redshift now supports altering sort keys on tables in zero-ETL integration

Amazon Redshift now lets you alter sort keys of tables replicated through zero-ETL integration. Sort keys play a crucial role in determining the physical sorting order of rows within a table, and optimizing them can significantly enhance query performance, especially for queries using range-bound filters on sort key columns.

Amazon Redshift's zero-ETL integration helps you derive holistic insights across many applications and break down data silos in your organization, making it simpler to analyze data from different operational databases. You can now modify the sort keys of your tables replicated through the zero-ETL integration, and achieve faster and more efficient querying of your replicated data in Amazon Redshift. Furthermore, you can even set the sort key of zero-ETL tables to AUTO and allow Amazon Redshift to observe your workload and automatically set a sort key based on your evolving workload and data patterns.

To learn more and get started with zero-ETL integration, visit the getting started guides for Amazon Redshift. To learn more about how Amazon Redshift sort's data, see documentation.

Amazon EKS support in Amazon SageMaker HyperPod to scale foundation model development

We are excited to announce the general availability of Amazon EKS support in SageMaker HyperPod which enables customers to run and manage their Kubernetes workloads on SageMaker HyperPod, a purpose-built infrastructure for foundation model (FM) development which reduces time to train models by up to 40%.

Many customers use Kubernetes to orchestrate their ML workflows due to its portability, scalability, and rich ecosystem of tools. These customers want to continue using Kubernetes' familiar interface, but still want an automated way to manage hardware failures. EKS support in HyperPod combines the benefits of SageMaker HyperPod offering self-healing performant clusters with the containerization capabilities of Amazon EKS, a managed Kubernetes service. With this launch, customers can run deep health checks during cluster creation to reduce failures during training. Further, HyperPod automatically replaces faulty nodes and resumes training from your last checkpoint on both AWS Trainium and Nvidia GPU at a scale of more than a thousand accelerators. Customers have the flexibility to use either the new HyperPod CLI, or their preferred tools, to submit, manage, and monitor workloads. The persistent cluster environment offers ssm access and the ability to customize the cluster. EKS orchestrated HyperPod clusters also integrate with CloudWatch Container Insights to provide out-of-the-box observability, by auto-discovering HyperPod node health status and visualizing them in curated dashboards.

This release is generally available in the AWS Regions where SageMaker HyperPod is available except Europe (London).

To learn more, see the following list of resources: Webpage, AWS News Blog, Documentation, Github repository.

Amazon EMR on EC2 improves cluster launch experience with intelligent subnet selection

Starting today, Amazon EMR on EC2 offers improved reliability and cluster launch experience for instance fleet clusters through enhanced subnet selection. With this feature, EMR on EC2 reduces cluster launch failures caused due to IP address shortages.

Amazon EMR is a cloud big data platform for data processing, interactive analysis, and machine learning using open-source frameworks such as Apache Spark, Apache Hive, and Presto. Previously, the subnet selection for EMR clusters only considered the available IP addresses for the core instance fleet. Amazon EMR now employs subnet filtering at cluster launch and selects one of the subnets that have adequate available IP addresses to successfully launch all instance fleets. If EMR cannot find a subnet with sufficient IP addresses to launch the whole cluster, it will prioritize the subnet that can at least launch the core and primary instance fleets. In this scenario, EMR will also publish a CloudWatch warning event to notify the user. If none of the configured subnets can be used to provision core and primary fleet, EMR will fail the cluster launch and provide a critical error event. These CloudWatch events enables you to monitor your clusters and take remedial actions as necessary.

Customers will benefit from this feature on all EMR 5.12.1 and later releases when launching EMR instance fleet clusters using allocation strategies. No further action is needed from your end. This capability is available in all AWS Regions, including the AWS GovCloud (US) Regions, where Amazon EMR on EC2 is available. To learn more, please refer to the documentation here.

Container Insights now announces SageMaker HyperPod node health observability on EKS

Amazon CloudWatch Container Insights now auto-discovers the health status of your SageMaker HyperPod nodes running on EKS and visualizes them in curated dashboards to help you monitor your node availability for operational excellence. Using out-of-the-box dashboards, you can identify unhealthy nodes easily and mitigate quickly to achieve efficient training durations.

Container Insights works with SageMaker to collect deep health check test results for HyperPod nodes and displays them in preset dashboards to help you understand the health and performance of your nodes, and identify if they are ready for scheduling. Container Insights assists you in optimizing training durations by classifying failing nodes as "pending reboot" and "pending replacement," and guiding you on maintaining node health in case automatic node replacement is disabled. If auto-recovery is enabled, you can gain visibility into your node mutations, delays in your training jobs, and understand how your tasks resume from the last check-point.

Getting started with Container Insights is easy. You can onboard either by installing CloudWatch Observability EKS Add-on or the latest CloudWatch agent into your clusters, or upgrading your Helm charts with the latest CloudWatch Agent version. Once configured you can navigate to Container Insights console and view your SageMaker Hyperpod node health status out-of-the-box.

SageMaker HyperPod node health observability is now available in Container Insights for EKS in all commercial regions where SageMaker HyperPod is present. HyperPod node health metrics follow observation based pricing – see Container Insights pricing page for details. For further information, see the Container Insights user guide.

Amazon MSK enhances cross-cluster replication with support for identical topic names

Amazon MSK Replicator now supports a new configuration that enables you to preserve original Kafka topic names while replicating streaming data across Amazon Managed Streaming for Apache Kafka (Amazon MSK) clusters. Amazon MSK Replicator is a feature of Amazon MSK that lets you reliably replicate data across MSK clusters in the same or different AWS region(s) with just a few clicks. The new configuration reduces the need for you to reconfigure client applications during setup and makes it even more simple to operate multi-cluster streaming architectures, while continuing to benefit from MSK Replicator's reliability.

With Amazon MSK Replicator, you can easily build regionally resilient streaming applications for business continuity, share data with partners, aggregate data from multiple clusters for analytics, and serve clients globally with lower latency. With the new configuration, you can retain topic names during replication while automatically avoiding the risk of infinite replication loops that comes with using third-party or open-source tools for replication. If you setup active-passive cluster architecture to build regionally resilient streaming applications, where one cluster handles live traffic while another acts as a standby, the new configuration also streamlines the failover process. Applications can seamlessly failover to the standby cluster without requiring reconfiguration, as topic names remain intact.

Support for the new configuration is available in all regions where Amazon MSK Replicator is available. To see all the regions where Amazon MSK Replicator is available, see the AWS Region table. To learn more, visit our developer guide or product page.

Amazon OpenSearch Service now supports OpenSearch version 2.15

You can now run OpenSearch version 2.15 in Amazon OpenSearch Service. With OpenSearch 2.15, we have made several improvements in the areas of search performance, query optimization, and added capabilities to help you to build AI-powered applications with greater flexibility and ease.

This launch includes radial search that allows you to search points in a vector space that reside within a specified maximum distance or minimum score threshold from a query point, offering greater flexibility for various applications like anomaly detection and geospatial searches. In addition, this release includes performance optimizations like two-phase processor for neural sparse search, and conditional scoring logic and optimized data handling for hybrid search. These performance improvements now allow you to run complex queries on larger datasets more efficiently.

OpenSearch now supports reindex workflow, allowing users to enable vector and hybrid search on existing indexes to reduce time and resources spent on re-indexing from source indexes. In addition, you can configure remote models to serve as guardrails to detect harmful, offensive, or inappropriate content (toxicity) more accurately. Finally, a new ML inference processor enables users to enrich ingest pipelines using inferences from OpenSearch-provided pretrained models.

For information on upgrading to OpenSearch 2.15, please see the documentation. OpenSearch 2.15 is now available in all AWS Regions where Amazon OpenSearch Service is available.

AWS Elastic Beanstalk adds support for IPv6 inbound traffic to service endpoints

AWS Elastic Beanstalk now supports dual-stack public service endpoints and dual-stack VPC endpoints, including VPC endpoints integrated with AWS PrivateLink.

This capability allows you to configure your Elastic Beanstalk VPC endpoints to accept dual-stack incoming traffic (via IPv6 and IPV4). You can also send requests to the Elastic Beanstalk service using the AWS CLI or the AWS SDK specifying an IPv4 endpoint or a dual-stack endpoint. For a list of public endpoints, see Elastic Beanstalk service endpoints in the Amazon Web Services General Reference.

Elastic Beanstalk support for IPv6 and IPv4 dual-stack functionality is available in all of the AWS Commercial Regions and AWS GovCloud (US) Regions that Elastic Beanstalk supports. For a complete list of regions and service offerings, see AWS Regions.

For more information about Elastic Beanstalk dual-stack traffic support, see IPV6 support in the AWS Elastic Beanstalk Developer Guide.

Amazon Aurora now supports R7g Graviton3-based instance family in 15 additional regions

AWS Graviton3-based R7g database instances are now generally available for Amazon Aurora with PostgreSQL compatibility and Amazon Aurora with MySQL compatibility in 15 additional regions, including US West (N. California), Canada (Central), South America (Sao Paulo), Europe (Stockholm), Europe (Frankfurt), Europe (London), Europe (Milan), Europe (Spain), Asia Pacific (Mumbai), Asia Pacific (Hyderabad), Asia Pacific (Seoul), Asia Pacific (Tokyo), Asia Pacific (Singapore), Asia Pacific (Sydney), and Asia Pacific (Hong Kong). AWS Graviton3 instances provide up to 30% performance improvement and up to 20% price/performance improvement over Graviton2 instances for Amazon Aurora, depending on the database engine version and workload.

AWS Graviton3 processors are custom-designed AWS Graviton processors built on the AWS Nitro System. The Graviton3 processors offer several improvements over the second-generation Graviton processors. R7g database instances offer up to 30Gbps enhanced networking bandwidth and up to 20 Gbps of bandwidth to the Amazon Elastic Block Store (Amazon EBS).

You can spin up Graviton3 R7g database instances in the Amazon RDS Management Console or using the AWS CLI. Upgrading a database instance to Graviton3 requires a simple instance type modification. For more details, refer to the Aurora documentation.

Amazon Aurora is designed for unparalleled high performance and availability at global scale with full MySQL and PostgreSQL compatibility. It provides built-in security, continuous backups, serverless compute, up to 15 read replicas, automated multi-Region replication, and integrations with other AWS services. To get started with Amazon Aurora, take a look at our getting started page.

Secondary sensor support for AWS IoT SiteWise Edge through CloudRail

Today, we're announcing the general availability of secondary sensor support for AWS IoT SiteWise. Through an integration with AWS Partner CloudRail, customers can now ingest data from over 12,000 sensors from vendors like ifm, SICK, Turck, and Pepperl+Fuchs using either IO-Link or Modbus TCP/IP protocols. Secondary sensors enable data collection from isolated brownfield equipment and for customers to digitally integrate it with their other operational data. Previously, ingesting data from brownfield equipment required either upgrading equipment or manual processes for data collection resulting in manual errors, additional cost and time to value.

Through a simple drop-down selection in the AWS Console, users add AWS Partner CloudRail software as a data source on their AWS IoT SiteWise Edge gateway and configure the desired sensor signals and protocols in the partner application. After deploying configurations, the equipment data flows to AWS IoT SiteWise Edge for local monitoring, storage, and access and on to AWS IoT SiteWise for integration with other industrial data and other AWS Cloud services.

AWS IoT SiteWise is a managed service that makes it easy to collect, store, organize and monitor data from industrial equipment at scale. AWS IoT SiteWise Edge extends cloud capabilities to on-premises applications.

This feature is generally available in the following AWS Regions: US East (N. Virginia), US East (Ohio), US West (Oregon), Europe (Ireland), Europe (Frankfurt), Asia Pacific (Mumbai), Asia Pacific (Tokyo), Asia Pacific (Singapore), Asia Pacific (Seoul), Asia Pacific (Sydney) and Canada (Central).

To learn more, visit the AWS IoT SiteWise user guide.

Amazon IVS Real-Time Streaming now supports RTMP ingest

Starting today, you can use RTMP (Real-Time Messaging Protocol) and the encrypted version, RTMPS, to broadcast to your Amazon Interactive Video Service (Amazon IVS) stages. This new protocol complements the currently supported WHIP (WebRTC-HTTP Ingestion Protocol). RTMP ingest enhances compatibility with a wide range of software and hardware encoders for increased flexibility in your broadcasting.

Amazon IVS is a managed live streaming solution that is designed to be quick and easy to set up, and ideal for creating interactive video experiences. Video ingest and delivery are available around the world over a managed network of infrastructure optimized for live video. Visit the AWS region table for a full list of AWS Regions where the Amazon IVS console and APIs for control and creation of video streams are available.

To learn more, please visit the Amazon IVS RTMP ingest documentation page.

