Data Lake powered BI solution for Better Decision Making

- a case study

Client -

A Public Transportation Agency in the US, operational from the last 30 years, across 80 routes, 500 fleet, having an annual ridership of 2.5 billion and generating terabytes (TBs) of data on a daily level.

The "Old Bad" Days for Client -

- Multiple software systems to supplement various operations and departments.
- Consolidating all the data at a common place to have a holistic view of the business.
- Lots of manual effort and chances of data inconsistencies.

Scope of Work -

- To create a robust and scalable Business Intelligence/Analytics (BI) solution, powered by a cloud-based data lake.
- The BI solution, capable of running across multiple departments, and present different KPIs according to different user roles.
- The lake, capable of consuming a large amount of data generated from different sources in near real-time, empowering the real-time dashboards built on top of it, for KPI monitoring and decision-making.
- Data lake, to be supplemented with various individual data marts for their respective departments' data segregation and their respective dashboards.

Challenges Involved -

Since the operations were being run since many years and also at a very large scale, the client had deployed multiple legacies and new systems were put in place to handle the various processes internally. As a result, the following challenges, of a very critical scale, were observed while implementing the complete solution -

- Different systems generated different types and formats of data - at different places.
- Compatibility issues for the top management to make sense out of all the data at a combined level, hindering decision-making, overshadowing the opportunities to improve further on.
- Client IT team not as updated with tech advancements, to provide access to different sources keeping in mind the updated frequency of data.
- With the PCI data security protocol in place, data access a major difficulty.
- Keeping the data lake and corresponding data marts consistent and updated with new data.
- Efficient data models to store large volumes of data in an optimized format.
- Keeping the overall solution cost-effective for the client.

Solution Approach -

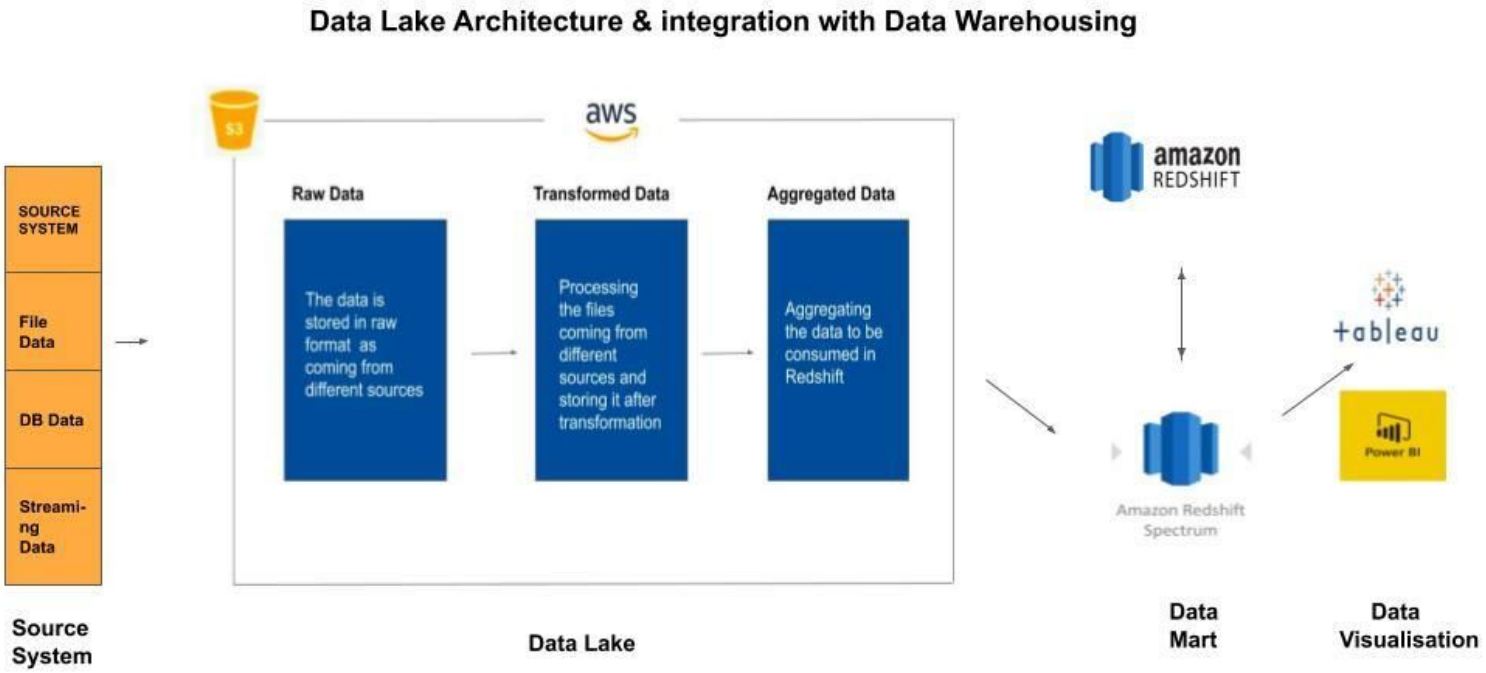
The complete pipeline for creating a data lake and using it to create data marts and respective dashboards was done as per the steps listed here:

- Understanding the individual systems, generating data in different formats, storage structure and the frequencies.
- Narrowing down to a single format for individual data streams to be captured from the client, keeping in mind the one-time load, incremental load and Change Data Capture (CDC).
- Understanding the type of KPIs to be visualised for each department.

- Finalize data lake architecture, and setting up individual data pipelines (ETL), keeping in mind the variety, velocity and veracity of data.

Technical Architecture -

Due to on-cloud deployment, Amazon Web Services (AWS) was the most obvious choice for creating the lake, considering the supplementary services provided by AWS.
Following is the most optimized architecture of the data lake created:



The "Good New" Days/Result -

Cloud-based data lake BI solution helped in: data management as well as in achieving the following objectives -

- Time for Data Analysis and Decision Making reduced from days to minutes.
- Automated orchestration of data from disparate sources - to eliminate manual intervention.
- Analytics delivered to business users with extensive drill-down options for deeper analysis
- More timely, accurate and less laborious access to high-value reporting and KPIs