The Deloitte On Cloud Podcast

Host, Gary Arora, Chief Architect for Cloud and AI Solutions, Deloitte Consulting LLP

Title: **Gary Arora breaks down the top three breakthroughs at AWS re:Invent 2024**

Description: AWS re:Invent 2024 is a wrap! In this episode, Gary Arora shares his top three highlights from this year's incredible event. Innovations include Amazon S3 updates with managed Iceberg tables and real-time metadata; the cutting-edge Trainium2 AI chips; and AWS Nova, a multimodal foundational model for generating text, image, and video. From faster analytics to trailblazing AI performance, these innovative offerings are setting new standards for digital transformation technology.

Duration: **00:04:51**

Gary Arora:
Hey, welcome back to the On Cloud Podcast. I'm your host, Gary Arora, live from AWS re:Invent 2024 in Las Vegas. It's been a busy week with big announcements from compute and databases to chips and AI. Here are my top three picks.

First up are the developments on Amazon S3 with S3 Tables and S3 Metadata. Amazon's S3, as we all know, is a simple storage service for storing objects, and one of the most widely used engines for querying the Parquet files in Amazon S3 is Apache Iceberg, and with S3 Tables, you now have fully managed support for Apache Iceberg for faster analytics, storing and managing tabular data at scale.

Think about running queries across billions of files containing petabytes, or even exabytes, of data. With Amazon S3 Metadata, it automatically captures metadata as the objects are uploaded real time into S3 Buckets. Metadata like size, source of objects, and even custom ones like tags or product SKUs, transaction IDs, and content rating. And it keeps this metadata updated as the objects change, making it easier to discover and understand your data for business analytics and real time inference applications.

Next up on my list is the next generation of AI chips with Trainium2. Trainium chips are a family of AI chips designed to train and deploy your most demanding models, like large language models, multimodal models, and diffusion transformers. And the race here has always been to reduce training times and inference latency. That's the time between when an AI system receives an input and generates the corresponding output.

As a reference point, the first-gen Trainium chip was launched in 2022, and Amazon used 80,000 Trainium and inferential chips at their most recent Prime Day to power Rufus, their shopping assistant. Trainium2 delivers up to four times the performance of its predecessor, and they are already powering latency-optimized versions of Llama 3.1, and Claude 3.5 models on Amazon Bedrock. But here's where it got really interesting. Apple joined the stage at AWS re:Invent to reveal how they are using Trainium1 for services like search and achieving a 40% boost in performance. And with Trainium2, Apple is exploring pre-training new models to build Apple Intelligence.

And topping my list at number one is Amazon Nova—AWS's brand new foundational model for generating text, images, and video. These models understand and generate content in over 200 languages. They can also understand up to 30 minutes of video content. And they can generate six-second studio quality video clips through text prompts and reference images on Nova Reel, which even allows you to customize camera motions like pans and 360-degree rotations.

And with the Nova Pro model, you have a context window of 300,000 tokens, which, for reference it, means it can process code bases with over 15,000 lines of code. This is setting new standards in multimodal intelligence and agentic workflows that require calling APIs and tools to complete complex workflows.

So, that's all for today's episode. Stay tuned for more tech talks, and thanks for listening to the Deloitte On Cloud podcast.

Operator:
This podcast is produced by Deloitte. The views and opinions expressed by podcast speakers and guests are solely their own and do not reflect the opinions of Deloitte. This podcast provides general information only and is not intended to constitute advice or services of any kind. For additional information about Deloitte, go to Deloitte.com/about.

Visit the On Cloud library
www.deloitte.com/us/cloud-podcast