Jillian Forde (00:00):

Welcome to the AWS Podcast. I'm your host, Jillian Forde. Today's episode is going to cover all of the latest announcements from day two of re:Invent, including those mentioned in Adam Selipsky's keynote. Let's dive right in.

(00:15):

We've got a new storage class for S3, the Amazon S3 Express One Zone storage class is purpose-built to deliver the fastest cloud object storage for performance-critical applications that demand consistent, single-digit millisecond request latency. S3 Express One Zone can improve data access speeds by 10x, and reduce request costs by 50% compared to S3 Standard, and skills to process millions of requests per minute for your most frequently accessed datasets. It enables workloads such as machine learning training, interactive analytics, and a media content creation to achieve single-digit millisecond data access speeds with high durability and availability. S3 Express One Zone uses a new bucket type, S3 directory buckets, to support hundreds of thousands of requests per second, and uses purpose-built hardware and software optimized for low latency.

(01:21):

While you've always been able to choose a specific AWS region to store your S3 data, with S3 Express One Zone, you can now select a specific AWS availability zone within an AWS region to store your data. Super exciting. S3 Express One Zone is generally available in the US East Northern Virginia, US West Oregon, Europe, Stockholm, and Asia-Pacific Tokyo AWS regions. You can now use Mountpoint for Amazon S3 to access objects stored in the new Amazon S3 Express One Zone storage class using file system operations. Using Amazon S3 Express One Zone with Mountpoint for Amazon S3, this accelerates file-based applications that make random data access requests by up to 6x compared to S3 Standard. This speeds up machine learning training jobs for vision models, completing them faster and reducing compute costs.

(02:20):

Additionally, with S3 Express One Zone's integrations with Mountpoint for Amazon S3 and big data platforms, like Amazon EMR, you can prepare your training data using Amazon EMR and run training jobs using Mountpoint all on the same data set, which simplifies your architecture. Mountpoint for Amazon S3 supports sequential and random read operations, and sequential right operations for creating new files. For details on supported file system operations with this release, and to get started, read the documentation.

(02:56):

Starting today, you can now use Amazon Athena to query data stored in the Amazon S3 Express One Zone storage class for up to 2.1x faster query performance than S3 Standard. If you have latency sensitive use cases for your data lakes, such as business intelligence analytics and reporting, financial risk monitoring, or sensor data processing, use Athena with S3 Express One Zone to deliver fast query results. You can now accelerate data processing and analysis with Apache Spark applications by up to 4x than data in S3 Standard using Amazon EMR and, you guessed it, the Amazon S3 Express One Zone storage class.

(03:41):

Amazon Aurora Postgres zero-ETL integration with Amazon Redshift is in public preview. Within seconds of transactional data being written into Aurora, the data is available in Amazon Redshift. You don't have to build and maintain complex data pipelines to perform extract, transform, and load operations. But wait, there's more in zero-ETL. Now, AWS announces Amazon RDS for MySQL zero-ETL integration with Amazon Redshift, and it's in public preview for the Redshift Serverless and Ra3 instance types.

(04:16):

Even more, DynamoDB zero-ETL integration with OpenSearch Service. This zero-ETL integration uses OpenSearch ingestion to synchronize the data from DynamoDB tables to OpenSearch. OpenSearch ingestion is able to automatically understand the format of the data in DynamoDB tables, and map the data to your index mapping templates in OpenSearch, to yield the most performance search results. Furthermore, customers can synchronize data from multiple DynamoDB tables into one OpenSearch managed cluster, or Serverless collection, to offer holistic insights across several applications.

(04:59):

Amazon DynamoDB now supports, yes, zero-ETL integration with Amazon Redshift, enabling customers to run high performance analytics on their DynamoDB data. This zero-ETL integration has no impact on production workloads running on DynamoDB. As data is written into a DynamoDB table, it is seamlessly made available in Amazon Redshift, eliminating the need for customers to build and maintain complex data pipelines for performing extract, transform, and load operations. So much zero-ETL goodness.

