GCP to AWS Migration Case Study for Quikr

aws partner network

Executive Summary

Introduction

Quikr is a prominent online marketplace founded in 2008, headquartered in Mumbai, India. As one of the largest classified platforms in the country, Quikr offers a wide range of products and services for individuals and businesses to buy and sell. Users can conveniently access Quikr's platform through desktop, laptop, and mobile devices, enabling them to post ads across various categories including real estate, vehicles, electronics, jobs, and more. With its user-friendly interface and extensive reach, Quikr has transformed the online buying and selling experience in India, connecting millions of users, and facilitating seamless transactions.

Customer Challenges

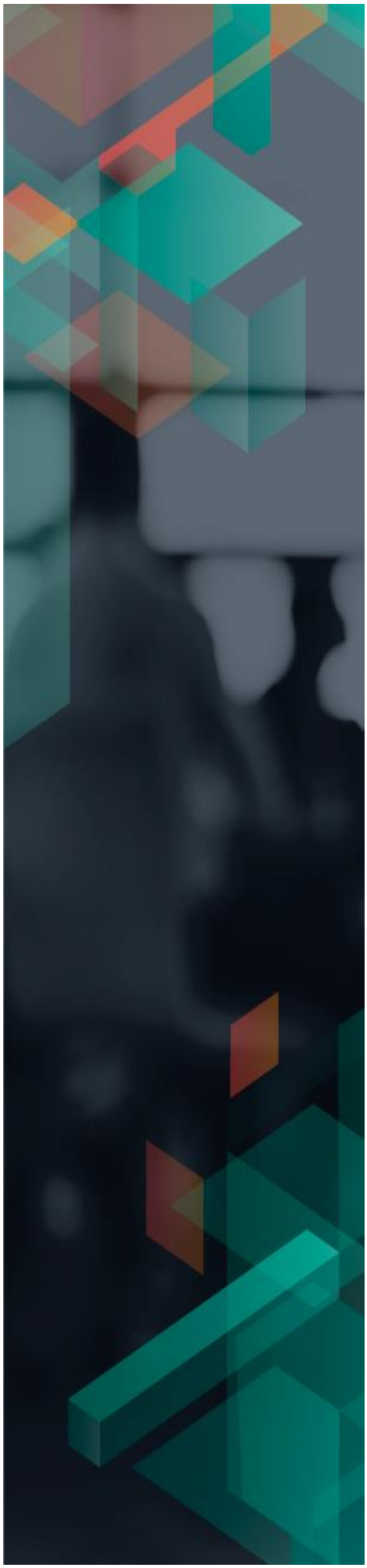
Quikr aimed to enhance platform availability, lower costs, and meet the evolving requirements of their vast user base of over 30 million unique users. The migration process involved migrating a massive data set consisting of 47,965 BigQuery tables and 82 Google Cloud Storage (GCS) buckets, totaling 200 TB, from Google Cloud Platform (GCP) to Amazon Web Services (AWS). The conversion of BigQuery data to Parquet format for improved analytics posed a challenge. Throughout the migration process, Quikr aimed to minimize disruptions to their online marketplace platform while ensuring the secure and reliable transfer of data.

Solution

- Amazon EMR clusters were used to migrate the data from GCP BigQuery to Amazon S3
- The Amazon EMR cluster was set up on VPC, which includes spot instances to run task nodes. Amazon EMR has default functionality for scheduling YARN jobs so that running jobs do not fail when task nodes running on Spot Instances are terminated. Amazon EMR allows application master processes to run only on core nodes.
- We utilized spot Amazon EC2 instances with Amazon EMR to reduce the cost by up to 90% compared to the on-demand EC2 instances.
- We have generated CSVs that list 47,965 BigQuery tables that total 150 TB of Analytics data from GCP BigQuery to Amazon S3.
- For GCS Coldline data migration, we have used the EMR cluster to migrate data from 82 buckets of 50TB data into a single Amazon S3 bucket per client requirements.
- Custom Python Scripts kept on Amazon EMR Notebooks, have verified rows, columns, and Schema of all the tables before and after migration
- After migration, we have created AWS Glue with Terraform for each database which will generate a table automatically configure the 8000+ AWS Glue Crawlers to create the databases and tables for Amazon Athena.