AWS IoT SiteWise Edge adds support for 100+ protocols through Litmus Edge

Today, we're announcing the general availability of expanded industrial protocol support for AWS IoT SiteWise. Through a new integration with AWS Partner Litmus, customers can now ingest data from 100+ additional industrial protocols including proprietary protocols from companies like Allen-Bradley, Beckhoff, Emerson, Fanuc, Mitsubishi, Omron, and Yaskawa along with many others. Previously, ingesting data from these protocols required acquiring, provisioning, and configuring infrastructure and middleware for data collection resulting in additional cost and time to value.

Through a simple drop-down selection in the AWS Console, users add AWS Partner Litmus Edge software as a data source on their AWS IoT SiteWise Edge gateway. Users then configure the protocols, build data flows, and configure data processing in the partner application. After configurations are deployed, the equipment data flows to AWS IoT SiteWise Edge for local monitoring, storage, and access. It is also sent to AWS IoT SiteWise for integration with other industrial data and usage in other AWS Cloud services.

AWS IoT SiteWise is a managed service that makes it easy to collect, store, organize and monitor data from industrial equipment at scale. AWS IoT SiteWise Edge extends the cloud capabilities to on-premises applications.

This new feature is generally available in the following AWS Regions: US East (N. Virginia), US East (Ohio), US West (Oregon), Europe (Ireland), Europe (Frankfurt), Asia Pacific (Mumbai), Asia Pacific (Tokyo), Asia Pacific (Singapore), Asia Pacific (Seoul), Asia Pacific (Sydney) and Canada (Central).

To learn more, visit the AWS IoT SiteWise user guide.

AWS Resource Access Manager now supports AWS PrivateLink

AWS Resource Access Manager (AWS RAM) now supports AWS PrivateLink, allowing you to create and manage your resource shares from within your Amazon Virtual Private Cloud (VPC) without traversing the public internet.

AWS RAM helps you securely share your resources across your organization, with specific organizational units (OUs), or with individual AWS accounts. You can centrally create a resource and then share that resource using AWS RAM to reduce the operational overhead of managing resources in a multi-account environment.

AWS RAM support for AWS PrivateLink is available in the AWS Commercial Regions, the AWS GovCloud (US) Regions, and the China Regions. To get started with using AWS RAM to share resources, visit the AWS Resource Access Manager Console.

AWS Elemental MediaConnect adds support for input thumbnail images

You can now monitor your sources and get instant visual feedback for AWS Elemental MediaConnect flows with thumbnail images via the AWS Management Console or API. Thumbnails are also available in the Workflow Monitor tool.

Input thumbnails provide a visual representation of your live content. Rather than relying solely on metadata or metrics, you can now glance at a thumbnail to verify that your sources are operating as expected. This makes it easier to detect issues, troubleshoot problems, confirm the right input source is being sent, and ensure the quality of your live video. For more information on thumbnail access, visit the MediaConnect documentation. Input thumbnails in MediaConnect are available at no additional cost.

AWS Elemental MediaConnect is a reliable, secure, and flexible transport service for live video that enables broadcasters and content owners to build live video workflows and securely share live content with partners and customers. MediaConnect helps customers transport high-value live video streams into, through, and out of the AWS Cloud. MediaConnect can function as a standalone service or as part of a larger video workflow with other AWS Elemental Media Services, a family of services that form the foundation of cloud-based workflows to transport, transcode, package, and deliver video.

Visit the AWS Region Table for a full list of AWS Regions where MediaConnect is available. To learn more about MediaConnect, please visit here.

Amazon EC2 P5e instances are generally available via EC2 Capacity Blocks

Today, AWS announces the general availability of Amazon Elastic Compute Cloud (Amazon EC2) P5e instances, powered by the latest NVIDIA H200 Tensor Core GPUs. Available via EC2 Capacity Blocks, these instances deliver the highest performance in Amazon EC2 for deep learning and generative AI inference.

You can use Amazon EC2 P5e instances for training and deploying increasingly complex large language models (LLMs) and diffusion models powering the most demanding generative AI applications. You can also use P5e instances to deploy demanding HPC applications at scale in pharmaceutical discovery, seismic analysis, weather forecasting, and financial modeling.

P5e instances feature 8 H200 GPUs which have 1.7x GPU memory size and 1.5x GPU memory bandwidth than H100 GPUs featured in P5 instances. They provide market-leading scale-out capabilities for distributed training and tightly coupled HPC workloads with up to 3,200 Gbps of networking using second-generation Elastic Fabric Adapter (EFA) technology. To address customer needs for large scale at low latency, P5e instances are deployed in Amazon EC2 UltraClusters.

P5e instances are now available in the US East (Ohio) AWS Region in the p5e.48xlarge sizes through EC2 Capacity Blocks for ML.

To learn more about P5e instances, see Amazon EC2 P5e Instances.

Amazon Kinesis Data Streams now supports FIPS 140-3 enabled interface VPC endpoint

Starting today, Amazon Kinesis Data Streams supports adding a VPC endpoint using AWS PrivateLink that connects through the regional endpoint that has been validated under the Federal Information Processing Standard (FIPS) 140-3 program. With this new launch, you can easily use AWS PrivateLink with Kinesis Data Streams for those regulated workloads that require a secure connection using a FIPS 140-3 validated cryptographic module.

FIPS compliant endpoints helps companies contracting with the US federal governments meet the FIPS security requirement to encrypt sensitive data in supported Regions. To create an interface VPC endpoint that connects to a Kinesis Data Streams FIPS endpoint, see using Amazon Kinesis Data Streams with Interface VPC Endpoints.

This new capability is available in all AWS Regions in the United States, including the AWS GovCloud (US) Regions. To learn more about AWS PrivateLink, see accessing AWS services through AWS PrivateLink. To learn more about FIPS 140-3 at AWS, visit FIPS 140-3 Compliance.

Amazon RDS Custom for SQL Server is now available in 3 additional AWS Regions

Amazon RDS Custom for SQL Server is a managed database service that allows you to bring your own licensed SQL Server media and have access to the underlying operating system and database environment. This service is now available in the AWS Regions of US West (N. California), Asia Pacific (Osaka), and Europe (Paris).

By using Amazon RDS Custom for SQL Server, you can benefit from the agility of a managed database service, including features such as managed high availability (MAZ), automated backups and point-in-time recovery, and cross-region snapshot copying. You can choose to deploy RDS Custom for SQL Server using your own licensed SQL Server media or with License Included (LI) instances. It can help you save time on the undifferentiated heavy lifting of database management and focus on more business-impacting tasks.

To learn more about Amazon RDS Custom, please see the Amazon RDS Custom for SQL Server User Guide and Amazon RDS Custom Pricing for pricing details and availability in the additional regions.

CloudWatch Application Signals now supports request based Service Level Objectives (SLOs)

CloudWatch Application Signals, which helps troubleshoot issues quickly by providing out-of-the-box dashboards that correlate telemetry across metrics, traces, and logs for your applications and their dependencies, now supports Service Level Objectives (SLOs) calculated based on the request count i.e. the fraction of good or bad requests of the total operations performed by any service. This allows for more precise tracking of how many requests met the defined SLO criteria (e.g., latency under 200ms) out of the total requests received by a service.

SLOs represent a commitment to maintain a certain level of service quality, typically expressed as percentage attainment of performance goals within a given time interval. The existing offering of period-based SLOs measures performance during each time period in an interval and aggregates proportion of good periods that meet a specified performance threshold. The release of request-based SLOs allows you to track error tolerance in terms of requests. This enables fine-grained monitoring based on customer traffic, particularly useful for applications prone to high fluctuations in request volume, where performance shortfall during period of low traffic can lead to a quick erosion of the error budget.

Request-based SLOs are available in all regions where Application Signals is generally available - 28 commercial AWS Regions except CA West (Calgary) Region. For pricing, see Amazon CloudWatch pricing.

See SLO documentation to learn more, or refer to the user guide and AWS One Observability Workshop to get started with Application Signals.

PostgreSQL 17 RC1 is now available in Amazon RDS Database preview environment

Amazon RDS for PostgreSQL 17 Release Candidate 1 (RC1) is now available in the Amazon RDS Database Preview Environment, allowing you to evaluate the pre-release of PostgreSQL 17 on Amazon RDS for PostgreSQL. You can deploy PostgreSQL 17 RC1 in the Amazon RDS Database Preview Environment that has the benefits of a fully managed database.

PostgreSQL 17 includes updates to vacuuming that reduces memory usage, improves time to finish vacuuming, and shows progress of vacuuming indexes. With PostgreSQL 17, you no longer need to drop logical replication slots when performing a major version upgrade. PostgreSQL 17 continues to build on the SQL/JSON standard, adding support for `JSON_TABLE` features that can convert JSON to a standard PostgreSQL table. The `MERGE` command now supports the `RETURNING` clause, letting you further work with modified rows. PostgreSQL 17 also includes general improvements to query performance and adds more flexibility to partition management with the ability to SPLIT/MERGE partitions. Please refer to the PostgreSQL community announcement for more details.

Amazon RDS Database Preview Environment database instances are retained for a maximum period of 60 days and are automatically deleted after the retention period. Amazon RDS database snapshots that are created in the preview environment can only be used to create or restore database instances within the Preview Environment. You can use the PostgreSQL dump and load functionality to import or export your databases from the Preview Environment.

Amazon RDS Database Preview Environment database instances are priced as per the pricing in the US East (Ohio) Region.

Amazon ECS now supports AWS Graviton-based Spot compute with AWS Fargate
Amazon Elastic Container Service (Amazon ECS) now supports AWS Graviton-based compute with AWS Fargate Spot. This capability helps you run fault-tolerant Arm-based applications with up to 70% discount compared to Fargate prices. AWS Graviton processors are custom-built by AWS to deliver the best price-performance for cloud workloads.

Amazon ECS with AWS Fargate enables customers to deploy and build workloads at scale in a serverless manner. Customers can get better price-performance by running Arm-based workloads. Starting today, customers can further optimize for costs by running fault-tolerant Arm-based workloads on AWS Fargate Spot. To get started, just configure your task definition just like you do today with cpu-architecture = ARM64 and choose FARGATE_SPOT as the capacity provider to run your Amazon ECS service or a standalone task. Amazon ECS will leverage spare AWS Graviton-based compute capacity available in the AWS cloud for running your service or task. You can now get the simplicity of serverless compute with familiar cost optimization levers of Spot capacity with Graviton-based compute.

This capability is now available for Amazon ECS tasks running on AWS Fargate platform version 1.4.0 or higher in all commercial and the AWS GovCloud (US) Regions. To learn more about usage and pricing of AWS Fargate Spot, see Fargate Capacity Providers documentation and the AWS Fargate pricing page.

Amazon Timestream for InfluxDB is now available in the Canada, London and Paris AWS regions
You can now use Amazon Timestream for InfluxDB in the Canada (Central), Europe (London) and Europe (Paris) AWS regions. Timestream for InfluxDB makes it easy for application developers and DevOps teams to run fully managed InfluxDB databases on AWS for real-time time-series applications using open-source APIs.

To migrate to Timestream for InfluxDB from a self-managed InfluxDB instance, you can use our sample migration script by following this guide. Timestream for InfluxDB offers the full feature set available in the InfluxDB 2.7 release of the open-source version, and adds deployment options with Multi-AZ high availability and enhanced durability. For high availability, Timestream for InfluxDB allows you to automatically create a primary database instance and synchronously replicate the data to an instance in a different Availability Zone. When it detects a failure, Amazon Timestream automatically fails over to a standby instance without manual intervention.

With the latest release, customers can use Amazon Timestream for InfluxDB in the following regions: US East (Ohio), US East (N. Virginia), US West (Oregon), Canada (Central), Asia Pacific (Mumbai), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Tokyo), Europe (Paris), Europe (Frankfurt), Europe (Ireland), Europe (London), and Europe (Stockholm). To get started with Amazon Timestream, visit our product page.

Amazon RDS for MariaDB supports minors 10.11.9, 10.6.19, 10.5.26
Amazon Relational Database Service (Amazon RDS) for MariaDB now supports MariaDB minor versions 10.11.9, 10.6.19, and 10.5.26. We recommend that you upgrade to the latest minor versions to fix known security vulnerabilities in prior versions of MariaDB, and to benefit from the bug fixes, performance improvements, and new functionality added by the MariaDB community.

You can leverage automatic minor version upgrades to automatically upgrade your databases to more recent minor versions during scheduled maintenance windows. You can also leverage Amazon RDS Managed Blue/Green deployments for safer, simpler, and faster updates to your MariaDB instances. Learn more about upgrading your database instances, including automatic minor version upgrades and Blue/Green Deployments, in the Amazon RDS User Guide.

Amazon RDS for MariaDB makes it straightforward to set up, operate, and scale MariaDB deployments in the cloud. Learn more about pricing details and regional availability at Amazon RDS for MariaDB. Create or update a fully managed Amazon RDS database in the Amazon RDS Management Console.