(05:34):

But now, we are going to move on to Graviton 4. So now, starting today, new memory-optimized Ec2 R8g instances, powered by the latest generation custom-designed AWS Graviton 4 processors, are available in preview. R8g instances are built on the AWS Nitro System. It's a collection of hardware and software innovations designed by AWS. The AWS Nitro System enables the delivery of efficient, flexible, and secure cloud services, with isolated multi-tenancy, private networking, and fast local storage. R8g instances offer larger instances, with up to 3x more vCPUs and memory than our 7g instances. Amazon Ec2 R8g instances are ideal for memory intensive workloads, such as your databases, in-memory caches, and real-time big data analytics. AWS Graviton 4 is the latest in the Graviton family of processors that are custom designed by AWS to provide the best price performance for workloads in Amazon EC2. They provide up to 30% better compute performance compared to AWS Graviton 3 processors.

(06:56):

AWS announces the preview of a new generative AI-based capability in Amazon DataZone to improve data discovery, data understanding, and data 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 required for analysis, which enhances data discoverability, and cuts down on back and forth communications with data producers. Data consumers, these are your data analysts, data engineers, data scientists, I see you, who are listening, you've got 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. Amazon DataZone AI recommendations for descriptions are available in preview, and Amazon DataZone domains provisioned in the US East Northern Virginia and US West Oregon regions.

(08:18):

And for even more generative AI, continued pre-training in Amazon Bedrock is a new capability that allows you to train Amazon Titan Text Express and Amazon Titan Text Lite foundation models, and customize them using your own unlabeled data in a secure and managed environment. As models are continually pre-trained on data spanning different topics, genres and contexts over time, they become more robust and learn to handle out-of-domain data better by accumulating wider knowledge and adaptability, creating even more value for your organization.

(08:59):

Organizations want to build domain-specific applications that reflect the terminology of their business. However, many foundation models are trained on large amounts of publicly available data, and are not suited to highly specialized domains. To adapt foundation models with knowledge more relevant to a domain, you can engage continued pre-training, which leverages vast sets of unlabeled data. Continued pre-training in Bedrock helps address out-of-domain learning challenges by exposing models to new diverse data beyond their original training. With continued pre-training, you can expand the models' understanding to include the language used in your domain, and improve the model's overall competency for your business.

(09:47):

And now generally available is fully managed Agents for Amazon Bedrock. This enables generative AI applications to execute multistep tasks across company systems and data sources. Agents can plan and perform business tasks, such as answering questions about product availability or taking orders. Customers can create an agent in just a few clicks by writing a few instructions in natural language, providing access to a company's systems, and defining AWS Lambda functions. Agents analyze the user requests, and break it down into a logical sequence using the foundation model's reasoning capabilities to determine what information is needed. The API is to call and the sequence of execution to fulfill this request. After creating the plan, agents call the right APIs and retrieve the information needed from the company's systems and data sources to provide accurate and relevant responses. Agents automatically perform this process in the background and securely by encrypting data in transit and at rest each time. This relinquishes customers from having to engineer prompts, train models, or manually connect systems.

(11:01):

With Agents for Amazon Bedrock, customers can easily integrate generative AI into their businesses, simplifying and accelerating how they perform and execute tasks without the undifferentiated heavy lifting. Amazon Bedrock now supports fine-tuning for Meta Llama 2 and Cohere Command Lite, along with Amazon Titan Text Lite and Amazon Titan Text Express foundation models. So you can use labeled data sets to increase model accuracy for particular tasks.

(11:31):

Organizations with small labeled data sets that want to specialize a model for a specific task use a process called fine-tuning, which adapts a model's parameters to produce outputs that are more specific to their business. Parameters represent what the model has learned during training, and adjusting them can refine the model's knowledge and capabilities to make decisions within an organization's context. Using a small number of labeled examples in Amazon S3, you can fine-tune a model without having to annotate large volumes of data. Bedrock makes a separate copy of the base foundation model that is accessible only by you, and trains this private copy of the model. None of your content is used to train the original base models, you can configure your Amazon VPC settings to access Amazon Bedrock APIs, and provide model fine-tuning data in a secure manner.

