clearscale

aws partner network

Premier Consulting Partner

Big Data Solution Transformation with Custom API and Serverless Architecture

mediacompany

When companies have important data scattered throughout different data centers that is used to drive key business operations, it often proves challenging to aggregate the data in a way that allows all parts of the business to quickly and efficiently make decisions with the most up-to-date information. This can be very detrimental to an organization as it attempts to rise to the challenge of meeting competitors head-on, getting key metrics around business processes or health, or deliver content to a broad audience.

The Challenge

A recent client of ClearScale, an AWS Premier Consulting Partner, struggled with this exact challenge. They asked for an evaluation of what they had and requested suggestions on where ClearScale could take them utilizing AWS Services. Stored in two different data centers located in two separate geographical locales, the information warehoused in these locations was hosted within data sources such as Drupal, WordPress, and others. To get vital information out the data had to be aggregated, a time- and resource-consuming operation, so that applications could pull relevant data sets which ultimately resulted in latency and data integrity issues. This negatively impacted the client from a performance perspective as well as provided unreliability in accessing the data. When the company decided that their subsidiary locations needed to use the same data set, they discovered that the processes they had in place would not allow them to quickly aggregate nor deliver that data without significant business or technical impact.

Not having the AWS experience necessary to find a suitable alternative, they engaged ClearScale an AWS Premier Consulting Partner, for a viable solution. They knew that to get data consistently to their subsidiaries they would need to develop a global API that could be leveraged to allow these groups to pull data directly. Another challenge was to get the data consolidated in one location, but also to make certain that the API had access controls in place with a goal of preventing the exposure of the data these subsidiaries should not have access to.

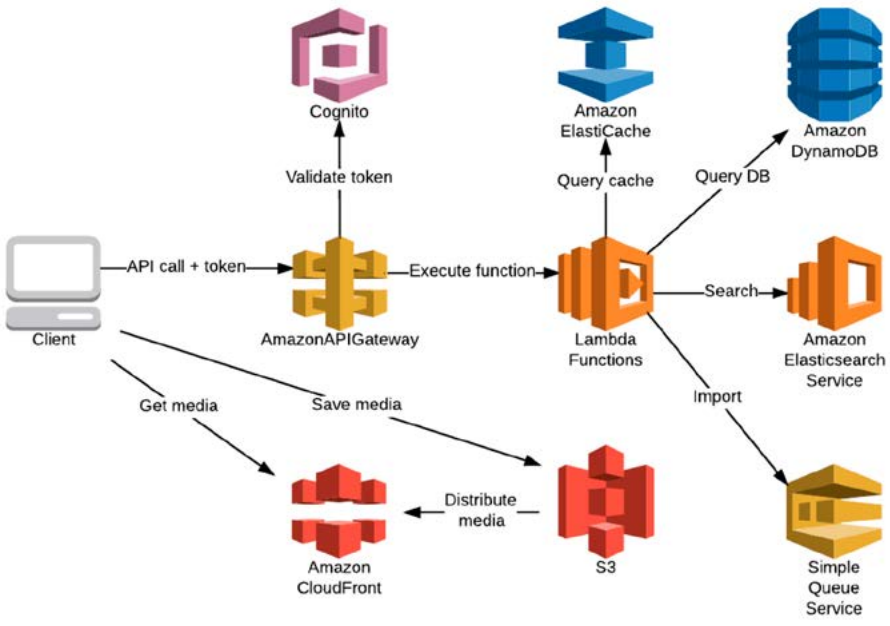
The ClearScale Solution

ClearScale immediately recognized the opportunity to migrate the customer into a serverless architecture environment within the AWS Cloud. ClearScale evaluated the specific requirements the client had concerning the need to create a global API that would allow all their subsidiaries and various applications to access the data reliably. To execute successfully on this deliverable, ClearScale worked with the client to understand what precise data they would need to extract from the data warehouse, that would then need to be created in AWS. Once identified, an API was designed, built, and deployed into AWS API Gateway which leveraged Lambda functions. Using node.js, the API was built in such a way that separate calls could be made to focus on a key set of data querying functions and each set of calls deployed individually within the API had their own set of deploy and caching protocols using AWS ElastiCache.

