REPORT REPRINT

AWS re:Invents lake house architecture for data and analytics

JANUARY 5 2021

By Matt Aslett

As always, Amazon Web Services used its re:Invent customer event to deliver a bewildering number of announcements both major and minor. At the heart of it all was an expanded commitment to the concept of the lake house architecture.

451 Research

S&P Global
Market Intelligence

Introduction

As always, Amazon Web Services used its re:Invent customer event to deliver a bewildering number of announcements involving major new and updated product releases, as well as minor feature and functionality enhancements across a variety of product categories. Data and analytics were no exception, with numerous announcements that spanned the breadth of AWS's product portfolio, including relational and nonrelational operational databases, data warehouses, machine learning, big-data processing, interactive query, real-time analytics, and business intelligence. At the heart of it all was an expanded commitment to the concept of the lake house architecture, which provides a strategic vision of how multiple AWS data and analytics services can be combined into a multi-purpose data processing and analytics environment.

451 TAKE

AWS now offers a broad portfolio of data-related services. While this supports a customer's ability to choose their own preferred approach for any given workload, it also presents the potential for complexity as customers become increasingly reliant on a combination of various interdependent data processing and analytics services. While the company's reference to the lake house architecture is not new, it is interesting to see a difference in the way it is being presented to customers – not as a term to describe the functional coexistence of data warehousing and data lake concepts, but as a strategic vision of how multiple data and analytics services can be combined into a coherent, multifaceted data processing and analytics environment. We expect this architectural vision will play well with enterprises as they continue to increase their investment in AWS, providing confidence that the company can serve as a strategic partner for large-scale data and analytics initiatives, as well as a provider of multiple discreet data-related services.

Context

When it comes to databases and data processing, AWS's perspective can be summed up as 'one size fits nothing at all.' Rather than try and standardize on a single database or data platform for multiple workloads, AWS maintains that a better approach for customers is to choose from a portfolio of purpose-built databases accessed via fully managed APIs.

The thinking is that this approach enables the user to understand the specific requirements of their use case, data type and access patterns, and then pick the right database to support that individual workload. At the same time, there are potential disadvantages to using multiple services (not least management and data flow complexity). That is where the company's commitment to the lake house architecture, tying a variety of services into a coherent data architecture, comes in.

Building out the lake house

AWS originally started using the term 'lake house' a year ago in relation to Amazon Redshift Spectrum, its service that enables users of the Amazon Redshift data-warehouse service to apply queries to data stored in Amazon Simple Storage Service (Amazon S3). While Amazon Redshift continues to play a role in the lake house architecture, it was clear from this year's re:Invent that AWS has expanded its conceptual view of the architecture beyond the data warehouse.

At the heart of the lake house architecture, as it was described at this year's re:Invent, is a combination of the Amazon S3 cloud storage and Amazon Athena interactive query services (reflecting what we have described as an abstracted data architecture), along with the AWS Glue data integration service and AWS Lake Formation.

The latter was first launched in 2018 and facilitates building and governing Amazon S3-based data lakes. Key components of AWS Lake Formation include a catalog of assets in the data lake, blueprints to ingest data from a variety of sources and a centralized permissions model that supports fine-grain security policies. At re:Invent, AWS announced the preview release of AWS Lake Formation transactions, row-level security and acceleration. The former is enabled by the introduction of governed tables – a new Amazon S3 table type that supports atomic, consistent, isolated and durable (ACID) transactions, which will allow multiple users to concurrently insert, delete and modify data stored in governed tables.

Meanwhile, AWS also announced the gated preview release of AWS Glue Elastic Views, a new capability of AWS Glue that enables users to create materialized views to combine and replicate data across multiple data stores (initially Amazon DynamoDB, Amazon S3, Amazon Redshift and Amazon Elasticsearch Service, with Amazon RDS, Amazon Aurora and others to follow). AWS Glue Elastic Views goes beyond simple federated query by adding the ability to precompute results and update the target data store as underlying data in the source data store changes. AWS Glue Elastic Views therefore enables customers to continue to use dedicated data stores for particular use cases (such as Amazon Redshift for reporting on large volumes of structured data), and enables the analysis of that data alongside unstructured data in the data lake, as required.

While the foundations of the lake house architecture are Amazon S3, Amazon Athena, AWS Glue and AWS Lake Formation, there continues to be an important role for Amazon Redshift as one of the various purpose-built database services that can be applied to data in the data lake, alongside relational and nonrelational operational databases (including Amazon RDS, Amazon Aurora and Amazon DynamoDB), machine learning (Amazon SageMaker), big-data processing (Amazon EMR), log analytics (Amazon Elasticsearch), and real-time analytics (Amazon Kinesis). In addition to expanding the lake house architecture, AWS also announced a variety of new features and capabilities for each of these services.