(12:32):

Knowledge Bases for Amazon Bedrock is now generally available. Fully managed Knowledge Bases for Amazon Bedrock securely connects foundation models to internal company data sources for retrieval augmented generation to deliver more relevant, context-specific, and accurate responses. Knowledge Bases extend the foundation model's powerful capabilities to make it more knowledgeable about your business, customers, and offerings. To equip the foundation model with up-to-date and proprietary information, organizations use retrieval augmented generation, which is a technique that fetches data from company data sources, and enriches the prompt to provide more relevant and accurate responses.

(13:16):

Implementing retrieval augmented generation requires an organization to perform several cumbersome steps to convert the data into embedding vectors, store the embeddings in a specialized vector database, and build custom integrations into the database to search and retrieve text relevant to the user's query. This can be time-consuming and inefficient. Knowledge Bases for Amazon Bedrock is a fully managed retrieval augmented generation capability that allows you to fully customize foundation model responses with contextual and relevant company data. Simply point to the location of your data in Amazon S3, and Knowledge Bases for Amazon Bedrock takes care of the entire ingested workflow into your vector database. But don't worry if you don't have an existing vector database, Amazon Bedrock creates an Amazon OpenSearch Serverless vector store for you. Simply point to your location of your data in S3, and Knowledge Bases for Bedrock takes care of the entire ingested workflow into your vector database.

(14:20):

And another exciting announcement from Amazon Bedrock is Guardrails. This is in preview, and enables customers to implement safeguards across foundation models based on their use cases and responsible AI policies. Customers can create multiple Guardrails tailored to different use cases and apply them across foundation models, providing a consistent user experience, and standardizing safety across generative AI applications. Customers need to safeguard their generative AI applications for a relevant and safe user experience. While many foundation models have built-in protections to filter undesirable and harmful content, customers may want to further tailor interactions specific to their use cases and adhering to their responsible AI policies.

(15:08):

One example is a bank might want to configure its online assistant to refrain from providing investment advice and limit harmful content. With Guardrails, customers can define a set of denied topics that are undesirable within the context of their application, and configure thresholds to filter harmful content across categories such as hate, insult, sexual and violence. Guardrails evaluate user queries and foundation model responses against the denied topics and content filters, helping to provide content that falls into restricted categories. This allows customers to closely manage user experiences based on application specific requirements and policies. Guardrails are supported for English content across text-based foundation models and fine-tuned models on Amazon Bedrock, as well as Agents for Amazon Bedrock.

(16:15):

And there's a new generative AI-powered assistant called Amazon Q, and this is in preview, and it's specifically designed for work that can be tailored to your business, to have conversations, solve problems, generate content, take actions, using the data and expertise found in your company's information repositories, code, and enterprise systems. Amazon Q has a broad base of general knowledge and domain-specific expertise. It is secure and private in practice and by design. Amazon Q is designed to help customers meet stringent enterprise requirements. With Amazon Q, employees can ask questions and get answers from knowledge spread across disparate content repositories, summarize lengthy reports, write articles, take actions, and much more, all within their company's connected content repository. Amazon Q offers over 40 built-in connectors to popular enterprise systems. It generates responses only from the content that each user is permitted to access with enterprise-grade access control. It also provides references and citations so users can trace which documents were used to provide a response.

(17:28):

Now, onto other really cool things that Amazon Q can help you with is a lot of AWS-related tasks. So Amazon Q can help you select Amazon EC2 instances. Amazon Q uses machine learning to help customers take quick and cost-effective decisions for their compute instance type before building their workloads. Amazon Q can provide expert assistance when building, deploying, and operating applications and workloads on AWS. Conversations with Q can help you use best practices to architect applications, explain source code, and implement application functionality right in the IDE, upgrade application versions, and troubleshoot errors, like network connectivity issues. It's like having a conversation with me and Simon, I love it.