Amazon S3 Access Grants introduce the ListCallerAccessGrants API
Amazon S3 Access Grants now support ListCallerAccessGrants, a new API that allows AWS Identity and Access Management (IAM) principals and AWS IAM Identity Center end users to list all S3 buckets, prefixes, and objects they can access, as defined by their S3 Access Grants. Customers can use ListCallerAccessGrants to build applications that identify and then take action on data that is accessible to specific end users. For example, the Storage Browser for Amazon S3, an open source UI component that customers can add to their applications to provide end users with a simple interface for data stored in S3, uses ListCallerAccessGrants to present end users with the data that they have access to in S3, based on their S3 Access Grants.

S3 Access Grants map identities in AWS IAM or Identity Providers (IdPs) to your datasets in S3. When customers call the ListCallerAccessGrants action, S3 identifies the IAM principal or IAM Identity Center user and their associated groups. The API then returns the S3 Access Grants for the end user and their groups based on group membership in AWS IAM or an IdP.

The ListCallerAccessGrants API is available in all AWS Regions where AWS IAM Identity Center is available. For pricing details, visit Amazon S3 pricing. To learn more about S3 Access Grants, visit the S3 User Guide.

Amazon WorkSpaces Pools now allows you to bring your Windows 10 or 11 licenses
Amazon Web Services announces the ability to setup Amazon WorkSpaces Pools powered by Microsoft Windows 10 and 11 operating systems using Bring Your Own License (BYOL). Now, customers can bring their Windows 10 or 11 licenses (provided they meet Microsoft's licensing requirements) to support their eligible Microsoft 365 Apps for enterprise, providing a consistent desktop experience to their users when they switch between on-premise and virtual desktops.

WorkSpaces Pools simplifies management across a customer's WorkSpaces environment by providing a single console and set of clients to manage the various desktop hardware configurations, storage, and applications for their users. With BYOL, the operating system is hosted on hardware that is dedicated to admin's AWS account and they can automatically scale a pool of virtual desktops based on real-time usage metrics or predefined schedules. WorkSpaces Pools offers pay-as-you-go hourly pricing that excludes any operating system fees, providing significant savings.

To take advantage of this option, the organization must meet the Microsoft licensing requirements, and must commit to running a minimum number of WorkSpaces in a given AWS Region each month. BYOL for Windows 10 and 11 is supported for Standard, Performance, Power and PowerPro bundles. For the best experience with video conferencing we recommend using Power or PowerPro bundles only. For Region availability details, see AWS Regions and Availability Zones for WorkSpaces Pools. To learn more about this option and the eligibility requirements, please see the Amazon WorkSpaces BYOL documentation and FAQs on BYOL.

Amazon Connect Contact Lens now supports new ways to automate agent performance evaluations
You can now automatically mark a performance evaluation question as not applicable based on conversational insights (e.g., detected call reason, etc.), and you can now use additional contact metrics (e.g., longest hold duration, number of holds, agent interaction duration including holds) to automatically fill answers to questions in the evaluation form. With this launch, you can automatically complete only the applicable evaluation questions, under specific conditions. For example, you could check if an agent explained new account benefits and pricing, only for those customers who called to open an account. Additionally, you can automatically evaluate whether the agent was able to resolve the customer's issue efficiently (e.g., resolved the customer's issue within 10 minutes) and did not put the customer repeatedly on hold.

This feature is available in all regions where Contact Lens performance evaluations is already available. To learn more, please visit our documentation and our webpage. For information about Contact Lens pricing, please visit our pricing page.

Amazon EC2 R7i instances are now available in additional AWS region
Starting today, Amazon Elastic Compute Cloud (Amazon EC2) R7i instances are available in Europe (Milan).

Amazon EC2 R7i instances are powered by custom 4th Generation Intel Xeon Scalable processors (code-named Sapphire Rapids), available only on AWS, offer up to 15% better performance over comparable x86-based Intel processors utilized by other cloud providers.

R7i instances deliver up to 15% better price-performance versus R6i instances. These instances are SAP certified and are a great choice for memory-intensive workloads, such as SAP, SQL and NoSQL databases, distributed web scale in-memory caches, in-memory databases like SAP HANA, and real time big data analytics like Hadoop and Spark. They offer larger instance sizes, up to 48xlarge, and two bare metal sizes (metal-24xl, metal-48xl) for high-transaction and latency-sensitive workloads. These bare-metal sizes support built-in Intel accelerators: Data Streaming Accelerator, In-Memory Analytics Accelerator, and QuickAssist Technology, allowing customers to facilitate efficient offload and acceleration of data operations and optimize performance for workloads.

R7i instances support the new Intel Advanced Matrix Extensions (AMX) that accelerate matrix multiplication operations for applications such as CPU-based ML. In addition, customers can now attach up to 128 EBS volumes to an R7i instance (vs 28 EBS volume attachments on R6i). This allows processing of larger amounts of data, scale workloads, and improve performance over R6i instances.

To learn more, visit Amazon EC2 R7i Instances.

Amazon Managed Service for Apache Flink now supports Apache Flink 1.20
Amazon Managed Service for Apache Flink now supports Apache Flink minor version 1.20. This version is expected to be be the last 1.x minor version released by the Flink community before Flink 2.0. We recommend that you upgrade to Flink 1.20 to benefit from bug fixes, performance improvements, and new functionality added by the Flink community. You can use in-place version upgrades for Apache Flink to upgrade your existing application to this new version.

Amazon Managed Service for Apache Flink makes it easier to transform and analyze streaming data in real time with Apache Flink. Apache Flink is an open source framework and engine for processing data streams. Amazon Managed Service for Apache Flink reduces the complexity of building and managing Apache Flink applications and integrates with Amazon Managed Streaming for Apache Kafka (Amazon MSK), Amazon Kinesis Data Streams, Amazon OpenSearch Service, Amazon DynamoDB streams, Amazon S3, custom integrations, and more using built-in connectors. Create or update an Amazon Managed Service for Apache Flink application in the Amazon Managed Service for Apache Flink console.

You can learn more about Apache Flink 1.20 in Amazon Managed Service for Apache Flink in our documentation. To learn more about open source Apache Flink visit the official website. For Amazon Managed Service for Apache Flink region availability, refer to the AWS Region Table.

Amazon RDS for Oracle now supports OEM and OLS options with Multitenant configuration
Amazon RDS for Oracle now supports Oracle Enterprise Manager (OEM) and Oracle Label Security (OLS) option with Oracle Multitenant configuration. OEM enables monitoring and managing the Oracle infrastructure from a single console. OLS provides fine-grained control of access to individual tables or rows, and helps you enforce regulatory compliance with a policy-based administration.

To enable OEM on your Amazon RDS for Oracle database instance with multitenant configuration, you can add an option to a new or existing option group. You can use OEM option for OEM Database Express for a lightweight tool for managing a single database instance, or OEM_AGENT for OEM Cloud Control for an enterprise-level tool for managing a large environment. To learn more about enabling OEM, please refer to Amazon RDS for Oracle documentation.

To enable the OLS option on your Amazon RDS for Oracle database instance with multitenant configuration, you can add the option to a new or existing option group. In order to use the OLS option, you need to have a valid Oracle Label Security option license, in addition to an Oracle Enterprise Edition license with "Software Update License & Support". To learn more about enabling OLS, please refer to Amazon RDS for Oracle documentation.

Amazon RDS for Oracle allows you to set up, operate, and scale Oracle database deployments in the cloud. See Amazon RDS for Oracle Pricing for up-to-date pricing and regional availability.

AWS Gateway Load Balancer now supports configurable TCP idle timeout
Today AWS Gateway Load Balancer (GWLB) is launching a new capability that allows you to align the TCP idle timeout value of GWLB with clients and target appliances. Using this capability you can now perform uninterrupted stateful inspection and fine tuning of the applications that use long-lived flows, such as financial applications, databases and ERP systems, when using GWLB.

Prior to this launch, TCP idle timeout was a fixed value of 350 seconds, which could interrupt long-lived traffic flows of some applications. With this launch, you now have the flexibility to configure GWLB's TCP idle timeout to be a value between 60 seconds and 6000 seconds, with the default remaining at 350 seconds for backward compatibility. This configuration can help reduce interruptions in traffic flows by maintaining target stickiness for the optimal duration based on the needs of your application. You can configure the TCP idle timeout value using the 'tcp.idle_timeout.seconds' listener attribute on your existing and newly created GWLBs.

This capability is available in all AWS commercial and the AWS GovCloud (US) Regions.

For more information on how to use this feature, see this AWS blog post and product documentation.

Amazon RDS Custom for SQL Server supports Cross-region Snapshot Copying

Amazon RDS Custom for SQL Server now supports copying database snapshots, either created automatically or manually, across commercial AWS Regions. This enables you to seamlessly move database snapshots for your RDS Custom for SQL Server instances to different Regions, which can be used to build a robust disaster recovery solution for your mission-critical data.

RDS Custom for SQL Server is a managed database service that allows you to bring your own licensed SQL Server software and customize the underlying operating system. With today's launch, you can manually copy RDS Custom for SQL Server database snapshots to other AWS Regions, enabling data retention, compliance, and disaster recovery for your applications deployed across multiple Regions. For more information see Copying an Amazon RDS Custom for SQL Server DB snapshot.

When you create a snapshot of your RDS Custom for SQL Server instance, it captures the entire database instance, not just individual databases. To copy the snapshot in a different AWS region, select the snapshot, choose "**Copy Snapshot"** from the **Snapshot actions** menu, pick a destination region, and enter a name for the new snapshot. You can initiate the snapshot copy process from the AWS Management Console, the AWS Command Line Interface (CLI), or through the Amazon RDS Custom for SQL Server APIs.

Amazon RDS Custom for SQL Server with cross-region snapshot copying is available in all commercial AWS Regions where RDS Custom for SQL Server is offered.

Announcing Storage Browser for Amazon S3 for your web applications (alpha release)

Amazon S3 is announcing the alpha release of Storage Browser for S3, an open source component that you can add to your web applications to provide your end users with a simple interface for data stored in S3. With Storage Browser for S3, you can provide authorized end users access to easily browse, download, and upload data in S3 directly from your own applications. Storage Browser for S3 is available in the AWS Amplify JavaScript and React client libraries.

Storage Browser for S3 only displays the data that your end users are authorized to access and automatically optimizes requests to deliver high throughput data transfer. You can control access to your data based on your end user's identity using AWS security and identity services or your own managed services. You can also customize Storage Browser for S3 to match your existing application's design and branding.

We are making the alpha release of Storage Browser for S3 available to collect early feedback and incorporate community input into the design and implementation. To get started with Storage Browser for S3, visit the GitHub page.

Bedrock Agents on Sonnet 3.5

Agents for Amazon Bedrock enable developers to create generative AI-based applications that can complete complex tasks for a wide range of use cases and deliver answers based on company knowledge sources. In order to complete complex tasks, with high accuracy, reasoning capabilities of the underlying foundational model (FM) play a critical role.

Today, Amazon Bedrock customers in US East (N. Virginia), US West (Oregon), Europe (Frankfurt), Asia Pacific (Tokyo), and Asia Pacific (Singapore) can leverage Claude 3.5 Sonnet with their Bedrock Agents.

Claude 3.5 Sonnet is Anthropic's latest foundation model and ranks among the best in the world. Claude 3.5 Sonnet delivers improved speed, performance and agentic reasoning compared with Claude 3 Opus. Additionally, with this model, Bedrock Agents now supports the Anthropic recommended tool use for function calling which leads to an improved developer and end user experience.

To learn more, read the Claude in Amazon Bedrock product page and documentation. To get started with Claude 3.5 Sonnet in Amazon Bedrock, visit the Amazon Bedrock console. To learn more about the list of models supported on Bedrock Agents, visit the documentation page.

AWS AppSync enhances API monitoring with new DEBUG and INFO logging levels

Today, AWS announces the addition of DEBUG and INFO logging levels for AWS AppSync GraphQL APIs. These new logging levels provide more granular control over log verbosity and make it easier to troubleshoot your APIs while optimizing readability and costs.

With DEBUG and INFO levels, alongside the existing ERROR and ALL levels, customers now have greater flexibility to capture relevant log information at the appropriate level of detail. This allows customers to more precisely pinpoint and resolve issues by sending just the right amount of information to their Amazon CloudWatch Logs. Customers can now log messages from their code with the "error", "log", and "debug" functions and configure the level at which logs will be sent to CloudWatch Logs on their API. The API logging level can be changed at any time without having to change any resolver or function code. For example, an API's logging level can be set to DEBUG during development and troubleshooting but changed to INFO in production. The logging level can be set to ALL to see additional trace information.

The new logging levels are available in all AWS Regions where AppSync is supported.

To learn more about AppSync's new logging levels and how to implement them in your GraphQL APIs, see the AWS AppSync Developer Guide.