About Quikr

Quikr is an online marketplace that allows individuals and businesses to buy and sell products and services. It was launched in 2008 and has since become one of the largest classifieds platforms in India. Quikr operates leading transaction marketplaces built on top of India's largest classifieds platform for online buying and rentals, of which over 30 million unique users are used monthly.



- Amazon S3 is used for all the sales orders, inventory, and trends data to be uploaded from multiple online locations as batch updates for staging.
- Utilized Amazon Glue for availability, durability, and scalability of processing for ETL.
- Processed data from Glue jobs are loaded into Amazon Redshift, which is further used for reporting, business apps, and business intelligence by the client.

Third-party applications or solutions used

- A custom PySpark file that will take parameters such as source and destination to migrate data from one place to another was created.
- A python script was created to pass the values from CSV and assign PySpark Jobs accordingly to the Amazon EMR cluster to migrate the table on Amazon S3 to a specific destination.
- We have DistCp (distributed copy) to migrate data from GCS to S3, and that was assigned using Python script to automate the job assignment

Architecture Diagram

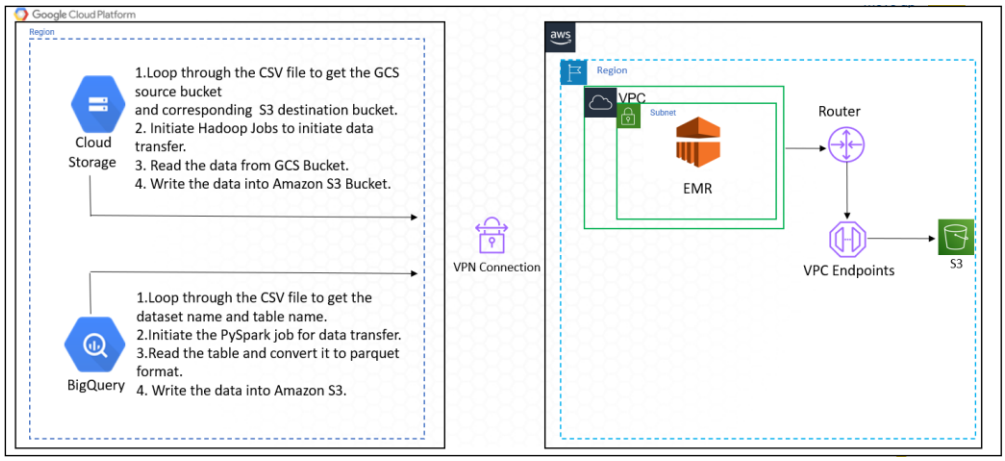
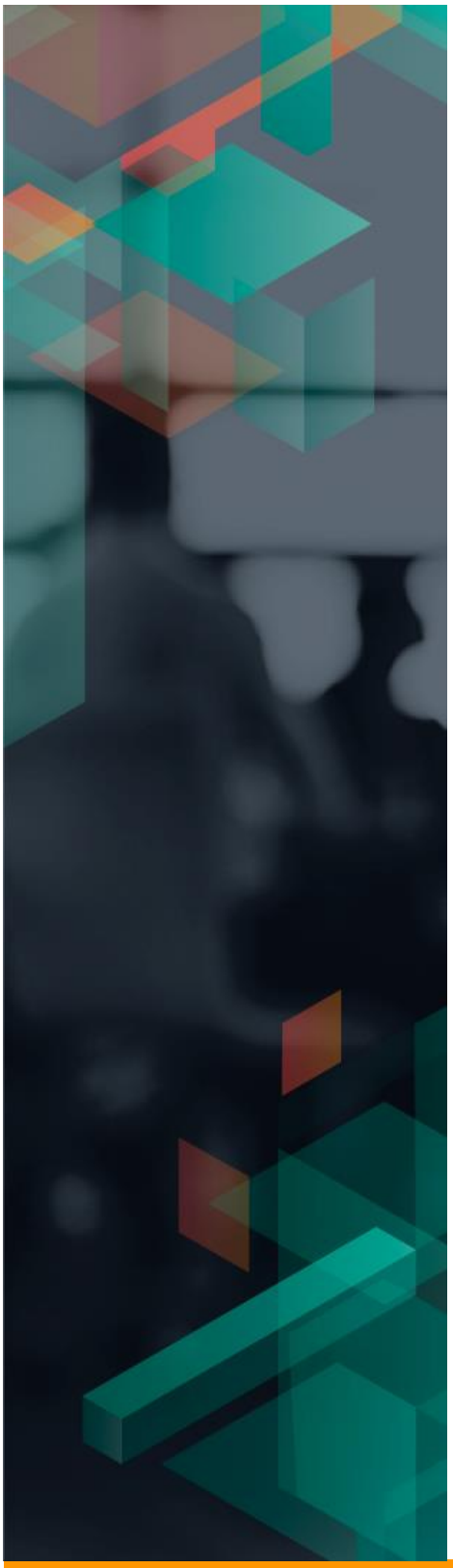