Amazon Redshift announcements

In addition to Amazon Redshift Spectrum, other important capabilities of Amazon Redshift, in relation to the lake house architecture, are Data Lake Export, which saves the results of a Redshift query to S3, and Federated Query, which enables Amazon Redshift to query data directly in Amazon RDS and Aurora PostgreSQL stores. This year's re:Invent conference also saw the company announce the preview release of the ability to perform federated querying of data in Amazon Aurora MySQL and Amazon RDS for MySQL.

AWS made a number of other announcements related to Amazon Redshift, not least the preview of Amazon Redshift ML, which enables customers to use SQL commands to launch Amazon SageMaker to train machine learning models with data from Redshift and subsequently deploy the model in Redshift for inference. Amazon Redshift ML users can choose their preferred models or use Amazon SageMaker Autopilot to automatically select models, and can generate results from the trained models using standard SQL queries.

AWS also announced the addition of a new SUPER data type to store semi-structured data or documents (such as JSON) with Amazon Redshift (taking advantage of the PartiQL SQL-compatible query language), as well as the preview of AQUA (Advanced Query Accelerator) for Amazon Redshift. AQUA is a high-speed cache that can be used to push and compute to the storage layer for acceleration, taking advantage of RA3 managed storage, AWS Nitro custom hypervisor, security and I/O acceleration hardware, and FPGAs.

The company also announced the preview of data sharing with Amazon Redshift, making it possible to provide read access to data in Amazon Redshift across different clusters without the need to copy or move data. Redshift data sharing also takes advantage of Amazon Redshift RA3 managed storage, which separates compute and storage for Redshift.

The company's long-term vision is to enable sharing with other services (such as Amazon Athena, SageMaker or EMR), while the ability to share data across different accounts (e.g., with partners or suppliers) is also part of the preview, as a complement to the AWS Data Exchange data marketplace.

Amazon EMR, Aurora, Kinesis and QuickSight

While Amazon Redshift perhaps had the lion's share of data-related announcements at re:Invent 2020, there were plenty of other announcements, including the availability of Amazon Neptune ML, which uses graph neural networks (GNNs) to provide predictions on graph data in AWS's Neptune graph database. AWS also announced the ability to run Apache Spark on Kubernetes with Amazon EMR on Amazon EKS, as well as the preview release of Amazon EMR Studio, a managed Jupyter notebook environment that enables developers and data scientists to develop and run applications and code on Amazon EMR on EC2 or Amazon EMR on Amazon EKS.

Amazon's Aurora database service was also the focus of announcements, including the preview of Babelfish for Amazon Aurora, a new translation layer that provides support for the Microsoft SQL Server wire protocol and T-SQL query language, enabling applications written for SQL Server to run on Aurora PostgreSQL. AWS also announced the preview release of Aurora Serverless v2, with scaling and capacity improvements, as well as support for multi-availability zones, global database configuration and read replicas.

AWS's QuickSight business intelligence service also received a boost with the addition of the Amazon QuickSight Q natural language query functionality, as well as the ability to visualize data from Amazon Elasticsearch Service. The company also recently announced long-term retention of data in Amazon Kinesis Data Streams, enabling users of the data streaming service to analyze historical data (up to a year old), as well as real-time data streams.

Additionally, AWS made a series of announcements related to Amazon SageMaker and its machine learning and artificial intelligence services, which will be covered in a forthcoming report.

Competition

The closest competition for AWS comes from Microsoft Azure, Google Cloud, Alibaba Cloud, and established industry giants including IBM, Oracle, TIBCO and SAP. For data processing and analytics workloads in particular, the company also faces a variety of specialist competitors in addition to those already mentioned. This list is almost endless, but a few examples would include Snowflake, Teradata, Cloudera and Yellowbrick Data in data warehousing; MongoDB, DataStax, Couchbase and Actian in operational databases; Databricks and Qubole in big-data processing; Starburst and Ahana in distributed SQL analytics; Informatica, Alation and Hitachi Vantara in data integration and data management; and Salesforce's Tableau and Qlik in business intelligence. As these vendors build out cloud services to complement existing on-premises offerings, they have the potential advantage of leveraging on-premises customer relationships. While the large industry players similarly have diverse portfolios that could be compared with the ranges of services offered by AWS as part of its lake house architecture, the other vendor that is significantly committed to the lake house (or in its case 'lakehouse') concept is Databricks.

SWOT Analysis

STRENGTHS

AWS is the go-to provider of cloud services for many companies – having taken the lead in establishing itself as an alternative to on-premises datacenters – and has built out a broad portfolio of data-related services.

WEAKNESSES

The breadth of diverse options AWS provides for data and analytics services is large enough to potentially be bewildering to the uninitiated, and complex to manage for large-scale adopters.

OPPORTUNITIES

The lake house as an architectural vision could provide greater confidence that AWS can serve as a strategic partner for large-scale data and analytics initiatives, as well as a provider of multiple discreet data-related services.

THREATS

The company faces a variety of competitors, including industry giants and specialists, in multiple segments. Many of these have the theoretical advantage of leveraging existing relationships forged on-premises.