Use Apache Spark on Amazon EMR Serverless directly from Amazon Sagemaker Studio

You can now run petabyte-scale data analytics and machine learning on Amazon EMR Serverless directly from Amazon SageMaker Studio notebooks. EMR Serverless automatically provisions and scales the required resources, allowing you to focus on your data and models without having to configure, optimize, tune, or manage clusters. EMR Serverless automatically installs and configures open source frameworks and provides a performance-optimized runtime that is compatible with and faster than standard open source.

With this release, you can now visually create and browse EMR Serverless applications directly from SageMaker Studio and connect to them in a few simple clicks. Once connected to an EMR Serverless application, you can use Spark SQL, Scala, Python to interactively query, explore and visualize data, and run Apache Spark jobs to process data directly from Studio Notebooks. Jobs run fast because they use EMR's performance-optimized versions of Spark. For e.g. Spark on EMR 7.1 is 4.5x faster than it's open source equivalent. EMR Serverless offers fine-grained automatic scaling, which provisions and quickly scales the compute and memory resources to match the requirements of your application and you pay for only what you use.

These features are supported on SageMaker Distribution 1.10 and above, and are generally available in all AWS Regions where SageMaker Studio is available. To learn more, read the blog Use LangChain with PySpark for Processing documents at massive scale with Amazon SageMaker Studio and EMR Serverless, or the SageMaker Studio documentation here.

Amazon SES announces enhanced onboarding with adaptive setup wizard and Virtual Deliverability Manager

Today, Amazon Simple Email Service (SES) launched enhancements to its onboarding experience to help customers easily discover and activate key SES features. The SES console now features an adaptive setup page that brings recommendations for optimal setup to the forefront. Additionally, the update introduces the option to enable the Virtual Deliverability Manager (VDM) within the initial onboarding wizard, offering maximum guidance from the beginning of the setup process.

Previously, the SES onboarding process focused primarily on the initial steps of authenticating domains and attaining production access. With an on-demand advisor check, this workflow guides customers to begin sending authenticated email that meets mailbox provider requirements with DKIM-alignment, SPF-alignment, and a DMARC policy. Now, the enhanced onboarding experience empowers customers to optimize their email deliverability from the start. Customers can easily configure email monitoring, provision dedicated IP addresses, and adjust sending limits to meet their projected volume - all through the adaptive setup page that detects their current configuration and provides tailored recommendations. With these enhancements, SES customers can be confident they are setting up their email infrastructure for long-term success from the beginning of their SES journey.

SES offers a guided onboarding experience in all AWS regions where SES is available.

For more information, please visit the documentation for getting started with SES.

Stability AI's Top 3 Text-to-Image Models Now Available in Amazon Bedrock

Stable Image Ultra, Stable Diffusion 3 Large (SD3 Large), and Stable Image Core models from Stability AI are now generally available in Amazon Bedrock. These models will empower customers in various industries, including media, marketing, retail, and game development, to generate high-quality visuals with unprecedented speed and precision.

All three of these models are capable of generating stunningly photo-realistic images with exceptional detail, color accuracy, and lifelike lighting. Each model caters to diverse use cases:

- Stable Image Ultra – produces the highest quality, photo-realistic outputs, making it perfect for professional print media and large-format applications. This model excels at rendering exceptional detail and realism.
- Stable Diffusion 3 Large – strikes an ideal balance between generation speed and output quality, making it ideal for creating high-volume, high-quality digital assets like websites, newsletters, and marketing materials.
- Stable Image Core – optimized for fast and affordable image generation, making it great for rapidly iterating on concepts during the ideation phase.

These models enable customers to streamline creative processes, swiftly adapt to market trends, drive innovation through visual brainstorming, and gain a competitive advantage by boosting productivity, reducing costs, and improving visual communication across business functions.

Stability AI's Stable Image Ultra, SD3 Large, and Stable Image Core models are now available in Amazon Bedrock in the US West (Oregon) AWS region. To learn more read the AWS News Blog or visit the Stability AI in Amazon Bedrock product page, and documentation. To get started with SD3 in Amazon Bedrock, visit the Amazon Bedrock console.

Amazon Timestream for InfluxDB now supports enhanced management features

We are excited to announce the launch of enhanced management options for Amazon Timestream for InfluxDB, allowing you to scale your instance sizes up or down as needed and update your deployment configuration between Single-AZ and Multi-AZ, giving you greater flexibility and control over your time-series data processing and analysis.

Timestream for InfluxDB is extensively used in applications that require high-performance time-series data processing and analysis. You can quickly respond to changes in data ingestion rates, query volumes, or other workload fluctuations by scaling your instances sizes up and down, ensuring that your Timestream for InfluxDB instances always have the necessary resources to handle your workload and cost effectively. You can also change your availability configuration by moving between Single-AZ and Multi-AZ configurations depending on your needs and budget. This means you can focus on building and deploying your applications, rather than worrying about instance sizing and management.

Amazon Timestream for InfluxDB is available in the following AWS Regions: US East (Ohio), US East (N. Virginia), US West (Oregon), Asia Pacific (Tokyo), Asia Pacific (Sydney), Asia Pacific (Singapore), Asia Pacific (Mumbai), Europe (Ireland), Europe (Frankfurt), and Europe (Stockholm).

You can create a Amazon Timestream Instance from the Amazon Timestream console, AWS Command line Interface (CLI), or SDK, and AWS CloudFormation. To learn more about compute scaling for Amazon Timestream for InfluxDB, visit the product page, documentation, and pricing page.

Amazon EC2 X2idn instances now available in Middle East (Bahrain) region

Starting today, memory-optimized Amazon Compute Cloud (Amazon EC2) X2idn instances are available in Middle East (Bahrain) region. These instances, powered by 3rd generation Intel Xeon Scalable Processors and built with AWS Nitro System, are designed for memory-intensive workloads. They deliver improvements in performance, price performance, and cost per GiB of memory compared to previous generation X1 instances. These instances are SAP-certified for running Business Suite on HANA, SAP S/4HANA, Data Mart Solutions on HANA, Business Warehouse on HANA, SAP BW /4HANA, and SAP NetWeaver workloads on any database.

With this launch, X2idn are available in the following AWS Regions: US East (Ohio, N. Virginia), US West (Oregon, N. California), Africa (Cape Town), Asia Pacific (Hyderabad, Jakarta, Malaysia, Mumbai, Osaka, Seoul, Singapore, Sydney, Tokyo), China (Beijing, Nginxia), Middle East (Bahrain, Dubai), Europe (Frankfurt, Ireland, London, Milan, Paris, Stockholm, Spain, Zurich), Canada (Central), South America (São Paulo), and the AWS GovCloud (US-East, US-West) Regions. X2idn is available for purchase with Savings Plans, Reserved Instances, Convertible Reserved, On-Demand, and Spot instances, or as Dedicated instances or Dedicated hosts. To learn more, visit the EC2 X2i Instances Page, or connect with your AWS Support contacts.

AWS Fault Injection Service introduces additional safety control

AWS Fault Injection Service (FIS) now provides additional safety control with a safety lever that, when engaged, stops all running experiments and prevents new experiments from starting. Customers can now prevent fault injection during certain time periods, such as sales events or product launches, or in response to application health alarms.

FIS has built-in safety guardrails, including "stop conditions" that automatically stop experiments and remove faults when alarms are triggered. Today we add another important guardrail, the "safety lever". When engaged, a safety lever stops all experiments running in the account in the region, including multi-account experiments. The safety lever will remain engaged until manually disengaged by the customer, such as when the application has returned to a healthy state or a planned peak event has concluded.

Safety levers are generally available in all AWS Regions where FIS is available, including the AWS GovCloud (US) Regions, at no additional costs.

To get started, visit the safety levers user guide.

AWS Glue now provides job queuing

Today, AWS adds job queuing for AWS Glue jobs. This new capability enables you to submit AWS Glue job runs without needing to manage account level quotas and limits.

AWS Glue job queuing monitors your account level quotas and limits. If quotas or limits are insufficient to start a Glue job run, AWS Glue will automatically queue the job and wait for limits to free up. Once limits become available, AWS Glue will retry the job run. Glue jobs will queue for limits like max concurrent job runs per account, max concurrent Data Processing Units (DPU), and resource unavailable due to IP address exhaustion in Amazon Virtual Private Cloud (Amazon VPC). AWS Glue job queuing can be enabled on your jobs via the AWS Management console or API/CLI.

AWS Glue job queuing is available in all AWS commercial regions where AWS Glue is generally available. To learn more, visit the AWS Glue product page, our documentation, or blog post.

Amazon EBS direct APIs now supports IPv6 in AWS PrivateLink

Amazon EBS direct APIs now support the Internet Protocol version 6 (IPv6) protocol when you connect your Virtual Private Cloud (VPC) to EBS Direct APIs using AWS PrivateLink. EBS direct APIs can help customers to simplify their backup and recovery workflows by directly creating and reading EBS snapshots via APIs. Through AWS PrivateLink, customers can access EBS direct APIs as if it were in your VPC. This change can support customers with their IPv6 compliance needs, integrate with existing IPv6-based on-premises applications, and remove the need for expensive networking equipment to handle the address translation between IPv4 and IPv6.

To use this new capability, you can configure your applications to use the Amazon EBS direct API IPv6 endpoints, or dual-stack Amazon EBS direct APIs endpoints which support both IPv4 and IPv6. When you make a request to a dual-stack Amazon EBS direct APIs endpoint, the endpoint resolves to an IPv6 or an IPv4 address, depending on the protocol used by your network and client.

You can access Amazon EBS direct APIs with the IPv6 protocol through AWS PrivateLink in all AWS Regions where EBS direct APIs are available. To learn more about EBS Direct APIs please visit our product page.

Introducing sagemaker-core: A New Object-Oriented SDK for Amazon SageMaker

Amazon SageMaker is excited to announce sagemaker-core, a new Python SDK that provides an object-oriented interface for interacting with SageMaker resources such as TrainingJob, Model, and Endpoint resource classes. The resource chaining feature in sagemaker-core lets developers pass resource objects as parameters, eliminating the need to manually specify complex parameters. The SDK also abstracts low-level details like resource state transitions and polling logic. It achieves full parity with SageMaker APIs, allowing developers to leverage all SageMaker capabilities directly through the SDK. Additional key usability improvements include auto code completion in popular IDEs, comprehensive documentation, and type hints.

The dedicated resource classes in sagemaker-core provide an intuitive object-oriented view of available functionalities, reducing cognitive load for developers and minimizing the need to manage complex parameter structures. Comprehensive documentation, and type hints help developers write code faster and with fewer errors without needing to navigate complex API documentation. By handling resource state management automatically, developers can focus on building and deploying machine learning models without getting bogged down by lower level resource monitoring tasks. When used with intelligent defaults, sagemaker-core alleviates the burden of repeatedly specifying common parameters. The combined effects of these features result in more readable and maintainable code along with increased developer productivity.

To get started, check out our example notebooks and technical documentation. We're excited to bring sagemaker-core to the SageMaker community and look forward to your contributions in making it even better.

Amazon DynamoDB announces support for Attribute-Based Access Control

Amazon DynamoDB now supports Attribute-Based Access Control (ABAC) for tables and indexes. ABAC is an authorization strategy that defines access permissions based on tags attached to users, roles, and AWS resources.

With ABAC, you can now use your tags to configure access permissions and policies. Tag-based access conditions can be used to allow or deny specific actions, when AWS Identity and Access Management (IAM) principals' tags match the tags on an Amazon DynamoDB table. With the flexibility of using tag-based conditions, you can now set more granular access permissions based on your organizational structures. ABAC allows you to scale your tag-based permissions to new employees and changing resource structures, without rewriting policies as organizations grow. ABAC is supported through the AWS Management Console, AWS API, AWS CLI, AWS SDK, and AWS CloudFormation.

Attribute-Based Access Control for Amazon DynamoDB is now available in limited preview in the US East (Ohio), US East (Virginia), and US West (N. California) Regions. To request access to the limited preview, visit the preview page.

AWS announces session reuse with Amazon Redshift Data API

Today, Amazon Redshift launches session reuse feature in Data API that enables you to access data efficiently from Amazon Redshift data warehouses by eliminating the need to manage database drivers, connections, network configurations, data buffering, and more. Data API session reuse allows you to retain the context of a session from one query execution to another, which reduces connection setup latency on repeated queries to the same data warehouse.

With session reuse, you can utilize session context on objects like variables or temporary tables, which you create once and use multiple times for various queries. This reduces the overall execution time for your queries. To reuse sessions, you must specify in seconds how long a session should be kept for after a query finishes and any subsequent queries can reference this session until the time expires, or it's extended.

