motherson

SAMRX

SAMRX Data Lake using Amazon EMR & various other AWS services

About SAMRX

Samvardhana Motherson Hamakyorex Engineered Logistics Limited (SAMRX), a joint venture company between Samvardhana Motherson International Ltd. (SAMIL), India and Hamakyorex Co. Ltd., Japan. SAMRX aims to revolutionize the transportation of finished vehicles in India by developing a modern, technology enabled and equitable socially conscious solution for the OEM customers. Building on the expertise of Hamakyorex and Motherson Group, SAMRX will work towards considerably scaling business operations over a period of next 3 years.

About Challenge

SAMRX uses currently 20 self-owned trucks and 42 drivers for the transportation of finished vehicles in India by developing a modern, technology enabled and equitable socially conscious solution for the OEM customers. The company plans to scale up its fleet of trucks to 50 by next quarter and have plan of over 1,500 and employ 3,200 drivers by FY 2023. Moreover, number of sensors are also going to be increase. Since Volume, Velocity and Variety of data will increase over the time need to have efficient & cost-effective solution for analyzing of various sensor data to gain business insights from data available for improving efficiency and cost-optimization.

We needed a cost-effective solution that will reduce the complexity associated with the process of analyzing the data received every day as Cassandra being used as database gets being choked for frequent read operations.

AWS Services used

Amazon EMR

Amazon S3

AWS Step Function

AWS Lake Formation

AWS Lambda

AWS Glue

Amazon Athena

Amazon QuickSight

Proposed Solution

We researched on various solution options e.g. Debezium Tool, Data Pipeline etc. but due to various constraints e.g. Data Pipeline was not available in Mumbai Region, we finalized following approach where we used Spark Cassandra Connector to access data from Cassandra DB.

We proposed an automated serverless solution on AWS to address this requirement and performed following steps for implementation of solution

- One-time migration of initial data of given tables to Amazon S3 by running Spark Jobs on EMR Cluster. We used Spark Cassandra Connector for getting data from Cassandra DB.
- In the Spark job itself converted the data format to Parquet as being columnar storage it gave huge cost benefit later when querying the data stored in data lake
- Implemented Step functions to sequence AWS Lambda and SNS, Lambda will be used to invoke EMR Cluster to run Spark Jobs for Incremental Data Transfer on daily basis.
- Used Spot instances for Core Nodes in EMR to save cost
- Configured Glue Jobs to run after incremental data transfer on daily basis to load partitions on regular basis
- Used AWS Lake Formation to set up and secure Data Lake of S3 buckets.
- Created various Dashboards using Amazon Quick Sight that helped the customer to gain business in sights for improving efficient and cost optimization.

Solution Outcome

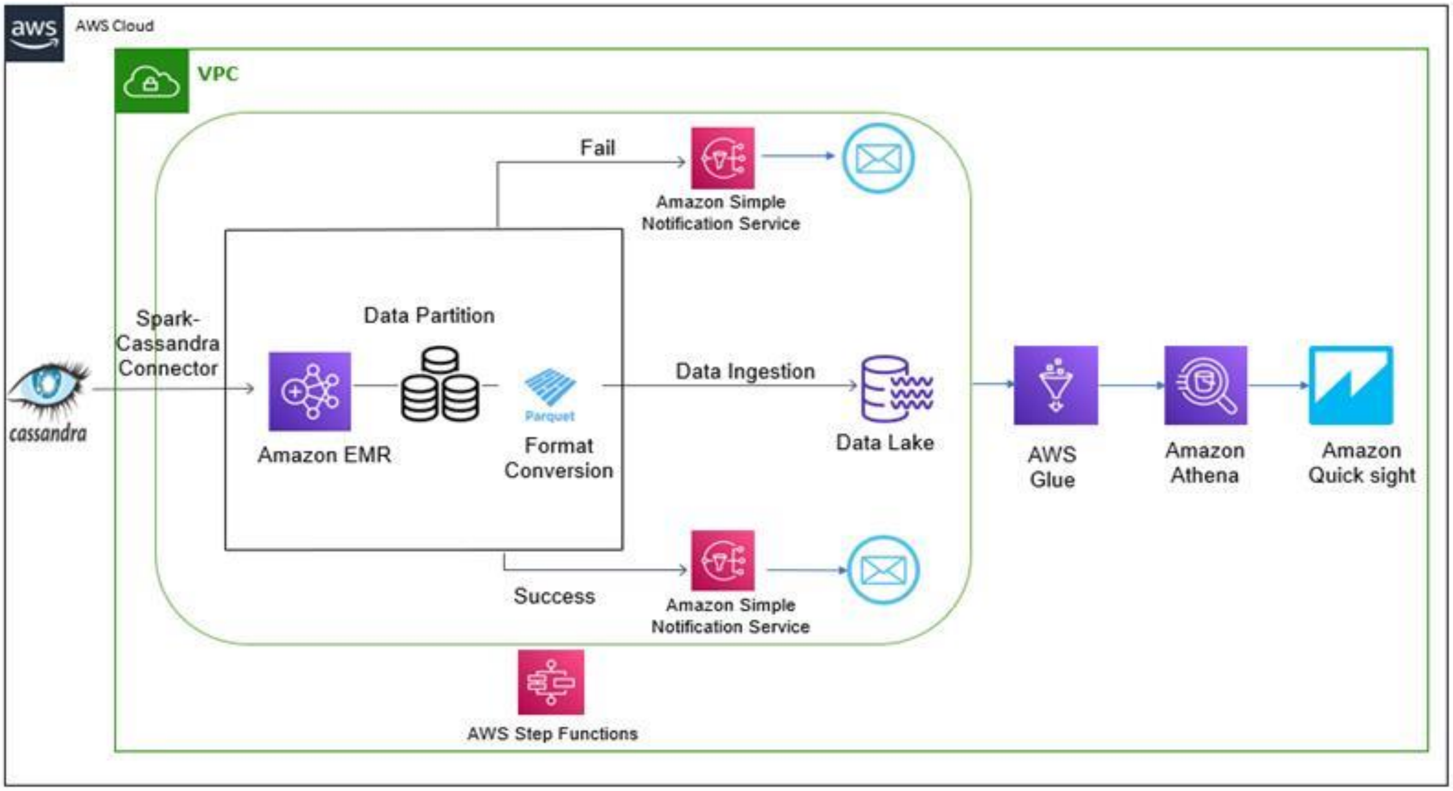
Data Lake on AWS was created successfully, and we migrated raw data of all sensors both one time and incremental on daily basis. Scheduled a Job that transfers required data on daily basis. Following objectives were met

