AWS Announces Two New Capabilities to Move Toward a Zero-ETL Future on AWS

Amazon Aurora zero-ETL integration with Amazon Redshift enables customers to analyze petabytes of transactional data in near real time, eliminating the need for custom data pipelines

Amazon Redshift integration for Apache Spark makes it easier and faster for customers to run Apache Spark applications on data from Amazon Redshift using AWS analytics and machine learning services

LAS VEGAS—Nov. 29, 2022—At AWS re:Invent, Amazon Web Services, Inc. (AWS), an Amazon.com, Inc. company (NASDAQ: AMZN), today announced two new integrations that make it easier for customers to connect and analyze data across data stores without having to move data between services. Today's announcement enables customers to analyze Amazon Aurora data with Amazon Redshift in near real time, eliminating the need to extract, transform, and load (ETL) data between services. Customers can also now run Apache Spark applications easily on Amazon Redshift data using AWS analytics and machine learning (ML) services (e.g., Amazon EMR, AWS Glue, and Amazon SageMaker). Together, these new capabilities help customers move toward a zero-ETL future on AWS. To learn more about unlocking the value of data using AWS, visit aws.amazon.com/data.

"The vastness and complexity of data that customers manage today means they cannot analyze and explore it with a single technology or even a small set of tools. Many of our customers rely on multiple AWS database and analytics services to extract value from their data, and ensuring they have access to the right tool for the job is important to their success," said Swami Sivasubramanian, vice president of Databases, Analytics, and Machine Learning at AWS. "The new capabilities announced today help us move customers toward a zero-ETL future on AWS, reducing the need to manually move or transform data between services. By eliminating ETL and other data movement tasks for our customers, we are freeing them to focus on analyzing data and driving new insights for their business—regardless of the size and complexity of their organization and data."

Data is at the center of every application, process, and business decision and is the cornerstone of almost every organization's digital transformation. But, real-world data systems are often sprawling and complex, with diverse data dispersed across multiple services and on-premises systems. Many organizations are sitting on a treasure trove of data and want to maximize the value they get out of it. AWS provides a range of purpose-built tools like Amazon Aurora, to store transactional data in MySQL and PostgreSQL-compatible relational databases, and Amazon Redshift, to run high-performance data warehousing and analytics workloads on petabytes of data. But to truly maximize the value of data, customers need these tools to work together seamlessly. That is why AWS has invested in zero-ETL capabilities like Amazon Aurora ML and Amazon Redshift ML, which let customers take advantage of Amazon SageMaker for ML-powered use cases, without moving data between services. Additionally, AWS provides seamless data ingestion from AWS streaming services (e.g., Amazon Kinesis and Amazon MSK) into a wide range of AWS data stores, such as Amazon Simple Storage Service (Amazon S3) and Amazon OpenSearch Service, so customers can analyze data as soon as it is available. Today's announcement builds on the strength and deep integrations of AWS's database and analytics portfolio to make it faster, easier, and more cost-effective for customers to access and analyze data across data stores on AWS.

Amazon Aurora zero-ETL integration with Amazon Redshift makes it easier to run petabyte-scale analytics on transactional data in Amazon Aurora in near real time with Amazon Redshift

The requirement for near real-time insights on transactional data (e.g., purchases, reservations, and financial trades) grows as organizations seek to better understand core business drivers and develop strategies to increase sales, reduce costs, and gain a competitive advantage. Many organizations today rely on a three-part solution to analyze their transactional data—a relational database to store data, a data warehouse to perform analytics, and a data pipeline to ETL data between the relational database and the data warehouse. Data pipelines can be costly to build and challenging to manage, requiring developers to write custom code and constantly manage the infrastructure to ensure it scales to meet demand. Some companies maintain entire teams just to facilitate this process. Additionally, it can take days before data is ready for analysis, and intermittent data transfer errors can delay access to time-sensitive insights even further, leading to missed business opportunities.

With Amazon Aurora zero-ETL integration with Amazon Redshift, transactional data is automatically and continuously replicated seconds after it is written into Amazon Aurora and seamlessly made available in Amazon Redshift. Once data is available in Amazon Redshift, customers can start analyzing it immediately and apply advanced features like data sharing and Amazon Redshift ML to get holistic and predictive insights. Customers can replicate data from multiple Amazon Aurora database clusters into the same Amazon Redshift instance to derive insights across several applications. Now, customers can use Amazon Aurora to support their transactional database needs and Amazon Redshift to power their analysis, without building or maintaining complex data pipelines.

Amazon Redshift integration for Apache Spark makes it easier to use AWS analytics and ML services to build and run Apache Spark applications on data from Amazon Redshift

Many developers use Apache Spark, an open-source processing framework used for big data workloads, to support a broad range of analytics and ML applications. Today, AWS supports Apache Spark on Amazon EMR, AWS Glue, and Amazon SageMaker with a fully compatible, AWS-optimized runtime that is 3x faster than open source. Customers often want to analyze Amazon Redshift data directly from these services. This requires them to go through the complex, time-consuming process of finding, testing, and certifying a third-party connector to help read and write the data between their environment and Amazon Redshift. Even after they have found a connector, customers must manage intermediate data-staging locations, such as Amazon S3, to read and write data from and to Amazon Redshift. All of these challenges increase operational complexity and make it difficult for customers to use Apache Spark to its full extent.

