Healthcare Data Analysis

The project aims to build a single source of true data storage for large healthcare datasets using Spark and S3. Some dashboards are also made in this project for visualization.

Tech Stack

- Data lake: Amazon S3

- Data source: PostgreSQL

- Data read storage: MySQL on Amazon RDS

- Processing layer: Apache Spark on EMR

- Visualization: Power BI

Architecture

The architecture of this project is presented as follows:

- Data is sourced from PostgreSQL and ingested into the *raw zone* of Data Lake hosted on S3.
- Raw data is cleansed and standardized before moving to the *cleansed zone*.
- Cleansed data is transformed into reportable form and loaded into the *curated zone*.
- Publish data from the *curated zone* to Data read storage for higher performance reports when connected from the BI Tool.
- Reports are created in Power BI from the data in MySQL.

Data Source

The source of raw data is from
https://data.cms.gov/provider-summary-by-type-of-service

The data used is Medicare Part D.
- Data source in PostgreSQL has 4 tables, total size around 10 GB:
  - Prescriber_drug: ~ 25M rows
  - Prescriber: ~ 1.1M rows
  - Drug: ~115K rows
  - State: ~30K rows

Visualization

We have performed visualization using POWERBI. The folder contains folder name visualization with respective visualization files

Achievement in learning

Apache Spark

- Components of Spark and how Spark works.
- How to adjust resources (RAM, CPU, instances,...) for optimizing Spark performance and costs.
- Tuning Spark application by using partition
- Use Spark to implement a full data pipeline.
- Fundamental of how to write Spark correctly.
- Manage Jar files for JDBC connection

Project set up

- Implement logging and log files to track the Spark application
- Test the project on local mode before running on the cluster.

AWS

- Set up EMR for Spark
- Track the resource utilization in EMR
