IST 615 – Cloud Management – Project Report

NYC Taxi Trip Analysis

Ankita Vartak

Hrishi Telang

Nivesh Vaze

Table of contents

[Project objective 2](#_Toc120998579)

[Configuration of services used 2](#_Toc120998580)

[Data description 5](#_Toc120998581)

[Business Requirements 5](#_Toc120998582)

[Solution Architect 6](#_Toc120998583)

[Tasks Completed 6](#_Toc120998584)

# Project objective

The idea behind this project was to build a data pipeline to extract and transform NYC taxi data (trip, payment type, calendar, vendor, zone) and create a cleaned and aggregated dataset to build visualizations.

We leveraged **Azure Data Lake Storage Gen2** to configure container and store the raw data files, utilized built-in **Serverless SQL Pool** in Azure Synapse Analytics to extract and transform the data, and used **Power BI** embedded to integrate and build dashboard.

# Configuration of services used

1. Storage account – **synapse615dl**

Configured Azure Data Lake Gen2 storage account and created container nyc-taxi-data to store above mentioned raw data files.

Graphical user interface, text, application

Description automatically generated

1. Configured Synapse Analytics workspace – **synapse-project615-ws**

Configured Synapse Analytics workspace to leverage the built-in serverless SQL pool and develop SQL scripts for building data pipelines.

Graphical user interface, text, application, email

Description automatically generated

1. Configured the Linked Service within Synapse Analytics and integrated Power BI embedded to build report for NYC Taxi dataset.

Graphical user interface, text, application

Description automatically generated

# Data description

Data source - <https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page>

All the files are in different formats (CSV, TSV, and JSON). The objective is to blend the data together from different file formats and aggregate the data for analysis.

Graphical user interface, diagram

Description automatically generated

Trip Data – Single record per ride consisting of location, fare and passenger details.

Calendar – Consist of date key to decode the day and week number of each date

Trip Type – whether the trip type is Dispatch or Streep-hail

Payment Type – Whether the amount is paid in cash or card

Rate Code – Used in calculating the fare based on the location of the ride booked

Vendor – Tells whether the vendor is VeriFone Inc or Creative Mobile Technologies

# Business Requirements

The business team wants to get an overview of average trip amount trend, comparison of cash and credit card transactions, total trips per borough, and a report consisting of total trip distance and total revenue generated per borough aggregated by month and year.

# Solution architect

Below is the architecture of the solution built in this project. NYC Taxi Data is a raw dataset partitioned based on month and year and stored in a folder structure. Each month has more than 50k records. For the purpose of project, we have extracted 6 months of data of the year 2021.

The goal is to extract all the 300k (approx.. total of 6 months) records within the bronze schema, convert the raw data into PARQUET format for faster processing and blend the data together in the gold schema. The goal is to reduce the latency and could refresh the live dashboard within minimum delay.

Graphical user interface, application

Description automatically generated

# Tasks completed

Designed database named nyc\_taxi\_ldw and created bronze, silver and gold schemas within the database.

Created external file formats within the serverless SQL pool to handle different file formats such as CSV, TSV, JSON and PARQUET.

Created external tables namely bronze.taxi\_zone, bronze.trip\_type, bronze.trip\_data\_green\_csv and bronze.calendar within bronze schema and extracted the raw data files in the bronze schema.

Converted the raw file format into PARQUET format to improve the processing and loaded the data within silver schema tables namely silver.taxi\_zone, silver.trip\_type, silver.trip\_data\_green\_csv and silver.calendar.

Cleaned and aggregated the data present in PARQUET format, joined the data together and loaded the data in gold schema namely gold.vw\_trip\_data\_green. (At this step, we aggregated 300k records and converted them into a meaningful dataset consisting of 1099 records in the gold schema)

Integrated the synapse analytics with embedded Power BI service and built an interactive report for the business stakeholders.

Graphical user interface

Description automatically generated with medium confidence

# Script files

Below attached are the SQL scripts to create database and schemas mentioned above in the report. There are total 12 sql scripts for creating tables and views, and 2 scripts for creating user defined stored procedures.



# Issues encountered and lessons learned

1. We wanted to utilize **MS Azure Storage Explorer** to upload the folder structure of our raw data within the container, however, the storage explorer app was not able to detect the azure subscription. This might be because we do not have the admin rights of the subscription and the account is managed by the university.

Graphical user interface, application

Description automatically generated

Remedy – We manually replicated the folder structure within the container and uploaded individual partitioned scripts within the folders.

1. Even though the raw data files are partitioned on the base of year and month, while creating external tables in silver schema, the data was being written in single PARQUET file.

Remedy – Ideally Spark pool is recommended while working with partitioned data. However, we created external Views and utilized file filepath function which provided the ability to query the data based on year and month.