AWS Official Blog

Simplify custom contact center insights with Amazon Connect analytics data lake
Amazon Connect's analytics data lake centralizes customer data for powerful performance analysis without complex pipelines.

AWS analytics services streamline user access to data, permissions setting, and auditing
Streamline data access with trusted identity propagation: simplify sign-in, define fine-grained permissions, and audit using actual user identities across selected AWS analytics services.

Amazon EC2 high memory U7i Instances for large in-memory databases
Unleash unparalleled power with U7i instances: up to 32 TiB DDR5 memory, 896 vCPUs, and unmatched performance for in-memory databases like SAP HANA, Oracle, and SQL Server.

AWS Weekly Roundup: LlamaIndex support for Amazon Neptune, force AWS CloudFormation stack deletion, and more (May 27, 2024)
Last week, Dr. Matt Wood, VP for AI Products at Amazon Web Services (AWS), delivered the keynote at the AWS Summit Los Angeles. Matt and guest speakers shared the latest advancements in generative artificial intelligence (generative AI), developer tooling, and foundational infrastructure, showcasing how they come together to change what's possible for builders. You can […]

Optimized for low-latency workloads, Mistral Small now available in Amazon Bedrock
Mistral Small, an efficient large language model, joins Amazon Bedrock's lineup, excelling in low-latency tasks, multilingual support, and coding prowess while maintaining cost-effectiveness.

AWS Weekly Roundup: Application Load Balancer IPv6, Amazon S3 pricing update, Amazon EC2 Flex instances, and more (May 20, 2024)
AWS Summit season is in full swing around the world, with last week's events in Bengaluru, Berlin, and Seoul, where my blog colleague Channy delivered one of the keynotes. Last week's launches Here are some launches that got my attention: Amazon S3 will no longer charge for several HTTP error codes – A customer reported […]

New compute-optimized (C7i-flex) Amazon EC2 Flex instances
The vast majority of applications don't run run the CPU flat-out at 100% utilization continuously. Take a web application, for instance. It typically fluctuates between periods of high and low demand, but hardly ever uses a server's compute at full capacity. CPU utilization for many common workloads that customers run in the AWS Cloud today. […]

AWS Weekly Roundup: New capabilities in Amazon Bedrock, AWS Amplify Gen 2, Amazon RDS and more (May 13, 2024)
AWS Summit is in full swing around the world, with the most recent one being AWS Summit Singapore! Here is a sneak peek of the AWS staff and ASEAN community members at the Developer Lounge booth. It featured AWS Community speakers giving lightning talks on serverless, Amazon Elastic Kubernetes Service (Amazon EKS), security, generative AI, […]

A new generative engine and three voices are now generally available on Amazon Polly
With Amazon Polly, you can select various voice options, including neural, long-form, and generative voices, which deliver ground-breaking improvements in speech quality and produce human-like, highly expressive, and emotionally adept voices.

Build RAG and agent-based generative AI applications with new Amazon Titan Text Premier model, available in Amazon Bedrock
Titan Text Premier is the latest large language model (LLM) in the Amazon Titan family of models, further increasing your model choice within Amazon Bedrock.

Build generative AI applications with Amazon Bedrock Studio (preview)
Amazon Bedrock Studio accelerates the development of generative AI applications by providing a rapid prototyping environment with key Amazon Bedrock features, including Knowledge Bases, Agents, and Guardrails.

AWS Weekly Roundup: Amazon Q, Amazon QuickSight, AWS CodeArtifact, Amazon Bedrock, and more (May 6, 2024)
April has been packed with new releases! Last week continued that trend with many new releases supporting a variety of domains such as security, analytics, devops, and many more, as well as more exciting new capabilities within generative AI. If you missed the AWS Summit London 2024, you can now watch the sessions on demand, […]

Build RAG applications with MongoDB Atlas, now available in Knowledge Bases for Amazon Bedrock
Foundational models (FMs) are trained on large volumes of data and use billions of parameters. However, in order to answer customers' questions related to domain-specific private data, they need to reference an authoritative knowledge base outside of the model's training data sources. This is commonly achieved using a technique known as Retrieval Augmented Generation (RAG). […]

Stop the CNAME chain struggle: Simplified management with Route 53 Resolver DNS Firewall
Updated 2 May 2024: I removed the reference to Route53 Alias that was incorrectly referred as a chain Starting today, you can configure your DNS Firewall to automatically trust all domains in a resolution chain (such as aCNAMEor DNAMEchain). Let's walk through this in nontechnical terms for those unfamiliar with DNS. Why use DNS Firewall? […]

Add your Ruby gems to AWS CodeArtifact
Updated 8 May 2024: I updated the command to collect RUBYGEMS_HOST to suppress the trailing / Ruby developers can now use AWS CodeArtifact to securely store and retrieve their gems. CodeArtifact integrates with standard developer tools like gem and bundler. Applications often use numerous packages to speed up development by providing reusable code for common […]

Amazon Titan Text Embeddings V2 now available in Amazon Bedrock, optimized for improving RAG
Amazon Titan foundation models (FMs) offer a comprehensive suite of pre-trained image, multimodal, and text models accessible through a fully managed API. They're designed for a range of applications while adhering to responsible AI practices.

Amazon Q Business, now generally available, helps boost workforce productivity with generative AI
Amazon Q Business is now available with many new features, including custom plugins, and a preview of Amazon Q Apps, generative AI–powered customized and sharable applications using natural language in a single step for your organization.

Amazon Q Developer, now generally available, includes previews of new capabilities to reimagine developer experience
This generative AI–powered assistant helps developers build applications on AWS, research best practices, perform troubleshooting, and resolve errors.

Run scalable, enterprise-grade generative AI workloads with Cohere Command R & R+, now available in Amazon Bedrock
In November 2023, we made two new Cohere models available in Amazon Bedrock (Cohere Command Light and Cohere Embed English). Today, we're announcing the addition of two more Cohere models in Amazon Bedrock; Cohere Command R and Command R+. Organizations need generative artificial intelligence (generative AI) models to securely interact with information stored in their […]

AWS Weekly Roundup: Amazon Bedrock, AWS CodeBuild, Amazon CodeCatalyst, and more (April 29, 2024)
This was a busy week for Amazon Bedrock with many new features! Using GitHub Actions with AWS CodeBuild is much easier. Also, Amazon Q in Amazon CodeCatalyst can now manage more complex issues. I was amazed to meet so many new and old friends at the AWS Summit London. To give you a quick glimpse, […]