To validate the data being requested by a given entity, the API was integrated with API Keys and Cognito tokens. This not only allowed for data security from a read perspective, but also from a write perspective; without the proper API Key from the client, data could not be altered or deleted, thus further protecting data integrity. Since most of the queries made through the API were expected to be read queries, the API and the database had to be designed to handle a million+ requests a day. The write queries were only expected to have up to 10,000 queries a day. Every response given was subsidiary-independent so multiple calls could be made in conjunction with the confidence and knowledge that the response would deliver the required information to the appropriate requestor without data intermingling with other requests.

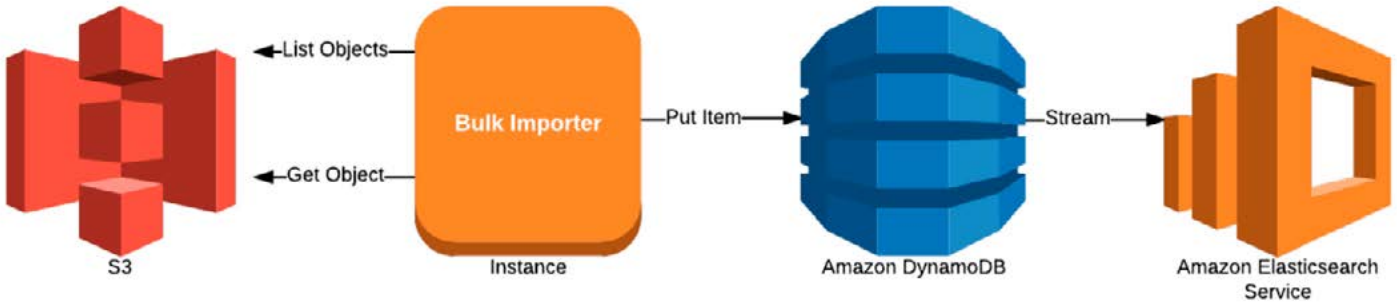
To further optimize the latency between the API request, the data query on the database and the return response, the tables within the database had to be configured so that the most heavily used tables had a higher read capacity value, with tables that were not as heavily used with a lower read capacity. This allowed for improved performance and scalability that would be needed in the future.

API Request Flow Diagram



Once the API had been completed and the requirements gathered for what was needed from a database perspective, ClearScale began the second phase of the project by setting up a DynamoDB NoSQL service and integrating it with Lambda Services, SQS, API Gateway, and S3 buckets for the client's account and a custom EC2 instance that handled the initial bulk data import. ClearScale performed a mass-migration of data from all the two on-premise data centers the client had established. The migration was a massive undertaking with approximately 250 million records transferred into the AWS S3 buckets, including 50+ TB of image data and associated database records into DynamoDB.

