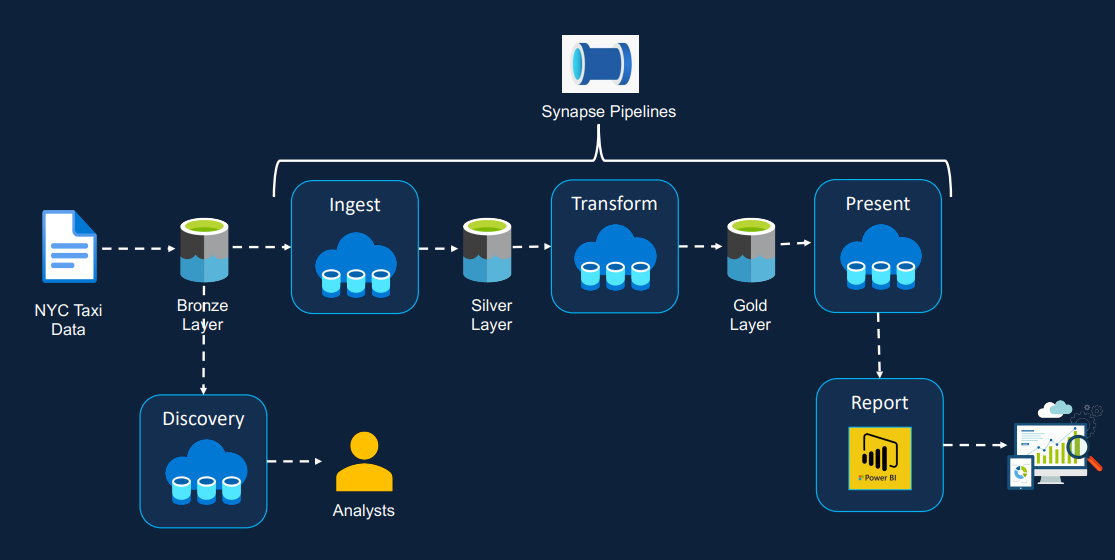
Microsoft Azure Synapse project document.



Solution Architect of the project

Main coding and transform tool ServerLess SQL pool

Data storing in data lake:

This data lake (nyc-taxi-data) is linked with azure synapse

A screenshot of a computer

Description automatically generated

Use openrowset function to read the data straight from the data lake

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

The data which storing in the data lake is current in mixing file type, some csv, some json, some parquet.

A close-up of a computer screen

Description automatically generated

To Query this data we need to use the openrowset which is kind of complicate

So i create a new database inside Azure Synapse include 3 schema which is stand for bronze (raw data storing), silver (raw data but in same format), gold (OLAP data that can be able to do data analyst task or ml task)

A screenshot of a computer

Description automatically generated

Bronze Data Create:

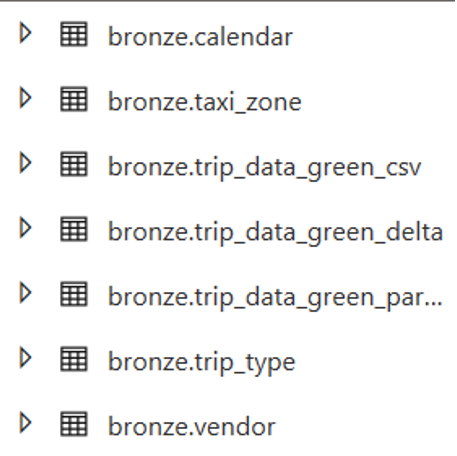
A screenshot of a computer code

Description automatically generated

A screenshot of a computer

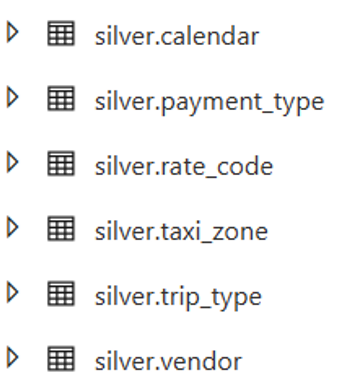
Description automatically generated

After create and import the data into the db inside the azure synapse the compliate query using openrowset is now become normal query.



OK, now the data has entered the Database environment and can be queried. The requirement now is to optimize the data in the data lake so that access in the future is no longer complicated, distinguish between file and file -> transfer all data to the same file type -> .parquet file -> produce a silver layer the data layer is changed to a parquet file and also saved as a table.