⏱ Analysis/Diagnostics of all available data in cost effective and timely manner. Approximate reduction of more than 70% processing time for analyzing the information comparing to earlier operations being directly performed on Cassandra Database.

💵 Created various dashboard giving insights to customer where cost optimizations can be done.

🎯 The biggest pain area of customer to identify period for which any sensor(s) was stopped or didn’t sent data, easily analyzed through data lake.

Architecture Diagram



How AWS services helped in the creating the data lake

Amazon S3 to store files in different stages

It is an object storage service that offers industry-leading scalability, data availability, security, and performance. In this solution it helped to store the partitioned data in parquet format.

AWS Lambda to run code serverless for EMR cluster setup and notifications

It let us run code without provisioning or managing servers. Lambda allowed us to run code for virtually any type of application or backend service - all with zero administration. Here, the lambda function was used to set up an EMR cluster which ensured the daily ingestion and partition of required data. As well as sending data to SNS according to the status of EMR cluster.

Amazon EMR to run Spark Jobs

It is an industry-leading cloud big data platform for processing vast amounts of data using open source tools such as Apache Spark, it helped in automating the time-consuming task of creating and provisioning the cluster. This cluster then executed a Spark script which partitioned the data and saved it in parquet format in S3 bucket.

AWS Lake Formation for creating Lake Formation

It is a centralized, curated, and secured repository that stores all your data, both in its original form and prepared for analysis. It helps in loading data from diverse source, configuring access control settings, granting access to data sets, and auditing access over time. We can then leverage these data sets with our choice of analytics and machine learning service like AWS Glue and AWS Athena.

AWS Step Function to orchestrate Lambda and SNS

It is a serverless function orchestrator that makes it easy to sequence AWS Lambda functions and multiple AWS services. We used step function to orchestrate the flow of starting the EMR cluster and then checking its status periodically to check the completion or failure of the job and notify the user accordingly using the AWS SNS service.

Amazon SNS for e-mail altering regarding EMR Cluster Status

Amazon Simple Notification Service (Amazon SNS) is a fully managed messaging service for both application-to-application (A2A) and application-to-person (A2P) communication.

AWS Glue for cataloging

It is a serverless ETL service that crawls your data, builds a data catalog, performs data preparation. The crawlers offered by AWS glue were used to create a data catalog and figure out the schema of the data received from Cassandra Db. The schema was used to query the data.

Amazon Athena for Querying Data Lake

Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL. Athena is serverless, so there is no infrastructure to manage, and you pay only for the queries that you run.

Amazon QuickSight for creating Dashboards

Amazon QuickSight is a scalable, serverless, embeddable, machine learning-powered business intelligence (BI) service built for the cloud. QuickSight lets you easily create and publish interactive BI dashboards that include Machine Learning-powered insights. QuickSight dashboards can be accessed from any device, and seamlessly embedded into your applications, portals, and websites.

About the Partner

MothersonSumi INfotech &Designs Ltd.

MothersonSumi INfotech & Designs Limited (MIND), a SEI CMMI Level 5 IT services company and the IT back bone of Motherson group. MIND is a trusted technology partner to over 200 clients globally. Our value proposition is in our strength in specific Industry segments and years of experience in the areas of intelligent warehousing, Supply chain enablement, software application development, smart ERP customization, infra managed services, cloud, IoT & Analytics. MIND is serving customers in 41+ countries with a strong team of 1500+ professionals.