Amazon Redshift integration for Apache Spark makes it easier for developers to build and run Apache Spark applications on data in Amazon Redshift using AWS-supported analytics and ML services. Amazon Redshift integration for Apache Spark is certified, packaged, and supported by AWS, eliminating the cumbersome and error-prone process associated with third-party connectors. Developers can begin running queries on Amazon Redshift data from Apache Spark-based applications within seconds using popular language frameworks (e.g., Java, Python, R, and Scala). Intermediate data-staging locations are managed automatically, eliminating the need for customers to configure and manage these in application code. To get started with Amazon Redshift integration for Apache Spark, visit aws.amazon.com/redshift/features/integration-for-apache-spark.

Adobe empowers everyone, from individuals and small businesses to government agencies and global brands, to create and deliver exceptional digital experiences. “Adobe’s mission is to change the world through digital experiences, and in today’s world, that means having analytics that can deliver both deep and real-time insights,” said Jack Lull, principal scientist for Adobe Acrobat Sign. “As an Amazon Aurora customer, we are excited for Amazon Aurora support for zero-ETL integration with Amazon Redshift, which will provide our growing Acrobat Sign customer base with new insights and faster analytics performance as their usage increases—all without the need for ongoing maintenance for our own teams."

Infor is a global leader in business cloud software and industry-specific enterprise resource planning solutions. "At Infor, we use AWS to build and deploy modern tools to help our customers transform their business and accelerate innovation. This includes a new managed data warehouse service for our customers' industry cloud data, which will help our customers make faster decisions with advanced analytics and ML," said Jim Plourde, senior vice president for Cloud Services at Infor. "We are excited for Amazon Aurora to support zero-ETL integration with Amazon Redshift, which will reduce our operational burden by making transactional data from Amazon Aurora available in Amazon Redshift in near real time. Now, we can benefit from the performance of Amazon Aurora as our relational database management system, while easily leveraging the analytics and ML capabilities in Amazon Redshift for our new managed data warehouse service."

GE Aerospace is a global provider of jet engines, components, and systems for commercial and military aircraft. The company has been designing, developing, and manufacturing jet engines since World War I. "Amazon Redshift is a focal point of our strategy to make data extremely accessible and usable across our organization," said Alcuin Weidus, senior principal data architect at GE Aerospace. "Data scientists, engineers, and developers leverage Apache Spark to build data products and run analytics workloads on Amazon EMR, AWS Glue, and third-party ML platforms hosted on AWS. We are excited for the Amazon Redshift integration for Apache Spark, which will streamline our developers' building process and help make applications more performant and secure."

The Goldman Sachs Group, Inc. is a leading global financial institution that delivers a broad range of financial services across investment banking, securities, investment management, and consumer banking to a large and diversified client base that includes corporations, financial institutions, governments, and individuals. "Our focus is on providing self-service access to data for all of our users at Goldman Sachs. Through Legend, our open source data management and governance platform, we enable users to develop data-centric applications and derive data-driven insights as we collaborate across the financial services industry," said Neema Raphael, chief data officer at Goldman Sachs. "With Amazon Redshift integration for Apache Spark, our data platform team will be able to access Amazon Redshift data with minimal manual steps—allowing for zero-code ETL that will increase our ability to make it easier for engineers to focus on perfecting their workflow as they collect complete and timely information. We expect to see a performance improvement of applications and improved security as our users can now easily access the latest data in Amazon Redshift."

About Amazon Web Services

For over 15 years, Amazon Web Services has been the world's most comprehensive and broadly adopted cloud offering. AWS has been continually expanding its services to support virtually any cloud workload, and it now has more than 200 fully featured services for compute, storage, databases, networking, analytics, machine learning and artificial intelligence (AI), Internet of Things (IoT), mobile, security, hybrid, virtual and augmented reality (VR and AR), media, and application development, deployment, and management from 96 Availability Zones within 30 geographic regions, with announced plans for 15 more Availability Zones and five more AWS Regions in Australia, Canada, Israel, New Zealand, and Thailand. Millions of customers—including the fastest-growing startups, largest enterprises, and leading government agencies—trust AWS to power their infrastructure, become more agile, and lower costs. To learn more about AWS, visit aws.amazon.com.

About Amazon
Amazon is guided by four principles: customer obsession rather than competitor focus, passion for invention, commitment to operational excellence, and long-term thinking. Amazon strives to be Earth's Most Customer-Centric Company, Earth's Best Employer, and Earth's Safest Place to Work. Customer reviews, 1-Click shopping, personalized recommendations, Prime, Fulfillment by Amazon, AWS, Kindle Direct Publishing, Kindle, Career Choice, Fire tablets, Fire TV, Amazon Echo, Alexa, Just Walk Out technology, Amazon Studios, and The Climate Pledge are some of the things pioneered by Amazon. For more information, visit amazon.com/about and follow @AmazonNews.