This does not mean that the bronze layer is useless, we can reuse the code of the bronze layer to create the silver layer, in addition, the creation of the silver layer is to help optimize the data at the storage (datalake).



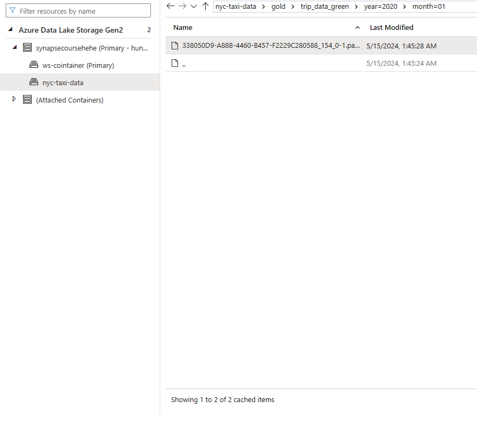
Data in silver layer store in data lake :

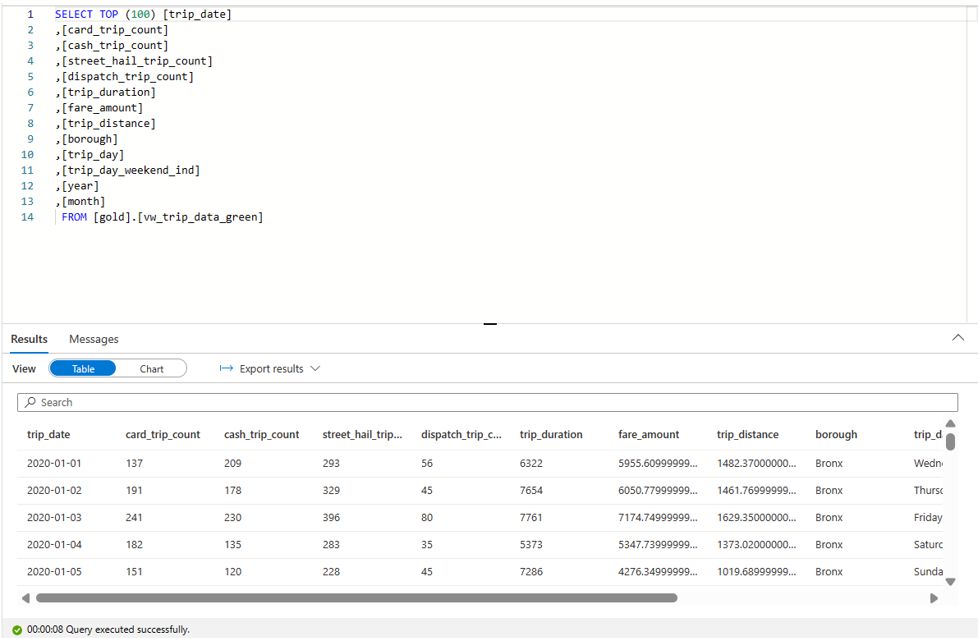
A screenshot of a computer

Description automatically generated

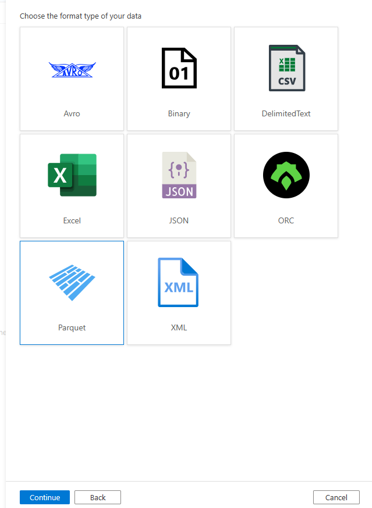
From the data in the silver layer, we can use it to generate queries for the purpose of analyzing the data and saving the results as views and storing the views in the gold layer, which is the presentation layer, so that the bi tools can access these outputs and make reports or dashboard.

The table in gold layer is divide by year and month that enrich the way of access and analyzing the data.





The current steps are ok, but you still have to manually trigger by hand, you have to build a pipeline for it to run from A - Z every action and set time for it to run.



The benefit of creating parquet files in the silver layer is that you can now link the data to Azure synapse pipeline to create a pipeline by simply selecting 1 type of dataset

All the previous action but run inside azure synapse pipeline

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated