AWS Announces Five New Database and Analytics Capabilities

Amazon DocumentDB Elastic Clusters scales customers' document workloads to support millions of writes per second and store petabytes of data

Amazon OpenSearch Serverless helps customers run search and analytics workloads without having to configure, scale, or manage underlying infrastructure

Amazon Athena for Apache Spark enables customers to get started with interactive analytics using Apache Spark in less than a second, instead of minutes

AWS Glue Data Quality cuts time for data analysis and rule identification from days to hours by automatically measuring, monitoring, and managing data quality in data lakes and across data pipelines

Amazon Redshift now supports a high availability configuration across multiple AWS Availability Zones

LAS VEGAS—Nov. 30, 2022—At AWS re:Invent, Amazon Web Services, Inc. (AWS), an Amazon.com, Inc. company (NASDAQ: AMZN), today announced five new capabilities across its database and analytics portfolios that make it faster and easier for customers to manage and analyze data at petabyte scale. These new capabilities for Amazon DocumentDB (with MongoDB compatibility), Amazon OpenSearch Service, and Amazon Athena make it easier for customers to run high-performance database and analytics workloads at scale. Additionally, AWS announced a new capability for AWS Glue to automatically manage data quality across data lakes and data pipelines. Finally, Amazon Redshift now offers support for a high availability configuration across multiple AWS Availability Zones (AZs). Today's announcement helps customers get the most out of their data on AWS by empowering them to access the right tools for their data workloads, operate at scale, and increase availability. To learn more about unlocking the value of data using AWS, visit aws.amazon.com/data.

"Data is inherently dynamic, and harnessing it to its full potential requires an end-to-end data strategy that can scale with a customer's needs and accommodate all types of use cases—both now and in the future," said Swami Sivasubramanian, vice president of Databases, Analytics, and Machine Learning at AWS. "To help customers make the most of their growing volume and variety of data, we are committed to offering the broadest and deepest set of database and analytics services. The new capabilities announced today build on this by making it even easier for customers to query, manage, and scale their data to make faster, data-driven decisions."

Organizations today create and store petabytes—or even exabytes—of data from a growing number of sources (e.g., digital media, online transactions, and connected devices). To maximize the value of this data, customers need an end-to-end data strategy that provides access to the right tools for all data workloads and applications, along with the ability to perform reliably at scale as the volume and velocity of data increase. To support customers designing their own end-to-end data strategies, AWS offers the industry's most comprehensive set of data services and solutions. This includes fully managed databases optimized for customers' most important use cases, such as Amazon Aurora for relational databases and Amazon DocumentDB for document databases. It also includes a broad range of analytics services to help customers gain valuable insights from their data, including Amazon OpenSearch Service for search and analytics workloads (e.g., real-time application monitoring, log analytics, and website search), Amazon Athena for interactive analytics, AWS Glue for data integration, and Amazon Redshift for data warehousing. Today's announcement builds on these services with advanced capabilities.