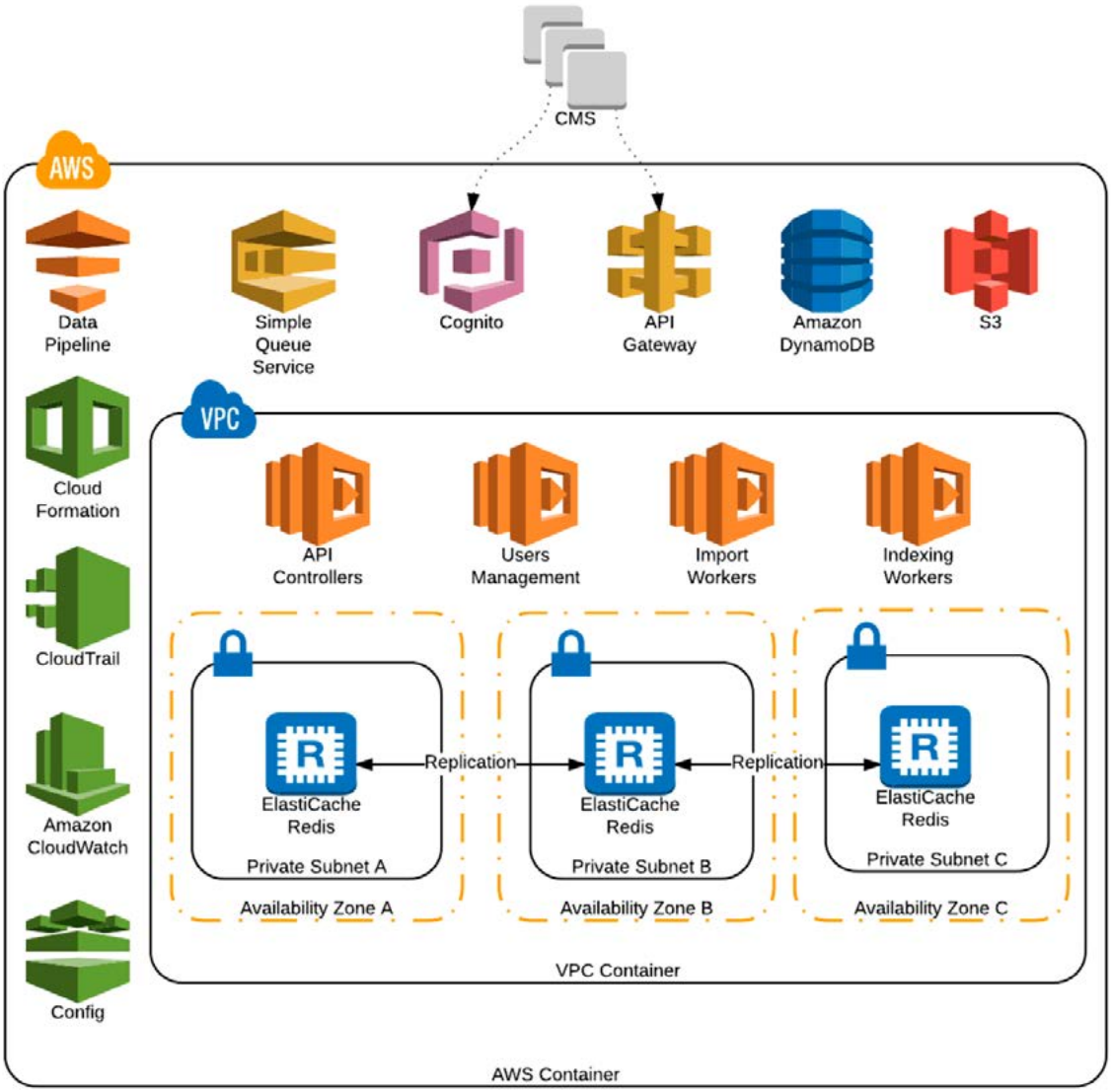
Initial Bulk Data Import Diagram



As information was inserted into DynamoDB, AWS ElasticSearch was leveraged to index each record. Combined with the API Gateway caching which was implemented earlier, this allowed for large volumes of data to be accessed, retrieved, and updated with low latency and increased speed. Once implemented and tested, ClearScale could demonstrate that between DynamoDB, ElastiCache, SQS, and API Gateway configurations, the setup was able to process one million reads and ten thousand writes per day, a clear improvement compared to the client's previous data center implementation.

Monitored by CloudWatch and CloudTrail and protected by AWS Identity Access Management (IAM), the resulting migration situated the formerly fragmented data sources into one cohesive serverless location in the cloud and was configured and optimized in such a way as to improve overall performance and reduced latency while providing a robust data redundancy and reliability.

Logical Architecture Diagram



The Benefits

ClearScale delivered an agreed-upon set of global APIs that the client could configure for each subsidiary's specific needs using client-defined API Keys. Additional training on how to maintain the database and ancillary AWS services was also provided. The end result was substantially improved backend services with highly deployable global APIs based on the client's needs. Not only was performance improved through the consolidation and aggregation of data from disparate data centers into one centralized serverless environment, but the reliability of obtaining data was immediately felt by those subsidiaries using the APIs created through the ClearScale/client partnership.