Amazon Connect now provides a weekly view of agent schedules

Amazon Connect now provides a weekly view of agent schedules, making it easier for contact center managers to get an at-a-glance view of staffing for an entire week. With this launch, you can now ensure there is required coverage each day via daily aggregated metrics including service level, occupancy, and forecasted versus scheduled hours. For example, from the weekly view you can easily identify if there is overstaffing on Wednesday and understaffing on Friday. You can then move agent shifts from Wednesday to Friday within the weekly view. Weekly view also makes it easy to verify that agents receive the appropriate shifts each day (e.g. each agent has an 8-hour shift) and that they are not working too many days in a row (e.g. each agent gets at least 2 days off every week). Weekly view improves manager productivity by reducing time spent on day to day management of agent shifts and makes it easier to review staffing for multiple days in a single view.

This feature is available in all AWS Regions where Amazon Connect agent scheduling is available. To get started with Amazon Connect agent scheduling, click here.

AWS Network Load Balancer now supports configurable TCP idle timeout

Today AWS Network Load Balancer (NLB) is launching a new capability that allows you to align the TCP idle timeout value of NLB with clients and target applications. Using this capability you can now reduce TCP connection retries and latency in applications that use long-lived flows, such as telemetry reporting devices, databases, streaming services and ERP systems, when using NLB.

Prior to this launch, TCP idle timeout was a fixed value of 350 seconds, which could cause TCP connection handshake retries for the long-lived traffic flows of some applications and add latency. With this launch, you now have the flexibility to configure NLB's TCP idle timeout to be a value between 60 seconds and 6000 seconds, with the default remaining at 350 seconds for backward compatibility. This configuration can help reduce latency for long-lived traffic flows by maintaining target stickiness for the optimal duration based on the needs of your application. You can configure the TCP idle timeout value using the 'tcp.idle_timeout.seconds' listener attribute on your existing and newly created NLBs.

This capability is available in all AWS commercial and AWS GovCloud (US) regions.

For more information on how to use this feature, see this AWS blog post and product documentation.

Amazon Connect now offers intraday forecasts

Amazon Connect forecasting, capacity planning, and agent scheduling now includes machine learning (ML) powered intraday forecast capabilities, available within the Amazon Connect Contact Lens dashboards. With intraday forecasts, you receive updates every 15 minutes with predictions for rest-of-day contact volumes, average queue answer time, and average handle time. These forecasts allow you to take proactive actions to improve customer wait time and service level. For example, if contact volume drops below expected levels, contact center managers can use the intraday forecast to predict how long that drop will continue, determine the required staffing levels, and shift the remaining agents into back office work or other higher volume queues.

This feature is available in all AWS Regions where Amazon Connect forecasting, capacity planning, and agent scheduling is available. To learn more see the Amazon Connect Administrator Guide.

AI recommendations for descriptions in Amazon DataZone expanded to more regions

AWS has expanded the AI recommendations for descriptions capability in Amazon DataZone to four new regions: South America (Sao Paulo), Europe (London), Asia Pacific (Sydney), and Canada (Central). This expansion helps improve data discovery, understanding, and usage by enriching the business data catalog. With a single click, data producers can generate comprehensive business data descriptions and context, highlight impactful columns, and include recommendations on analytical use cases.

With AI recommendations for descriptions in Amazon DataZone, data consumers can identify data tables and columns enhancing, which enhances data discoverability and cuts down on back-and-forth communications with data producers. Data consumers (such as data analysts, data engineers, and data scientists) have more contextualized data at their fingertips to inform their analysis. The auto-generated descriptions enable a richer search experience, as search results are now also based on detailed descriptions, possible use cases, and key columns. Data producers can also use APIs to programmatically generate descriptions for assets.

Amazon DataZone AI recommendations for descriptions is generally available in Amazon DataZone domains provisioned in the following AWS Regions: US East (N. Virginia), US West (Oregon), Europe (Frankfurt), Asia Pacific (Tokyo), South America (Sao Paulo), Europe (London), Asia Pacific (Sydney), and Canada (Central).

To learn more, see the Amazon DataZone Automate Data Discovery webpage, and User Guide. For pricing information, see the pricing page.

Amazon Connect Contact Lens can now generate transcriptions in 10 new languages

Amazon Connect Contact Lens can now generate transcriptions in 10 new languages that include Catalan (Spain), Danish (Denmark), Dutch (Netherlands), Finnish (Finland), Indonesian (Indonesia), Malay (Malaysia), Norwegian Bokmål (Norway), Polish (Poland), Swedish (Sweden), and Tagalog /Filipino (Philippines). With this launch, Contact Lens conversational analytics now provides transcription support for 33 languages.

Amazon Connect Contact Lens helps you to monitor, measure, and continuously improve contact quality and agent performance for a better overall customer experience. With Contact Lens conversational analytics, you can transcribe customer calls, analyze customer sentiment, discover top contact drivers, help redact sensitive data, and more, all natively within Amazon Connect.

Transcription support for these 10 new languages is available for post-call scenarios and across all regions where Contact Lens conversational analytics is available. To learn more, please visit our documentation and our webpage. This feature is included with Contact Lens conversational analytics at no additional charge. For information about Contact Lens pricing, please visit our pricing page.

Organizational Units in AWS Control Tower can now contain up to 1,000 accounts

AWS Control Tower now allows you to register Organizational Units (OUs) containing up to 1,000 accounts. With this launch, you can implement governance best practices and standardize configurations across the accounts in your OUs at greater scale. When you register an OU or enable the AWS Control Tower baseline on an OU, member accounts receive best practice configurations, controls, and baseline resources such as AWS IAM roles, AWS CloudTrail, AWS Config, AWS Identity Center, required for AWS Control Tower governance.

Until today, you could only register OUs with 300 accounts or less. Now, you can enroll up to 1,000 AWS accounts under AWS Control Tower governance in a single OU. This allows greater flexibility to preserve your existing OU structure when migrating to AWS Control Tower, and increased ability to scale in-place. Performance enhancements to the OU registration and re-registration processes also enable you to deploy AWS Control Tower baseline resources into your member accounts more efficiently.

The maximum number of accounts in an OU may differ depending on the number enabled controls and the number of regions you have under governance. To learn more, visit Limitations based on underlying AWS services in the AWS Control Tower User Guide. For a full list of AWS regions where AWS Control Tower is available, see AWS Region Table.

Amazon Redshift Serverless now supports AWS PrivateLink

Amazon Redshift Serverless now supports AWS PrivateLink (interface VPC endpoint) to connect to Amazon Redshift Serverless. You can now connect directly to the Amazon Redshift Serverless and Amazon Redshift Serverless API services using AWS PrivateLink in your virtual private cloud (VPC) instead of connecting over the internet.

When you use an AWS PrivateLink, communication between your VPC and Amazon Redshift Serverless is conducted entirely within the AWS network, which can provide greater security and protecting your sensitive information. An AWS PrivateLink endpoint connects your VPC directly to Amazon Redshift Serverless. The instances in your VPC don't need public IP addresses to communicate with the Amazon Redshift Serverless API. To use Amazon Redshift Serverless through your VPC, you have two options. One is to connect from an instance that is inside your VPC. The other is to connect your private network to your VPC by using an AWS VPN option or AWS Direct Connect. You can create an AWS PrivateLink to connect to Amazon Redshift Serverless using the AWS Management Console or AWS Command Line Interface (AWS CLI) commands. For more information, see Creating an Interface Endpoint.

Amazon Redshift Serverless support for AWS PrivateLink is available in all AWS regions where Amazon Redshift Serverless is available.

AWS IoT SiteWise models now support versioning

AWS IoT SiteWise now supports asset model and component model versioning. This new capability is designed to help industrial customers and integrators manage the evolution of their asset models and component models more effectively.

With asset model and component model versioning, industrial customers can now fetch the active version of their asset model and component model. This helps them recover from failed states and avoid losing changes due to conflicting updates. The versioning feature also enables optimistic locking when updating to safely make changes to models without the risk of overwriting each other's work.

The asset models and component models versioning feature is available in all regions where AWS IoT SiteWise is available.

AWS IoT SiteWise is a managed service that makes it easy to collect, store, organize, and monitor data from industrial equipment at scale to help you make data-driven decisions. To learn more about this new feature and how it can benefit your business, please visit the developer guide.

AWS Security Hub launches 8 new security controls

AWS Security Hub has released 8 new security controls, increasing the total number of controls offered to 423. With this release, Security Hub now supports controls for additional AWS services such as Amazon WorkSpaces and AWS DataSync . Security Hub also released new controls against previously supported services like AWS CodeBuild and Amazon Athena. For the full list of recently released controls and the AWS Regions in which they are available, visit the Security Hub user guide.

To use the new controls, turn on the standard they belong to. Security Hub will then start evaluating your security posture and monitoring your resources for the relevant security controls. You can use central configuration to do so across all your organization accounts and linked Regions with a single action. If you are already using the relevant standards and have Security Hub configured to automatically enable new controls, these new controls will run without taking any additional action.

To get started, consult the following list of resources:

- Learn more about Security Hub capabilities and features in the *AWS Security Hub user guide*
- Subscribe to the Security Hub SNS topic to receive notifications about new Security Hub features and controls
- Try Security Hub at no cost for 30 days

AWS Config conformance packs now available in additional AWS Regions

AWS Config conformance packs and organization-level management capabilities for conformance packs and individual AWS Config rules are now available in additional AWS Regions. Conformance packs allow you to bundle AWS Config rules and their associated remediation actions into a single package, simplifying deployment at scale. You can deploy and manage these conformance packs throughout your AWS environment. With this launch, AWS Config conformance packs are now available in 10 additional Regions, and organization-level management capabilities for conformance packs and individual AWS Config rules are now available in 12 additional Regions.

Conformance packs provide a general-purpose compliance framework designed to enable you to create security, operational, or cost-optimization governance checks using managed or custom AWS Config rules and AWS Config remediation actions. This allows you to monitor compliance based on your own groupings and also apply remediation automatically. With this launch, you can also manage the AWS Config conformance packs and individual AWS Config rules at the organization level which simplifies the compliance management across your AWS Organization.

To get started, you can either use the provided sample conformance pack templates or craft a custom YAML file from scratch based on a custom conformance pack. Conformance pack deployment can be done through the AWS Config console, AWS CLI, or via AWS CloudFormation.

You will be charged per conformance pack evaluation in your AWS account per AWS Region. Visit the AWS Config pricing page for more details. To learn more about AWS Config conformance packs, see our documentation.

AWS Backup extends support for Cross-Region backup with Amazon Neptune

Today, we are announcing the availability of AWS Backup support for cross-Region backup of Amazon Neptune backups in Asia Pacific (Hong Kong), Israel (Tel Aviv), and Middle East (Bahrain, UAE). Cross-Region backup enables customers to copy backups from one AWS Region to a different AWS Region, helping increase data resiliency. AWS Backup is a policy-based, fully managed and cost-effective solution that enables you to centralize and automate data protection of Amazon Neptune along with other AWS services (spanning compute, storage, and databases) and third-party applications.

With Cross-Region backup, customers can copy data from a source backup vault to a destination backup vault in another AWS Region, either on-demand or as part of a scheduled backup plan. Cross-Region backup helps customers meet their compliance requirements and disaster recovery needs by storing copies of backup data in a separate Region to their production data. Customers can also recover from backups in the new Region, reducing the risk of downtime and ensuring business continuity requirements are met.

AWS Backup for Amazon Neptune is available in all Regions where Amazon Neptune is available except for Africa (Cape Town) and Asia Pacific (Osaka). With today's launch, cross-Region backup is available for Amazon Neptune backups in all Regions where AWS Backup supports Amazon Neptune. For more information on regional availability, feature availability, and pricing, see the AWS Backup pricing page and the AWS Backup Feature Availability page.

To learn more about AWS Backup support for Amazon Neptune, visit AWS Backup's technical documentation. To get started, visit the AWS Backup console.

Amazon EMR Managed Scaling is now Application Master placement aware

Today, we are excited to announce a new enhancement in EMR Managed Scaling that improves application resiliency and scales the cluster based on executor and ApplicationMasters demand by adding support for Yarn Node Labels. Amazon EMR by default ensures that the processes that controls running jobs and needs to stay alive for the life of the job (ApplicationMasters) can run on both core and task nodes. However, many customers specially who use Spot Instances to run task nodes choose to run ApplicationMasters only on On-Demand core nodes to ensure running jobs do not fail if application masters running on Spot Instances are interrupted. With today's launch, EMR Managed Scaling will now scale the clusters based on the demand for the individual AM's or executors requests as defined by YARN node labels. Intelligently scaling the cluster based on AM's or executors demand leads to better performance, utilization and lower cost.

As part of today's launch, with EMR release 7.2 and later, Amazon EMR will also let you specify application level YARN node labels expressions by market type i.e. On-Demand vs Spot. Previously, customers were only able to specify YARN node labels expressions by node type level i.e. Core vs Task. Now, with this new enhancement customers will have the additional flexibility to better suit the needs of your cluster workloads.

This feature is available with Amazon EMR release 7.2 and above in all the AWS Regions where Amazon EMR Managed Scaling is available. Review our Managed Scaling documentation to learn more.