- **Amazon DocumentDB Elastic Clusters power petabyte-scale applications with millions of writes per second:** Tens of thousands of customers use Amazon DocumentDB to run their document workloads because it is fast, scalable, highly available, and fully managed. While each Amazon DocumentDB node can scale up to 64 tebibytes of data and support millions of read requests per second, a subset of customers with extremely demanding workloads needs the ability to scale beyond these limits to support millions of writes per second and store petabytes of data. Previously, these customers had to manually distribute data and manage capacity across multiple Amazon DocumentDB nodes. Amazon DocumentDB Elastic Clusters allow customers to scale beyond the limits of a single database node within minutes, supporting millions of reads and writes per second and storing up to 2 petabytes of data. As workload demands increase, Amazon DocumentDB Elastic Clusters take advantage of a distributed storage system to automatically divide large datasets across multiple nodes. This removes the need for customers to write custom code to distribute datasets and manually manage capacity across nodes. The underlying infrastructure is managed automatically, so customers can easily scale capacity based on their needs without needing to provision, scale, or manage database clusters. To learn more about Amazon DocumentDB Elastic Clusters, visit aws.amazon.com/documentdb/features/#elastic_clusters.
- **Amazon OpenSearch Serverless automatically scales search and analytics workloads:** To power use cases like website search and real-time application monitoring, tens of thousands of customers use Amazon OpenSearch Service. Many of these workloads are prone to sudden, intermittent spikes in usage, making capacity planning difficult. Amazon OpenSearch Serverless automatically provisions, configures, and scales OpenSearch infrastructure to deliver fast data ingestion and millisecond query responses, even for unpredictable and intermittent workloads. With Amazon OpenSearch Serverless, data ingestion and search resources scale independently, allowing these operations to run concurrently without any performance impact. Customers using Amazon OpenSearch Serverless get access to serverless benefits (e.g., automatic provisioning, on-demand scaling, and pay-for-use pricing), along with Amazon OpenSearch Service features, such as built-in data visualizations, that help them understand log data, identify anomalies, and see search relevance rankings. To learn more about Amazon OpenSearch Serverless, visit aws.amazon.com/opensearch-service/features/serverless.
- **Amazon Athena for Apache Spark accelerates startup of interactive analytics to less than one second**: Customers use Amazon Athena, a serverless interactive query service, because it is one of the easiest and fastest ways to query petabytes of data in Amazon Simple Storage Service (Amazon S3) using a standard SQL interface. Many customers are looking for that same ease of use when it comes to using Apache Spark, an open-source processing framework for big data workloads that supports popular language frameworks (i.e., Java, Scala, Python, and R). While developers enjoy the fast query speed and ease of use of Apache Spark, they do not want to invest time setting up, managing, and scaling their own Apache Spark infrastructure each time they want to run a query. Now, with Amazon Athena for Apache Spark, customers do not have to provision, configure, and scale resources themselves. Interactive Apache Spark applications start in less than one second and execute faster than open source using AWS's optimized Spark runtime. Because Amazon Athena is integrated with other AWS services, customers can query data from multiple sources, chain calculations together for complex analyses, and visualize the results. Amazon Athena for Apache Spark automatically determines the resources required based on application demand and scales as needed, so customers only pay for the queries they run. To get started with Amazon Athena for Apache Spark, visit aws.amazon.com/athena/spark.

- **AWS Glue Data Quality automatically monitors and manages data freshness, accuracy, and integrity**: Hundreds of thousands of customers use AWS Glue to build and manage modern data pipelines quickly, easily, and cost-effectively. Organizations need to monitor the data quality—a measure of the freshness, accuracy, and integrity of data—of the information in their data lakes and data pipelines to ensure it is high quality before using it to power their analysis or machine learning applications. But effective data-quality management is a time-consuming and complex process, requiring data engineers to spend days gathering detailed statistics on their data, manually identifying data-quality rules based on those statistics and applying them across thousands of datasets and data pipelines. Once these rules are implemented, data engineers must continuously monitor for errors or changes in the data to adjust rules accordingly. AWS Glue Data Quality automatically measures, monitors, and manages the data quality of Amazon S3 data lakes and AWS Glue data pipelines, reducing the time for data analysis and rule identification from days to hours. AWS Glue Data Quality computes statistics for customer datasets (e.g., minimums, maximums, histograms, and correlations) and uses them to automatically recommend rules to ensure data freshness, accuracy, and integrity. Customers can schedule AWS Glue Data Quality to run periodically as data changes, automatically analyzing the data and proposing changes to quality rules to ensure relevance. Data engineers can configure actions to alert users or stop data pipelines when quality issues occur, without having to write code. To learn more about AWS Glue Data Quality, visit aws.amazon.com/glue/features/data-quality.
- **Amazon Redshift now supports multi-AZ deployments**: Tens of thousands of AWS customers collectively process exabytes of data with Amazon Redshift every day. To support these customers' mission-critical workloads, Amazon Redshift offers capabilities that increase availability and reliability, such as automatic backups and the ability to relocate a cluster to another AZ in minutes. Many databases today use a primary-standby replication mode to support high availability where a single database serves live traffic, and standby copies replicate data from the live version in case they need to replace it. Building on these capabilities, Amazon Redshift now offers a high-availability configuration to enable fast recovery while minimizing the risk of data loss. With Amazon Redshift Multi-AZ, clusters are deployed across multiple AZs and use all the resources to process read and write queries, eliminating the need for under-utilized standby copies and maximizing price performance for customers. Since a multi-AZ data warehouse is still managed as a single Amazon Redshift data warehouse with one endpoint, no application changes are required to maintain business continuity. To learn more about Amazon Redshift Multi-AZ, visit aws.amazon.com/redshift/reliability.

