

CASE STUDY: IMAGINE COMMUNICATIONS

LEADING TELEVISION DIGITAL MARKETING AGENCY

Data Lake / Data Warehouse on Amazon Web Services

OVERVIEW

Imagine Communications is a Texas-based leading Television Broadcast Media company. Imagine Communications makes it possible for their clients to create digital content anywhere in the world and on any device or platform, but as the way people watch TV evolves and continues to change, media companies must maintain nimbleness in order to monetize their data.

THE CHALLENGE

The Imagine Communications Research and Development data team challenge is twofold: they first had to reduce prohibitive overhead and operational costs associated with developing a new product on Azure using Microsoft SQL Server, and deliver a new analytic platform built to leverage their existing analytic tools, provide instant elasticity, integrate multiple data sources, and increase time to deliver new products and services.

This analytics system transforms TV ratings data provided by Nielsen and ComScore spanning more than 200 geographic markets in a single serverless Data Lake.



THE SOLUTION

After a thorough consultation and evaluation, the Beyondsoft Big Data team proposed a solution built on AWS Big Data cloud native services, comprising of Athena, Glue, Redshift and Redshift Spectrum. The project also included the easy deployment of Redshift clusters and event-driven pipelines which allowed Imagine Communications to move large amounts of daily generated data to a Redshift cluster.

A tuning and revision of the former workflow was also recommended and executed. Some of the changes consisted of:

- Historical data and real-time data provided by Nielsen and ComScore, the rating agencies, is now stored in S3, consisting of more than 1 million files and more than 50 TB of uncompressed data.
- The input data comes in mainframe like format, so a Lambda/Python function was used to convert to S3 in Json format. Lambda functions were used for high concurrency data processing for S3 files and were invoked more than 1 million times to process more than 800,000 files in one hour.
- The S3/JSON data was converted to Parquet format using Spark which runs in Fargate containers. This process was done as a pure cost saving measure.
- The Glue Catalog has more than 2,000 tables that is queried by Athena.
- The Parquet formatted data stored in S3 is moved to Redshift Spectrum and Redshift Clusters via staging tables.
- The Redshift cluster, which is a 2 Node RA3 configuration, houses more than 20TB of raw data which is 4 years.
- The ETL into the Redshift cluster is done by Python/SQL using Fargate containers.
- The Reporting UI now communicates to the Redshift spectrum and Redshift cluster through C# API.

THE SOLUTION SUMMATION

By implementing the AWS managed services stack and architectural advancements, the Beyondsoft Big Data team successfully delivered a scalable, truly elastic, and highly performant data solution that also met Imagine Communications' stringent success criteria. Additionally, the overall cost to operate the new solution saw a drastic reduction from an average of \$200,000 month to month operating cost on another Cloud platform to an average of \$5,000 month on the AWS Cloud platform.

KNOWLEDGE TRANSFER

During the engagement with the Imagine Communications, the Beyondsoft team also provided education and training so that the Imagine Communications Data Analytics Team could take ownership of the newly developed environment. As a practice, the Beyondsoft Big Data Team developed and delivered a runbook so that Imagine Communications could, not only manage the newly provided data platform, but also add to the solution moving in the future.

BENEFITS

Migrating from Microsoft SQL Server on Azure to Redshift on Amazon Web Services provides Imagine Communications with increased performance, high scalability, and faster data processing by leveraging a sophisticated data lake solution.

TECHNOLOGY USED

Amazon Web Services

- Redshift • Spectrum • Athena • Glue • Catalog S3 • Lambda • Fargate • Apache • Spark • Parameter Store • Python

Beyondsoft Consulting Solutions

- ConvergDB

ARCHITECTURE