Amazon Personalize enhances automatic solution training

Amazon Personalize is excited to introduce the ability for developers to modify automatic training configurations after a Personalize solution has been created. With this launch, developers gain greater flexibility over the automatic training process for both new and existing solutions. Previously, changing configurations like training frequency required re-creating the solution entirely. Now, you can easily modify automatic training settings of any solution via API or console. When updating a solution's configuration, you can choose to enable or disable automatic retraining, as well as adjust the training frequency as needed.

Automatic training mitigates model drift and makes sure recommendations align with users' evolving behaviors and preferences. Modifying the configuration of solution training allows you to adapt model retraining to your evolving business needs and data volumes. For example, you can increase training cadence during peak seasons to further optimize the relevance of recommendations. This saves time and resources by making incremental adjustments as needed, rather than re-creating the entire solution. The automatic training will continue at the new cadence until you make another update or disable automatic training entirely. Existing solutions will remain unchanged unless you modify the training configuration.

Amazon Personalize helps companies elevate the customer experience with AI-powered personalization and deliver hyper- personalized user experiences in real-time with precision and scale to improve user engagement, customer loyalty, and business results. Modifications for automatic solution training is supported in all service regions. To learn more, visit our documentation.

AWS WAF enhances rate-based rules to support lower rate limits

AWS WAF now supports setting lower rate limit thresholds for rate-based rules. Customers can now configure rate-based rules with rate limits as low as 10 requests per evaluation window, compared to the previous minimum of 100 requests.

With AWS WAF rate-based rules, customers can count incoming requests and limit traffic that exceeds a defined request rate. Now, in addition to existing threshold options, customers can set rate-based rule thresholds as low as 10 requests per the evaluation time window. This granular control allows customers to more effectively detect and respond to traffic spikes targeting sensitive applications and APIs, enabling quicker mitigation of sudden usage increases or malicious activity.

To use lower rate thresholds, simply set the 'Rate limit' to any value between 10 and 100 when configuring rate-based rules. Existing rules will remain unchanged. To customize, edit your rule to select a lower threshold then save. To learn more, see the AWS WAF developer guide. There is no additional cost for using this feature, however standard AWS WAF charges still apply. For details, visit the AWS WAF Pricing page.

This feature is available in all AWS Commercial Regions, except Asia Pacific (Hyderabad), Europe (Spain), Australia (Melbourne), Europe (Zurich), Israel (Tel Aviv), US-GovCloud and China Regions.

AWS Amplify introduces new function capabilities with scheduled cron jobs and streaming logs

AWS Amplify now offers two new features for its Functions capability: Scheduled Cron Jobs and Streaming Logs. Cron Jobs allow developers to configure serverless functions to run at specific intervals, while Streaming Logs enable developers to quickly iterate and test function execution by streaming logs directly to their terminal.

The scheduling feature allows developers to use natural language or cron expressions to configure their serverless functions to run automatically at specified intervals or times. This is particularly useful for tasks such as data processing, batch operations, or scheduled maintenance. Additionally, streaming logs provides real-time visibility into function execution logs, enabling developers to monitor and debug their functions more effectively.

To learn more about Scheduling and Streaming Logs, visit the Amplify documentation. Explore the comprehensive guides, code samples, and best practices to get started with these new features.

Amazon RDS for MySQL announces Extended Support minor 5.7.44-RDS.20240808

Amazon Relational Database Service (RDS) for MySQL announces Amazon RDS Extended Support minor version 5.7.44-RDS.20240808. We recommend that you upgrade to this version to fix known security vulnerabilities and bugs in prior versions of MySQL. Learn more about the bug fixes and patches in this version in the Amazon RDS User Guide.

Amazon RDS Extended Support provides you more time, up to three years, to upgrade to a new major version to help you meet your business requirements. During Extended Support, Amazon RDS will provide critical security and bug fixes for your RDS for MySQL databases after the community ends support for a major version. You can run your MySQL databases on Amazon RDS with Extended Support for up to three years beyond a major version's end of standard support date. Learn more about Extended Support in the Amazon RDS User Guide and the Pricing FAQs.

Amazon RDS for MySQL makes it simple to set up, operate, and scale MySQL deployments in the cloud. See Amazon RDS for MySQL Pricing for pricing details and regional availability. Create or update a fully managed Amazon RDS database in the Amazon RDS Management Console.

AWS Deadline Cloud now supports Windows Server 2022 in service-managed fleets

Today, AWS announces support for running Windows Server 2022 on workers in service-managed fleets in AWS Deadline Cloud. AWS Deadline Cloud is a fully managed service that simplifies render management for teams creating computer-generated 2D/3D graphics and visual effects for films, TV shows, commercials, games, and industrial design.

Now you can build pipelines for 3D graphics and visual effects using digital content creation (DCC) software that requires Windows - like Adobe After Effects, and KeyShot - without having to set up, configure, or manage the worker infrastructure yourself. Deadline Cloud service-managed fleets can be set up in minutes to run either Windows or Linux, expanding the DCC software you can use within a fully managed render farm.

Windows Server 2022 is supported in service-managed fleets in all AWS Regions where Deadline Cloud is available.

For more information, please visit the Deadline Cloud product page, and see the Deadline Cloud pricing page for price details.

Announcing Validation API for AWS Step Functions

AWS Step Functions announces a new Validation API for your AWS Step Functions workflows. The Validation API enables you to perform semantic checks on your workflows as you author them, helping you find syntactical errors sooner, shortening development cycles. AWS Step Functions is a visual workflow service capable of orchestrating virtually any AWS service to automate business processes and data processing workloads.

With the Validation API, you can perform semantic checks on your workflows before you run or deploy them to catch issues sooner. Simply validate a workflow by creating, updating, or directly calling the ValidateStateMachineDefinition API. Now with the Validation API you can catch common syntactical errors in workflows such as missing .$ on a field that uses a JSONPath or Intrinsic Function. We have also included suggestions for field names that are incorrect, but have a close match. For example, in the event of potential case sensitivity errors when calling AWS services such as Amazon Simple Notification Service, or Amazon Simple Queue Service, we will suggest the correct alphabetical case for a service name when a mismatch is detected, saving you time.

You can get started with the Validation API in the AWS console, using the AWS-SDK, or the AWS Command Line Interface (CLI). To learn more, see the AWS Step Functions Developer Guide to get started.

AWS AppConfig now provides deletion protection for additional guardrails

Customers can now enable deletion protection on AWS AppConfig resources, including Configuration Profiles and Environments. AWS AppConfig helps engineers move faster and resolve issues more quickly with managed feature flags and dynamic configuration. However, deleting any configuration data, for application hygiene or compliance reasons, should always be done very carefully to avoid unexpected behavior. With AWS AppConfig deletion protection enabled, a customer's account will not be allowed to delete a recently-used resource without explicitly bypassing deletion protection in the AWS Management Console, CLI, or API call. In addition, customers can set the amount of time that is considered "recently-used" to tailor to their organization's workflows.

AWS AppConfig already has many safety guardrails to be able to update feature flags and configuration data with confidence. With AWS AppConfig, customers can gradually deploy changes to measure and limit impact; customers can set up an alarm to automatically rollback an in-process deployment; customers can validate configuration data syntactically and semantically prior to pushing out updates. With deletion protection, customers now have an additional safety guardrail to ensure their use of feature flags and dynamic configuration is as expected.

Deletion protection for AWS AppConfig resources is available in all AWS Regions, including the AWS GovCloud (US) Regions. To get started, use the AWS AppConfig Getting Started Guide, or learn about AWS AppConfig deletion protection.

Amazon OpenSearch Service now supports Graviton3 (C7g, M7g, R7g, R7gd) instances

Amazon OpenSearch Service now supports AWS Graviton3 instances, which deliver up to 25% better performance over Graviton2-based instances. The new instance types are compute optimized (C7g), general purpose (M7g), and memory optimized (R7g, R7gd) instances. You can update your domain to the new instances seamlessly through the OpenSearch Service console or APIs.

AWS Graviton3 processors are custom-designed AWS Graviton processors that enable the best price performance for workloads in Amazon Elastic Compute Cloud (Amazon EC2). They offer up to 30 Gbps enhanced networking bandwidth and up to 20 Gbps of bandwidth to the Amazon Elastic Block Store (Amazon EBS). To learn more about Graviton3 improvements, please see the blog.

Amazon OpenSearch Service Graviton3 instances support all OpenSearch versions and Elasticsearch (open source) versions 7.9 and 7.10. One or more than one Graviton3 instance types are now available on Amazon OpenSearch Service across 21 regions globally: US East (N. Virginia), US East (Ohio), US West (N. California), US West (Oregon), Asia Pacific (Hong Kong), Asia Pacific (Hyderabad), Asia Pacific (Mumbai), Asia Pacific (Seoul), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Tokyo), Canada (Central), Europe (Frankfurt), Europe (Ireland), Europe (London), Europe (Milan), Europe (Paris), Europe (Spain), Europe (Stockholm), Middle East (Bahrain), South America and (São Paulo).

To learn more about region specific instance type availability and their pricing, visit our pricing page. To learn more about Amazon OpenSearch Service, please visit the product page.

Amazon Redshift Serverless is now available in the AWS Asia Pacific (Jakarta) region

Amazon Redshift Serverless, which allows you to run and scale analytics without having to provision and manage data warehouse clusters, is now generally available in the AWS Asia Pacific (Jakarta) region. With Amazon Redshift Serverless, all users, including data analysts, developers, and data scientists, can use Amazon Redshift to get insights from data in seconds. Amazon Redshift Serverless automatically provisions and intelligently scales data warehouse capacity to deliver high performance for all your analytics. You only pay for the compute used for the duration of the workloads on a per-second basis. You can benefit from this simplicity without making any changes to your existing analytics and business intelligence applications.

With a few clicks in the AWS Management Console, you can get started with querying data using the Query Editor V2 or your tool of choice with Amazon Redshift Serverless. There is no need to choose node types, node count, workload management, scaling, and other manual configurations. You can create databases, schemas, and tables, and load your own data from Amazon S3, access data using Amazon Redshift data shares, or restore an existing Amazon Redshift provisioned cluster snapshot. With Amazon Redshift Serverless, you can directly query data in open formats, such as Apache Parquet, in Amazon S3 data lakes. Amazon Redshift Serverless provides unified billing for queries on any of these data sources, helping you efficiently monitor and manage costs.

To get started, see the Amazon Redshift Serverless feature page, user documentation, and API Reference.

Amazon SageMaker Projects now allows you to reuse names of previously deleted Projects

Amazon SageMaker Projects now allows you to reuse deleted project names. This launch enhances the project deletion process by removing the project names and metadata instead of only marking them as deleted. This capability is available via SageMaker Projects API, SageMaker Studio, and Amazon SageMaker Python SDK.

Customers use SageMaker Projects to preconfigure and bootstrap all the resources required to run their ML workflow (e.g. code repositories, model training instances, ML Pipelines). This enhances the productivity of data scientists and accelerates the journey of generative AI and ML workflows from prototyping to production. With this launch, whenever customers delete a SageMaker Project all the metadata associated with it will be automatically deleted. They can reuse the name for new SageMaker Projects created in the future.

This capability is available in all regions where Amazon SageMaker Projects is available. To learn more, see the SageMaker Projects Developer Guide

AWS Global Accelerator launches new edge location in Cairo, Egypt

AWS Global Accelerator now supports traffic through a new AWS edge location in Cairo, Egypt. With the addition of the edge location, Global Accelerator is now available through 121 Points of Presence globally and supports application endpoints in 29 AWS Regions.

AWS Global Accelerator is a service that is designed to improve the availability, security, and performance of your internet-facing applications. By using the congestion-free AWS network, end-user traffic to your applications benefits from increased availability, DDoS protection at the edge, and higher performance relative to the public internet. Global Accelerator provides static IP addresses that act as fixed entry endpoints for your application resources in one or more AWS Regions, such as your Application Load Balancers, Network Load Balancers, Amazon EC2 instances, or Elastic IPs. Global Accelerator continually monitors the health of your application endpoints and offers deterministic fail-over for multi-region workloads without any DNS dependencies.

To get started, visit the AWS Global Accelerator website and review its documentation.

Research and Engineering Studio on AWS Version 2024.08 now available

Today we're excited to announce Research and Engineering Studio (RES) on AWS Version 2024.08. This release adds new features such as Amazon S3 bucket mountpoints for Linux, allows creation of custom user roles and permission profiles, and offers the ability to adjust the list of Amazon EC2 instances available to launch as virtual desktops.

Amazon S3 mountpoints allow Linux desktops to access S3 buckets as a file system. Amazon S3 buckets are created using the AWS Console or AWS CLI and onboarded to RES by administrators using the S3 Buckets page in the web portal. You can mount buckets in either Read Only or Read/Write mode. Read/Write buckets have an optional setting to restrict data access by project, or both project *and* user. RES can also mount S3 buckets from other AWS accounts with the proper permissions.