Rippling brings together payroll, benefits, HR, IT, and more so their customers can manage employee operations in one place. "As our business continues to grow, we need the ability to scale beyond the limits of a single document database node," said Nitin Aggarwal, data engineering lead at Rippling. "Amazon DocumentDB Elastic Clusters will help us solve this challenge by enabling us to quickly and easily scale to support millions of reads and writes per second and store petabytes of data. We are excited to explore Amazon DocumentDB Elastic Clusters as our business and customer demands grow."

riskCanvas, a software as a service (SaaS) product offering from Genpact, is a financial crime compliance solution that leverages cutting-edge big data, automation, and machine learning technologies to deliver compliance, efficiency, and automation to its clients. "riskCanvas' Entity-Centric Monitoring incorporates transaction monitoring, external enrichment, watchlist screening, and negative news to automatically assess risk and alert high-risk customers only as the true risk of a customer exceeds predefined thresholds, substantially reducing the effort to meet regulatory compliance requirements. This requires

significant and varied analytic processing that often experiences spiky and unpredictable data load," said Ryan Skousen, chief technology officer at riskCanvas and vice president of technology at Genpact Financial Crimes. "We are excited about Amazon OpenSearch Serverless, which will scale automatically to meet the data ingestion and analytic processing requirements of our workloads, and then scale back down as demand decreases to reduce costs drastically—all with no reengineering or maintenance impact."

FINRA, a regulator for securities firms doing business with the public in the US, regulates trading in equities, bonds, and options. "At FINRA, we develop applications on Amazon Athena to enable analysts and business partners to securely query financial trading data with multiple terabytes in daily updates," said Ratnakar Korem, senior director at FINRA. "We are excited about Amazon Athena for Apache Spark, which will bring the speed and ease of use we enjoy with Amazon Athena to our on-demand and batch analytics. This serverless feature will enable FINRA to conduct analytics against Big Data without the overhead of explicitly defining compute resources and tuning Apache Spark performance. This ultimately helps regulatory users and data analysts quickly respond to changing market dynamics and share results with others in a cost-effective and timely manner."

United Airlines operates a large domestic and international route network, spanning cities large and small across the US and all six inhabited continents. "United Airlines is building hundreds of data- and analytics-driven tools for our customers and employees, which makes managing and maintaining data quality critical to our operations," said Sarang Bapat, director of Data Engineering at United Airlines. "We are excited about AWS Glue Data Quality, which will enable us to automatically identify, analyze, and act on data-quality issues in a matter of minutes. This will help us make informed, timely, and accurate decisions and save countless hours in manually identifying and fixing all data issues."

Janssen Pharmaceuticals, a subsidiary of Johnson & Johnson, researches and manufactures medicines with a focus on the changing needs of patients and the healthcare industry. "Janssen Pharmaceutical uses Amazon Redshift to enable critical insights that drive important business decisions for our data scientists, data stewards, business users, and external stakeholders," said Shyam Mohapatra, director of Information Technology at Janssen Pharmaceutical Companies of Johnson & Johnson. "With Amazon Redshift Multi-AZ, we can be confident that our data warehouse will be available without any disruptions that might delay or impact our ability to make important business decisions."

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.