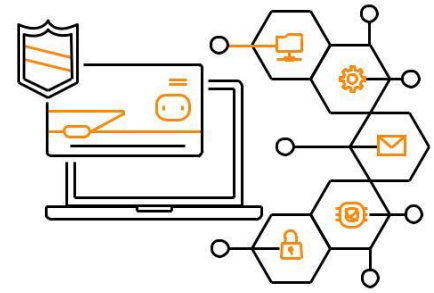# On-premises Big Data to AWS cloud migration to drive cost efficiency to drive
Case Study

Executive Summary

Rakuten Rewards migrated their on-premises Hadoop clusters to Amazon Web Services (AWS) using a combination of services like Amazon S3, Amazon Elastic Map Reduce (EMR), and Snowflake to achieve cost reduction and operational efficiency. AWS was selected after rigorous evaluation criteria which included a proof of concept. After successful data migration by Cloudwick, Rakuten has saved on operational costs and added new analytics capabilities like machine learning with data available in a data-lake for more complex analytics.

Customer Challenge

Rakuten was using on-premises Hadoop clusters which were extremely complex to manage and operate. They were running all data processing applications with data delivered to end users on Cloudera Impala. This resulted in less elasticity and more maintenance.

Also, the data volume increased over time and Hadoop clusters couldn't manage the scale accordingly. Operations engineers were spending most of their time in troubleshooting hardware and services failures. As a data driven company, timely processing of data and analytics is key to success. Rakuten needed highly scalable managed services to process the high volume of data rather than maintaining legacy Hadoop clusters.

Why AWS?

Rakuten considered multiple cloud vendors during the selection process. One of the determining factors in favor of AWS was the low latency connectivity between AWS and the on-premises clusters. During the POC, Rakuten achieved several times better data transfer speed between on-premises and AWS using direct connect connections. Moreover, migration of the existing Hadoop workloads to Amazon EMR is easier than the other cloud offerings.

About Rakuten Rewards

Rakuten Rewards, formerly known as Ebates, is a cash back and shopping rewards company headquartered in San Francisco Bay Area, CA.

Rakuten Rewards' revenue comes from affiliate network links. Members of the site click through affiliate links before shopping at a retailer's site.

Once the member makes a purchase, Rakuten Rewards receives an affiliate commission from the retailer which is then shared with the member.
.

Why did Rakuten select Cloudwick as AWS partner?

AWS recommended Cloudwick to Rakuten based on a successful similar project implementation. Cloudwick was able to complete a POC within 4 weeks. The POC included transferring 15TBs of data from on-premises Hadoop clusters to Amazon S3. Cloudwick also helped in resolving multiple Hadoop production issues during POC. Cloudwick's expertise in both Hadoop and AWS was the key to getting selected as an AWS partner of choice.

AWS and Cloudwick Solution

AWS Services used: Amazon S3, Snowflake, AWS Direct Connect, AWS Identity & Access Management (IAM), AWS CloudFormation, AWS CloudTrail, Amazon Elastic Container Services (ECS) and Amazon Virtual Private Cloud (VPC).

Cloudwick migrated Rakuten data warehouse workloads from their on-premises Cloudera Hadoop cluster to AWS using a combination of services like Amazon S3, AWS EMR, and Snowflake. The data marts were created using Snowflake on AWS. The on-premises enterprise Hadoop cluster consisted of roughly 100 TB of total storage. Cloudwick implemented a highly scalable data ingestion and data processing solution using AWS EMR and an orchestration layer using Amazon ECS so Rakuten can easily extend this solution to bring more data directly to AWS data lake.

Cloudwick team built a job orchestration platform to ingest data (original data store layer), process data (data warehouse layer) and store in data warehouse. This involved near real time jobs and batch jobs. The data ingestion layer ingested data from various sources like PostgreSQL, MySQL, Apache Kafka etc., to Snowflake on AWS. The data processing layer read data from snowflake an then it was processed and written to a final data warehouse layer. The data processing job types included Scala, Python, SQL etc. The *job orchestration framework* was designed keeping ease of usability and scalability in mind. Dockerizing the framework and deploying it on Amazon Elastic Kubernetes Service (EKS) allowed for graceful scaling. A new user to the system could schedule and configure a job without deeply getting involved in the underlying structure. A Scala *job migration framework* was also designed for real time user data pulls (login data, order transactions etc.,) and ingest to Snowflake. The framework supported various sources like PostgresSQL, MySQL, Apache Kafka etc., and target as Snowflake and Apache Kafka. Also, there were other jobs which read data from Snowflake, heavy data transformations, combine multiple tables and write data to Snowflake. Approximately 600 jobs were migrated and now run in the cloud using the job orchestration framework.

Customer Outcome

Cloudwick built a new framework for Rakuten where adding new workloads to the new data ecosystem was an easy process. Now data analytics team can generate reports faster with on-demand scalability. The IT department is no longer worried about Hadoop cluster management or troubleshooting. The migration allowed Rakuten to start leveraging other types of analytics such as machine learning on AWS as the data is already present in the data lake.

About Cloudwick

Cloudwick is an AWS certified Advanced Consulting Partner that specializes in building native data lakes that power faster, cheaper and more agile cloud analytics for IT, business users and data scientists.

Whether you need to migrate your analytics to the cloud, add decision automation to your business intelligence, improve customer experience with machine learning or want to build or buy a data lake Cloudwick has a complete portfolio of services and solutions.

Cloudwick