Fig 1: Migration from GCP to AWS

The source and destination are confirmed in CSV files. They are categorized based on size, estimated time, and cluster size for different categories. A Python script runs in the Amazon EMR machine, and then assigns a custom PySpark job that migrates data from BigQuery (source) and S3(destination) from CSV files. The DistCp tool transfers data from GCS to Amazon S3 on the Amazon EMR cluster. All the jobs assigned to Amazon EMR are automated using Python.



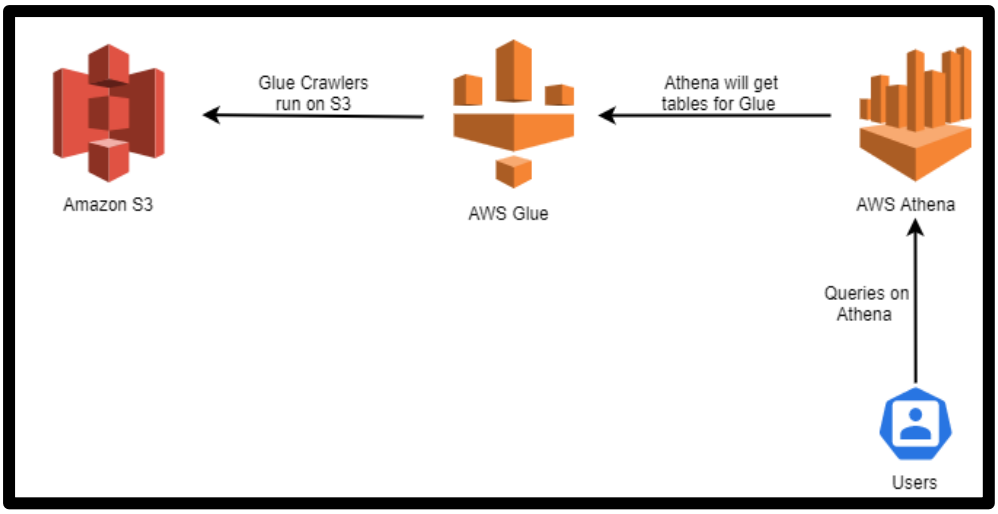


Fig 2: Query Execution on Amazon Athena

The data is stored on Amazon S3 and created AWS Glue crawlers according to the Datasets. All the AWS Glue crawlers were running on Amazon S3 and created tables accordingly which was then queried by Amazon Athena.

Conclusion

- The migration of data from GCP BigQuery and GCS buckets to Amazon S3 was completed in 5 days without any issues.
- All data files were stored on Amazon S3 in a way that would facilitate the creation of a Glue crawler on top of it.
- Over 8000 Glue crawlers were created, which contained a total of 47,965 tables.
- Amazon Athena is used to query the data for the purpose of post-verification, specifically to obtain the number of rows and columns.
- Apache Parquet, known for its efficiency and performance in flat columnar data storage, was used to convert the data into a cost-effective format for storing and querying on Amazon S3 using Amazon Athena.
- The verification of data files in AWS was done within five days of the completion of the migration process.
- The client has transitioned to AWS services for their analytics workloads and is utilizing Amazon Athena for the same purpose.