Custom permission profiles allow administrators to create unique permissions and assign them to users or groups. Start by modifying the default Project Member and Project Owner roles or create your own permission set to adjust permissions for project and virtual desktop infrastructure management. The list of allowed instance types is also now configurable from the RES UI. You can adjust this list to define the instances available for virtual desktops in your environment.

RES 2024.08 extends regional availability to Europe (Stockholm). See the regional availability page for a full list of regions where RES is available.

Check out additional release notes on Github to get started and deploy RES 2024.08 today.

Amazon Bedrock Knowledge Bases now supports Llama 3.1 405B, 70B, and 8B

Amazon Bedrock Knowledge Bases securely connects foundation models (FMs) to internal company data sources for Retrieval-Augmented Generation (RAG) to deliver relevant, context-specific, and accurate responses. Meta's Llama 3.1 family of foundation models (405B, 70B, and 8B) is now generally available on Knowledge Bases.

Llama 3.1 is the next generation of state-of-the-art models from Meta, supporting 128,000 tokens (roughly 96,000 words, or 192 pages of material) context length. Llama 3.1 models are well-suited for tasks that require complex reasoning, quick outputs, and RAG. Llama 3.1 is supported through the fully managed RetrieveAndGenerate API.

Amazon Bedrock Knowledge Bases on Llama 3.1 405B, 70B, and 8B models is now generally available in the US West (Oregon) AWS Region. To learn more, read the AWS news blog launch and visit the Meta Llama in Amazon Bedrock page. To get started, refer to the Knowledge Bases for Amazon Bedrock documentation and visit the Amazon Bedrock console.

AWS Network Firewall introduces GeoIP Filtering to inspect traffic based on geographic location

AWS Network Firewall now supports GeoIP Filtering on ingress and egress Amazon Virtual Private Cloud (VPC) traffic. This new feature makes it easy for customers to block traffic coming from or going to specific countries and meet compliance requirements. Previously, maintaining compliance with regulations was time-consuming because you have to maintain a list of IP addresses associated with specific countries and update your firewall rules regularly. GeoIP Filtering saves time and reduces operational complexity by enabling you to filter traffic on Network Firewall using the country name.

AWS Network Firewall is a managed firewall service that makes it easy to deploy essential network protections for all your Amazon VPCs. GeoIP Filtering allows you to enforce your AWS Network Firewall rules and policies consistently across your entire network, making it easier to meet business or regulatory compliance requirements and improve your network security posture.

GeoIP Filtering is supported in all AWS Regions where AWS Network Firewall is available today, including the AWS GovCloud (US) Regions. For more information about the AWS Regions where AWS Network Firewall is available, see the AWS Region table.

There is no additional cost to enable GeoIP Filtering on AWS Network Firewall. You can configure GeoIP Filtering using the AWS Management Console, AWS CLI, AWS SDK, or the AWS Network Firewall API. To learn more about configuring GeoIP Filtering, please refer to the service documentation.

Announcing AWS Parallel Computing Service

Today, AWS announces AWS Parallel Computing Service (AWS PCS), a new managed service that lets you run and scale high performance computing (HPC) workloads on AWS. The service enables you to build scientific and engineering models, and run simulations using your preferred HPC job scheduler (starting with Slurm). AWS PCS allows you to build complete HPC clusters that integrates compute, storage, networking, and visualization resources, and seamlessly scale from zero to thousands of instances. The service offers a fully managed Slurm scheduler with built-in technical support and a rich set of customization options, helping you tailor your HPC environment to your specific needs and integrate it with your preferred software stack.

With AWS PCS, you can take advantage of the scalability and flexibility of AWS for your HPC workloads, while maintaining compatibility with existing applications and job scripts. It simplifies cluster management by offering a unified set of APIs, AWS Management Console, SDK, and CLI tools for cluster provisioning and infrastructure updates. This helps you to reduce operational overhead, so you can focus more on your core engineering and scientific research, and less on compute infrastructure management.

AWS PCS is available in the following AWS Regions: US East (N. Virginia), US East (Ohio), US West (Oregon), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Tokyo), Europe (Frankfurt), Europe (Ireland), and Europe (Stockholm).

To learn more, visit AWS Parallel Computing Service and read the blog post.

Amazon Location Service announces Migration SDK

Amazon Location Service has launched a Migration SDK that enables users to easily and quickly migrate their existing application from Google Maps Platform to Amazon Location Service.

The Migration SDK provides an option for your application built using the Google Maps SDK for JavaScript to use Amazon Location Service without needing to rewrite any of the application or business logic if Amazon Location supports the capabilities used. Customers can compare their current Google Maps API usage with the Migration SDK's list of supported APIs to determine if the Migration SDK is right for them. The Migration SDK will receive updates as Amazon Location Service extends its Maps/Places/Routes feature set.

To learn more, visit the Migration SDK instructions and check for currently supported Google APIs here.

Amazon OpenSearch Serverless now available in the AWS GovCloud (US-West) Region

We are excited to announce that Amazon OpenSearch Serverless is expanding availability to the Amazon OpenSearch Serverless, now available in the AWS GovCloud (US-West) Region. OpenSearch Serverless is a serverless deployment option for Amazon OpenSearch Service that makes it simple to run search and analytics workloads without the complexities of infrastructure management. OpenSearch Serverless' compute capacity used for data ingestion, search, and query is measured in OpenSearch Compute Units (OCUs).

The support for OpenSearch Serverless is now available in 14 regions globally: US East (Ohio), US East (N. Virginia), US West (Oregon), Asia Pacific (Singapore), Asia Pacific (Sydney), Asia Pacific (Tokyo), Europe (Frankfurt), Europe (Ireland), Europe West (Paris), Europe West (London), Asia Pacific South (Mumbai), South America (Sao Paulo), Canada Central (Montreal) and the AWS GovCloud (US-West) Region. Please refer to the AWS Regional Services List for more information about Amazon OpenSearch Service availability. To learn more about OpenSearch Serverless, see the documentation.

AWS announces Amazon-provided contiguous IPv4 blocks

Starting today, customers can provision Amazon-provided contiguous IPv4 blocks using Amazon VPC IP Address Manager (IPAM), to simplify network management and security.

Customers want to use Amazon-provided contiguous IPv4 blocks in their networking and security constructs like access control lists, route tables, security groups, and firewalls as opposed to using many individual discontiguous public IPv4 addresses that can be cumbersome to manage. Within IPAM, customers can provision Amazon-provided contiguous IPv4 blocks into IPv4 publicly scoped regional pools. They can then create Elastic IP addresses from these pools, and use them with AWS resources, such as EC2 instances, Network Load Balancers and NAT gateways.

Customers are charged for all public IPv4 addresses in the Amazon-provided contiguous IPv4 block. To learn about pricing for this feature, please see the Amazon-provided contiguous IPv4 block tab in the VPC pricing page.

Amazon-provided contiguous IPv4 blocks are available in all AWS commercial regions and the AWS GovCloud (US) Regions, in both Free Tier and Advanced Tier of VPC IPAM. When used with the Advanced Tier of VPC IPAM, customers can share their Amazon-provided contiguous IPv4 blocks across two or more accounts. To get started please see the IPAM documentation page.

Amazon EC2 status checks now support reachability health of attached EBS volumes

Starting today, you can leverage Amazon EC2 status checks to directly monitor if the EBS volumes attached to your instances are reachable and able to complete I/O operations. You can use this new status check to quickly detect attachment issues or volume impairments that may impact the performance of your applications running on Amazon EC2 instances. You can further integrate these status checks within Auto Scaling groups to monitor the health of EC2 instances and replace impacted instances to ensure high availability and reliability of your applications. Attached EBS status checks can be used along with the instance status and system status checks to monitor the health of your instances.

Prior to today, you could only monitor the health of your EBS volume attachments by configuring and enabling a specific CloudWatch metric. Now with this capability, you can monitor the health of EBS volume attachments to your instance directly from EC2 Console or describe-instance-status API without any additional configuration or action from your side.

Polly Voices for two new locales: Czechia and Switzerland

Today, we are excited to announce the general availability of two female-sounding Amazon Polly voices for two new locales: Czechia and Switzerland.

Amazon Polly is a managed service that turns text into lifelike speech, allowing you to create applications that talk and to build speech-enabled products depending on your business needs.

We have developed a Swiss Standard German voice Sabrina and a Czech voice Jitka using a well-established neural TTS technology allowing our customers to synthesize natural human-like speech. With these voices, we bring Polly to two new countries and expand our language offerings. Sabrina and Jitka are conversational voices that can be used in Interactive Voice Response technology and more.

Sabrina and Jitka voices are accessible in all Polly regions and complement the other types of voices that are already available for developing speech products for a variety of use cases.

For more details, please read the Amazon Polly documentation and visit our pricing page.

Amazon EC2 C6gd and R6gd instances are now available in AWS Europe (Spain) region

Starting today, Amazon EC2 C6gd and R6gd instances are available in the AWS Europe (Spain) region. These instances are powered by AWS Graviton2 processors and are built on the AWS Nitro System. The Nitro System is a collection of AWS designed hardware and software innovations that enables the delivery of efficient, flexible, and secure cloud services with isolated multi-tenancy, private networking, and fast local storage.

C6gd instances are ideal for compute-intensive workloads such as high performance computing (HPC), batch processing, and CPU-based machine learning inference. R6gd instances are built for running memory-intensive workloads such as open-source databases, in-memory caches, and real time big data analytics. The local SSD storage provided on these instances will benefit applications that need access to high-speed, low latency storage, as well as for temporary storage of data such as batch and log processing, and for high-speed caches and scratch files. These instances offer up to 25 Gbps of network bandwidth, and up to 19 Gbps of bandwidth to the Amazon Elastic Block Store (Amazon EBS). C6gd and R6gd instances offer up to 3.8 TB of NVMe-based SSD storage.

To learn more about the instances, see Amazon EC2 C6gd and R6gd. To get started with AWS Graviton2-based instances, visit the AWS Management Console, AWS Command Line Interface (CLI), and AWS SDKs.

Amazon RDS for PostgreSQL announces Extended Support minor 11.22-RDS.20240808

Amazon Relational Database Service (RDS) for PostgreSQL announces Amazon RDS Extended Support minor version 11.22-RDS.20240808. We recommend that you upgrade to this version to fix known security vulnerabilities and bugs in prior versions of PostgreSQL. Learn more about the updates and patches in this Extended Support minor version in the Amazon RDS User Guide.

Amazon RDS Extended Support provides you more time, up to three years, to upgrade to a new major version to help you meet your business requirements. During Extended Support, Amazon RDS will provide critical security and bug fixes for your RDS for PostgreSQL databases after the community ends support for a major version. You can run your PostgreSQL databases on Amazon RDS with Extended Support for up to three years beyond a major version's end of standard support date. Learn more about Extended Support in the Amazon RDS User Guide.

You are able to leverage automatic minor version upgrades to automatically upgrade your databases to more recent minor versions during scheduled maintenance windows. Learn more about upgrading your database instances, including minor and major version upgrades, in the Amazon RDS User Guide.

Amazon RDS for PostgreSQL makes it simple to set up, operate, and scale PostgreSQL deployments in the cloud. See Amazon RDS for PostgreSQL Pricing for pricing details and regional availability. Create or update a fully managed Amazon RDS database in the Amazon RDS Management Console.

Amazon Bedrock now supports cross-region inference

Today, Amazon Bedrock announces support for cross-region inference, an optional feature that enables developers to seamlessly manage traffic bursts by utilizing compute across different AWS Regions. By using cross-region inference, Bedrock customers using on-demand mode will be able to get higher throughput limits (up to 2x their allocated in-region quotas) and enhanced resilience during periods of peak demand. By opting in, developers no longer have to spend time and effort predicting demand fluctuations. Instead, cross-region inference dynamically routes traffic across multiple regions, ensuring optimal availability for each request and smoother performance during high-usage periods.

Customers can control where their inference data flows by selecting from a pre-defined set of regions, helping them comply with applicable data residency requirements and sovereignty laws. Moreover, this capability prioritizes the connected Bedrock API source region when possible, helping to minimize latency and improve responsiveness. As a result, customers can enhance their applications' reliability, performance, and efficiency.

There's no additional routing cost for using cross-region inference and you will be charged based on the region you made the request in (source region). Please find the list of supported models and pre-defined regions here. To learn more about the feature and how to get started, refer to the Amazon Bedrock documentation or this blog.

AWS announces support for Microsoft Entra ID and Microsoft Intune on Amazon WorkSpaces Personal

Amazon WorkSpaces Personal now supports Microsoft Entra ID and Intune. With this launch, customers using Amazon WorkSpaces Personal can now provision virtual desktops joined with Entra ID and enrolled in Intune, without requiring Microsoft Active Directory. By integrating with AWS IAM Identity Center, the launch also allows customers the flexibility to use other cloud-based identity and endpoint management solutions with WorkSpaces including JumpCloud.