(18:17):

Amazon Q has been trained on 17 years of high-quality AWS examples and documentation to provide guidance for every step of the process of building, deploying, maintaining, and operating applications on AWS. Amazon Q combines its expertise at building software and knowledge of your code base to help you understand your code, generate tasks, fix bugs, and even help implement full features, all in a fraction of the time. Additionally, Amazon Q Code Transformation helps you upgrade your Java applications to the latest language version right in the IDE. Amazon Q is available everywhere builders need it. You can find Amazon Q in the AWS Management Console, the AWS Console mobile app, the AWS documentation, AWS website, your IDE through Amazon CodeWhisperer, or through the AWS chatbot in team chat rooms on Slack or Microsoft Teams.

(19:15):

Now, developers can assign a CodeCatalyst issue to Amazon Q, and Q performs the heavy lifting of converting a human problem to an actionable plan. Then, Q completes code changes and a pull request that is assigned to the requester. Q will then monitor any associated workflows and attempt to correct any issues. The user can preview code changes and merge the pull request. Development teams can utilize this new capability as an end-to-end streamlined experience within Amazon CodeCatalyst, without having to enter the IDE. This new capability enables teams using CodeCatalyst to scale with AI to assist developers in completing everyday software development tasks. Developers can now go from an idea in an issue, to a fully tested, merge-ready running application code with natural language inputs, in just a few clicks.

(20:07):

Today, Amazon Quicksight announces three new natural language capabilities enabled by Amazon Q for business users. Launching in preview, these capabilities can summarize dashboards, generate mini dashboards to answer data questions, and build stories explaining data. First, executive summaries provide business users an at-a-glance view of key insights, and automatically compare data period over data period to surface important trends. Second, a new data Q&A experience helps business users confidently answer questions beyond their dashboards and reports. Amazon Q and Quicksight suggests sample questions to start, generates mini dashboards, present related data to explain answer context, and provides best available answers for vague questions. For example, a prompt of marketing campaigns can summarize lead generation across marketing campaigns with a breakdown by individual campaigns over time.

(21:07):

Business users can now generate data stories, a new type of customizable interactive data document, using simple natural language prompts, to help better explain data. For example, a marketing manager might ask you to create a story that shows how our social marketing campaigns performed from November to January, and include recommended actions. QuickSight Q would then generate a story describing campaign performance and recommendations, with data visualizations and text summaries. This information can then be presented either in an interactive blog-like format or as a slide presentation.

(21:46):

Amazon Q in Connect offers generative AI-powered agent assistance in real time. Amazon Connect now offers in-app web and video calling. Amazon Connect now supports two-way SMS. Amazon Connect announces a preview of Analytics Data Lake, a zero-ETL analytics capability that empowers organizations to access the insights needed to understand and optimize key context-centered performance metrics via a unified data source and business of intelligence tool. With the Analytics Data Lake, records are deduped and ready to query, limiting the need to build and maintain complex data pipelines to perform extract, transform, and load operations, to access Amazon Connect data, to get it ready for analytics and AI workloads.

(22:39):

Amazon Connect Contact Lens now provides real-time conversational analytics for chat. Amazon Connect Customer Profiles now provides a generative AI-powered customer data mapping capability that significantly reduces the time needed to create unified profiles, enabling companies to help provide more personalized customer experiences. Amazon Connect now provides a no-code UI builder, enabled within the drag-and-drop workflow designer, that lets you create and manage the UI shown to agents in step-by-step guides. With this capability, you can design a guide that presents what the agent should review or do inside the Amazon Connect agent workspace at any moment during a customer interaction. Amazon Connect Contact Lens now provides generative AI-powered post-contact summarization, enabling contact center managers to more efficiently monitor and help improve contact quality and agent performance.

(23:41):

That is all from today. Tomorrow, I'll share the latest announcements from Re:Invent, and Simon will be back on Thursday to cover more announcements, including those from Werner Vogels' keynote. We'd love to get your feedback, so shoot us an email at awspodcast@amazon.com, and until next time, keep on building.