Conserve With Us Tries a New Approach to Data Pipeline Based on AWS Step Functions

CONSERVE
WITH US

Executive Summary

Conserve With Us is a technology platform that allows land conservancies to connect with people and communities to protect outdoor spaces. From local streams to majestic national parks, Conserve With Us is driven by a firm belief that greater investment in natural spaces will benefit all Americans of current and future generations.

Founded by a group of individuals with a deep and passionate interest in conserving land, Conserve With Us helps companies build brands that have a positive environmental impact while enhancing their web traffic. Through an innovated embedded technology, Conserve With Us makes it easy for customers to take action, like signing petitions and finding volunteer events, without leaving their favorite websites. With the continued rise of conscious consumerism, Conserve With Us allows consumers to simultaneously support the brands and lands they love.

The Challenge

When Conserve With Us anticipated a considerable increase in its workload, two bottlenecks stood in the way of the operation of its AWS-based technology platform: its PostgreSQL database and its analytics pipeline.

Conserve With Us needed a new, AWS-centric data pipeline solution that would mitigate the bottlenecks while offering greater scalability to handle increased workloads. A key requirement for the new solution was greater cost efficiency, requiring fewer staff resources for management and administration.

The ClearScale Solution

ClearScale, an AWS Premier Consulting Partner, was approached by Conserve With Us to solve their bottleneck and to reduce their costs. An early analysis determined that the AWS Data Pipeline in use by Conserve With Us lacked the necessary flexibility to handle greater workloads.

After evaluating various options, ClearScale took a unique approach by using AWS Step Functions, a general-purpose workflow management tool, for data pipeline orchestration. Used primarily for app development, AWS Step Functions typically orchestrates complex flows using Lambda functions, but ClearScale determined how to use its attributes for data pipeline orchestration.

AWS Step Functions coordinates multiple AWS services into serverless workflows. Consisting of a series of steps in which the output of one acts as the input for the next, Step Functions translate into easy-to-understand state machine diagrams. The state of each step is logged for quick and easy problem diagnosis and debugging. AWS Step Functions also helps achieve significant cost savings, with no costs or personnel required for provisioning, scaling, and managing servers.

Amazon Aurora was chosen for the database, providing speeds up to three times faster than standard MySQL/PostgreSQL databases with high security, availability, and reliability at 1/10th the cost. Aurora is fully managed by Amazon RDS, which automates tasks such as database setup, patching, and backups, and delivers high performance and availability including point-in-time recovery, continuous backup to Amazon S3, and replication across three Availability Zones. The Aurora cluster was set up with six copies of the database distributed across three Availability Zones The solution also uses Amazon Athena, a serverless, interactive query service that makes it easy to analyze data in Amazon S3 using well-known SQL.

In addition, the use of Aurora auto-scaling dynamically adjusts the number of Aurora replicas provisioned for an AuroraDB cluster using single-master replication. This enables the AuroraDB cluster to handle sudden workload increases. When the workload decreases, unneeded replicas are removed, so the customer doesn't have to pay for unused provisioned DB instances.

GraphQL code, which resides in Lambda functions was tuned to utilize only Aurora Reader replica for read-only queries, and Aurora Writer for requests that change state.

Additional infrastructure included Amazon S3 as the primary storage platform for the solution's associated data lake due to its virtually unlimited scalability, and Amazon Kinesis for data collection, processing, and analysis in real-time.

All data types can be stored in their native formats. It also integrates with services such as Amazon Athena and AWS Glue to query and process data, as well as with AWS Lambda serverless computing to run code without provisioning or managing servers. The solution uses an instrumental Glue feature to synchronize some of the computations from the Athena back to the Aurora Postgres.

To ensure that the created solution can support the high-traffic demand, an auto-scaling EC2 group was set up with the JMeter load testing suite, using Infrastructure-as-a-Code practices. This design allows for an almost unlimited test load.

The Solution Diagram
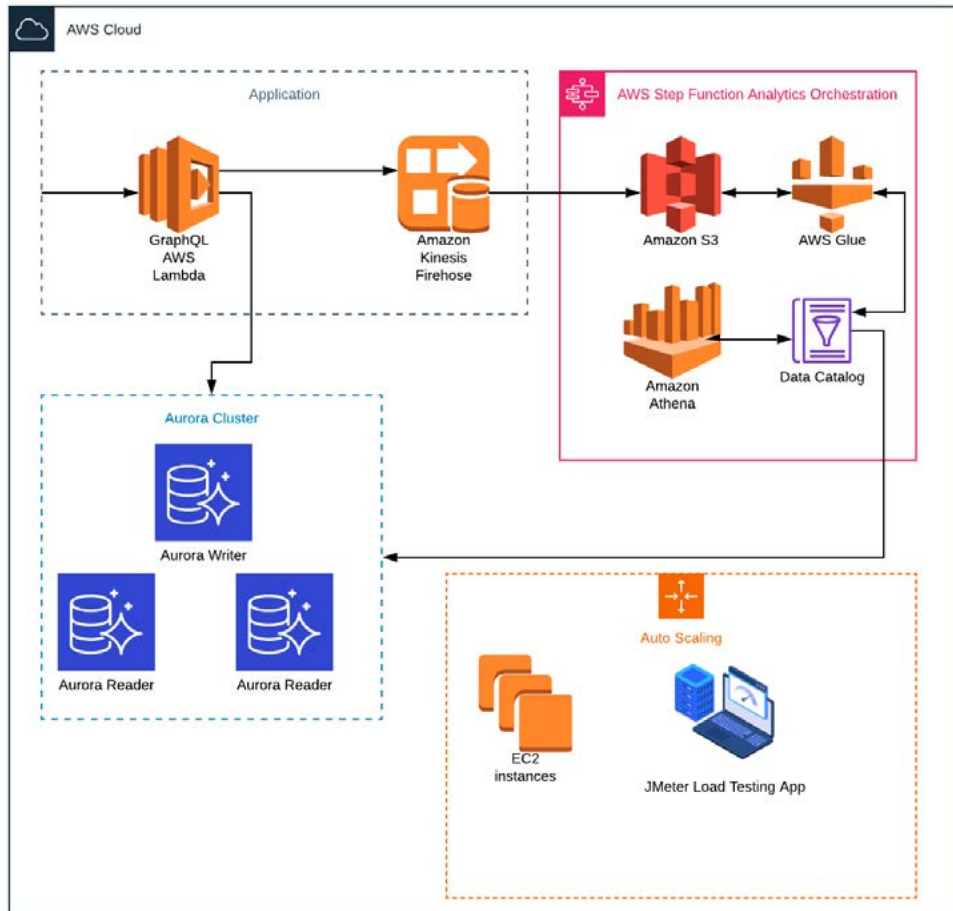
Benefits

ClearScale's innovative solution generated serverless analytics and data pipelines for the customer that have reduced its administrative costs for data management, which require fewer employees to maintain. Data processing is faster, analytics are far more granular and beneficial, and the overall process is more secure and reliable.

The solution also has repercussions for data management in general. ClearScale's creative approach is spurring discussions on and investigations into the use of AWS Step Functions for building "better data pipelines" — proving once again that ClearScale is truly at the forefront of the Big Data and app development industries.

The use of Step Functions allowed to significantly reduce operational costs of the solution. The overall serverless set up also plays a large role in the cost-reduction.