With the launch, WorkSpaces Personal now supports both AD and non-AD domain joined virtual desktops. For customers who want to use Entra ID for identity management, AWS IAM Identity Center is used to ensure user identity data is automatically synchronized from Entra ID to AWS. Leveraging Windows Autopilot user-driven mode, Windows 10 and 11 virtual desktops are automatically enrolled to Intune during provisioning and joined to Entra ID during Windows Out of Box Experience (OOBE). End users log into their virtual desktops as Entra ID users, so they can access Microsoft 365 Apps for enterprise without another Entra ID login. In addition, with non-AD domain joined WorkSpaces, customers now have the option to use JumpCloud which is a native cloud directory platform which provides identity, access, and device management.

The feature is generally available today in all regions where Amazon WorkSpaces Personal is offered, except for Africa (Cape Town), Israel (Tel Aviv), and China regions. There is no extra cost for using the feature and IAM Identity Center.

To learn more about the feature, see Amazon WorkSpaces Administration Guide. To get started with the feature, log on to AWS Management Console.

Amazon Q Business launches IAM federation for user identity authentication

Amazon Q Business is a fully managed, generative AI–powered assistant that enhances workforce productivity by answering questions, providing summaries, generating content, and completing tasks based on customers' enterprise data. Customers create and manage their workforce user identity using identity providers of their choice. Previously, customers had to sync their user identity information from their identity provider into AWS IAM Identity Center, and then connect their Amazon Q Business applications to IAM Identity Center for user authentication.

Starting today, customers can use the Amazon Q Business IAM federation feature to connect their applications directly to their identity provider to source user identity and user attributes for these applications.

At launch, Amazon Q Business IAM federation will support the OpenID Connect (OIDC) and SAML2.0 protocols for identity provider connectivity. Amazon Q Business applications built using IAM federation will support advanced features including custom plugins, Amazon Q Apps, and personalization. Amazon Q Business IAM federation is available in all AWS Regions where Amazon Q Business is available. To learn more, visit the documentation. To explore Amazon Q Business, visit the website.

Amazon Braket adds support for Rigetti's 84-Qubit Ankaa™-2 system, our largest gate-based superconducting device

Amazon Braket, the quantum computing service from AWS, now offers Rigetti Computing's latest 84-qubit Ankaa-2 system in the US West (N. California) Region. Ankaa-2 is the highest qubit-count gate-based quantum device available on Amazon Braket, enabling customers to tackle larger and more complex problems while pushing the boundaries of what's possible with today's quantum hardware. Ankaa-2 allows customers to submit their quantum tasks and hybrid jobs at any time, opening up new possibilities for uninterrupted experiments and a more efficient use of quantum hardware resources, all on a pay-as-you-go basis.

Braket's core mission is to accelerate research and innovation in quantum computing by making quantum hardware available to customers on-demand, with pay-as-you-go pricing, via a unified interface and access model. With this launch, customers can now use the familiar Braket SDK and APIs to access the latest 84-qubit system from Rigetti, which is designed to offer significantly faster gate operation times, improved fidelities, and four-fold qubit connectivity enabling customers to run deeper circuits. Researchers requiring access to lower levels of control over the hardware to explore use cases such as studying noise, developing new and more robust gates, and devising error mitigation schemes can access Ankaa-2 using Braket Pulse.

To get started with Rigetti Ankaa-2, refer to the Amazon Braket Rigetti device page for device details, the Amazon Braket Documentation for guides and resources, and the Amazon Braket Pricing page for pricing information.

Amazon QuickSight now supports sharing views of embedded dashboards

Amazon QuickSight now supports sharing views of embedded dashboards. This feature allows developers to enable more collaborative capabilities in their application with embedded QuickSight dashboards. Additionally, they can enable personalization capabilities such as bookmarks for anonymous users.

This feature enables embedded dashboards readers to collaborate easily. They can share a unique link that displays only their changes while staying within the application. When an embedded dashboard reader wants to share their view with another user of the application, the application developer can pass this intent to QuickSight SDK. This will generate a shareable reference to the current state of the embedded dashboard. When a different reader lands on that shareable link, developers can load the QuickSight dashboard with that particular shared state. For more information, click here.

This feature is now available in all supported Amazon QuickSight regions - US East (Ohio and N. Virginia), US West (Oregon), Asia Pacific (Mumbai, Seoul, Singapore, Sydney and Tokyo), Canada (Central), Europe (Frankfurt, Ireland and London), South America (São Paulo) and the AWS GovCloud (US-West) Region.

AWS announces a streamlined Federated and SSO sign in process for the AWS Console Mobile App

Amazon Web Services (AWS) is announcing a streamlined Federated and SSO sign in process for the AWS Console Mobile App. AWS customers who use Federated or SSO authentication methods with the AWS Console Mobile App can now select their sign in URL from a list of recently used URLs when they setup a subsequent identity to access an account.

The Console Mobile App lets users view and manage a select set of resources to stay informed and connected with their AWS resources while on-the-go. The sign in process supports device password managers and biometrics authentication, making access to AWS resources simple, secure, and quick.

Visit the product page for more information about the Console Mobile App.

Knowledge Bases for Amazon Bedrock supports Anthropic's Claude 3.5 Sonnet

Knowledge Bases for Amazon Bedrock securely connects foundation models (FMs) to internal company data sources for Retrieval Augmented Generation (RAG) to deliver relevant, context-specific, and accurate responses. Anthropic's Claude 3.5 Sonnet foundation model is now generally available on Knowledge Bases.

Anthropic's Claude 3.5 Sonnet—the first model in the forthcoming Claude 3.5 model family—has a 200,000 context window (roughly 150,000 words, or around 300 pages of material) and raises the industry bar for intelligence, outperforming other generative AI models on a wide range of evaluations. Claude 3.5 Sonnet is well-suited for tasks that require complex reasoning, quick outputs, and RAG. Additionally, Claude 3.5 Sonnet is supported through the fully managed RetrieveAndGenerate API.

Anthropic's Claude 3.5 Sonnet model support for Amazon Bedrock Knowledge Bases is now generally available in the US East (N. Virginia), US West (Oregon), Europe (Frankfurt) and Asia Pacific (Tokyo) AWS Regions. To learn more, read the AWS News launch blog and Claude in Amazon Bedrock product page. To get started, refer to the Knowledge Bases for Amazon Bedrock documentation and visit the Amazon Bedrock console.

Amazon Connect provides new ways to configure callbacks

Amazon Connect now allows you to configure flows to take actions on callbacks prior their creation and while they are in queue. For example, you can now automate sending a notification to a customer via SMS before calling them back, update callback attributes based on latest customer data for agents to reference, or even terminate the callback if the issue has already been resolved. You can also now run flows to dynamically re-prioritize and transfer callbacks to another queue based on customer information from Customer profiles or third-party applications, or if it's just taking too long for the callback queue to drain.

This feature is available in all AWS regions where Amazon Connect is offered. To learn more about this feature, see the Amazon Connect administrator guide. To learn more about Amazon Connect, the AWS cloud-based contact center, please visit the Amazon Connect website.

CloudFormation simplifies resource discovery and template review in the IaC Generator

Today, AWS CloudFormation announces two new enhancements to the IaC generator, which customers use to create infrastructure-as-code (IaC) from existing resources. Now, after the IaC generator finishes scanning the resources in an account, it presents a graphical summary of the different resource types to help customers more quickly find the resources they want to include in their template. After selecting resources, customers can also preview their template in AWS Application Composer, which visualizes the full application architecture with the resources and their relationships.

Customers use the best practice of IaC, where they specify and version infrastructure configuration using code, to easily replicate their environments, reliably deploy changes to them, and apply controls to enforce their security and governance policies. The IaC generator helps customers adopt IaC on the resources they created through other means, such as the AWS Management Console or CLI. After generating a template for selected resources, customers can import the resources into CloudFormation, download the template for deployment to other regions and accounts, or generate a CDK CLI command that converts the template into a CDK app in their preferred programming language, such as TypeScript or Python.

The IaC generator is available in AWS Regions where CloudFormation is available (please refer to the AWS Region table). To get started, open the CloudFormation console and select the IaC generator in the navigation panel. To learn more:

- CloudFormation documentation for Scan Summary
- CloudFormation documentation for Application Composer

Amazon Q now provides more details about user subscriptions and associated resources

The Amazon Q Console now provides administrators with greater visibility into how users are utilizing Amazon Q Developer Pro, Amazon Q Business Pro, and Amazon Q Business Lite subscriptions. This new feature enables administrators to view a list of subscribed users, their subscription status (e.g., active, pending, under free trial, canceled), and their corresponding associations. Associations refer to the applications, accounts, or services that a user has access to through their subscription. Organization administrators will have a view of all subscription associations across applications in various accounts, while member account administrators' visibility will be limited to only the applications within accounts they administer.

With this update, administrators can monitor subscribed users of the Amazon Q offerings. Key capabilities include listing all subscribed users, their status, the specific subscription types, the applications they have associations to (with AWS account numbers), filtering and searching users and associations, and generating a downloadable report.

The visibility into user subscriptions and associations is available across all AWS regions where the specific Amazon Q subscriptions are offered.

To learn more about Amazon Q's subscription management features, visit the Amazon Q Console or Amazon Q documentation.

Amazon DocumentDB (with MongoDB Compatibility) Global Clusters introduces Failover

Amazon DocumentDB now supports Global Cluster Failover, a fully managed experience for performing a cross-region failover to respond to unplanned events such as a regional outage. With Global Cluster Failover, you can convert a secondary region into the new primary region in typically a minute and also maintain the multi-region Global Cluster configuration. An Amazon DocumentDB Global Cluster is a single cluster that can span up to 6 AWS Regions, enabling disaster recovery from region-wide outages and low latency global reads.

Combined with Global Cluster Switchover, you can easily promote a secondary region to a primary region for both unplanned and planned events. Global Cluster Switchover is a fully managed cross-region database failover experience meant for planned events such as regional rotations. See our documentation to learn more about Global Cluster Failover and Switchover.

Amazon DocumentDB makes it easy and cost effective to operate critical document workloads in a highly-available manner at virtually any scale without managing infrastructure. To get started with Amazon DocumentDB, take a look at our getting started page.

AWS Amplify introduces multiple bucket support for Storage

AWS Amplify is launching multiple bucket support for Storage (JavaScript Only). You can now configure more than one storage bucket in your Amplify backend configuration. Amplify storage integrates with Amazon Simple Storage Service (Amazon S3) and provides an intuitive approach to managing cloud-based file storage. With this new feature you can upload and download files from and to multiple storage buckets, providing greater flexibility and control over your storage resources.

With multiple bucket support, you can better manage your storage resources by configuring additional storage buckets and applying existing access permissions to different paths within each bucket. Amplify Storage APIs for JavaScript provide APIs like upload, download, and more and now accept a bucket alias or bucket name and region. Currently, the UI components (Storage Image and Storage Manager) and Amplify console display content only from the default storage bucket.

To learn how to configure additional storage buckets for different use cases, follow our blog post. If you're already using Amplify Storage, check out our documentation to add additional buckets to your existing storage configuration. For those new to Amplify Storage, get started by following our guide to set up storage for your Amplify project.

Amazon Connect Contact Lens now provides an audit trail for changes to an agent performance evaluation

Amazon Connect Contact Lens now provides an audit trail to review the changes made to an agent performance evaluation when it is re-submitted. This launch displays the audit trail, which was previously available within a customer's S3 bucket, directly in the Amazon Connect UI. When an evaluator submits changes to an existing evaluation form, managers can now view an audit trail of who submitted the original evaluation, who re-submitted the evaluation, and what changes they made. Contact center managers can use this information to perform internal audits and improve consistency across evaluators.

This feature is available in all regions where Contact Lens performance evaluations is already available. To learn more, please visit our documentation and our webpage. For information about Contact Lens pricing, please visit our pricing page.

Amazon Data Firehose is now available in the AWS Asia Pacific (Malaysia) region

Starting today, you can use Amazon Data Firehose in the AWS Asia Pacific (Malaysia) region. Amazon Data Firehose is the easiest way to load streaming data into data stores and analytics tools. You can capture, transform, and deliver streaming data into Amazon S3, Amazon OpenSearch Service, Amazon Redshift, Snowflake, Apache Iceberg, and third party analytics applications such as Splunk and Datadog, enabling real-time analytics use cases.

With Amazon Data Firehose, you don't need to write applications or manage resources. You configure your data producers to send data to Amazon Data Firehose, and it automatically delivers the data to the destination that you specified. You can also configure Amazon Data Firehose to transform your data before delivering it. To get started, you need an AWS account. Once you have an account, you can create a delivery stream in the Amazon Data Firehose Console. To learn more, explore the Amazon Data Firehose Developer Guide.

For Amazon Data Firehose availability, refer to the AWS Region Table.