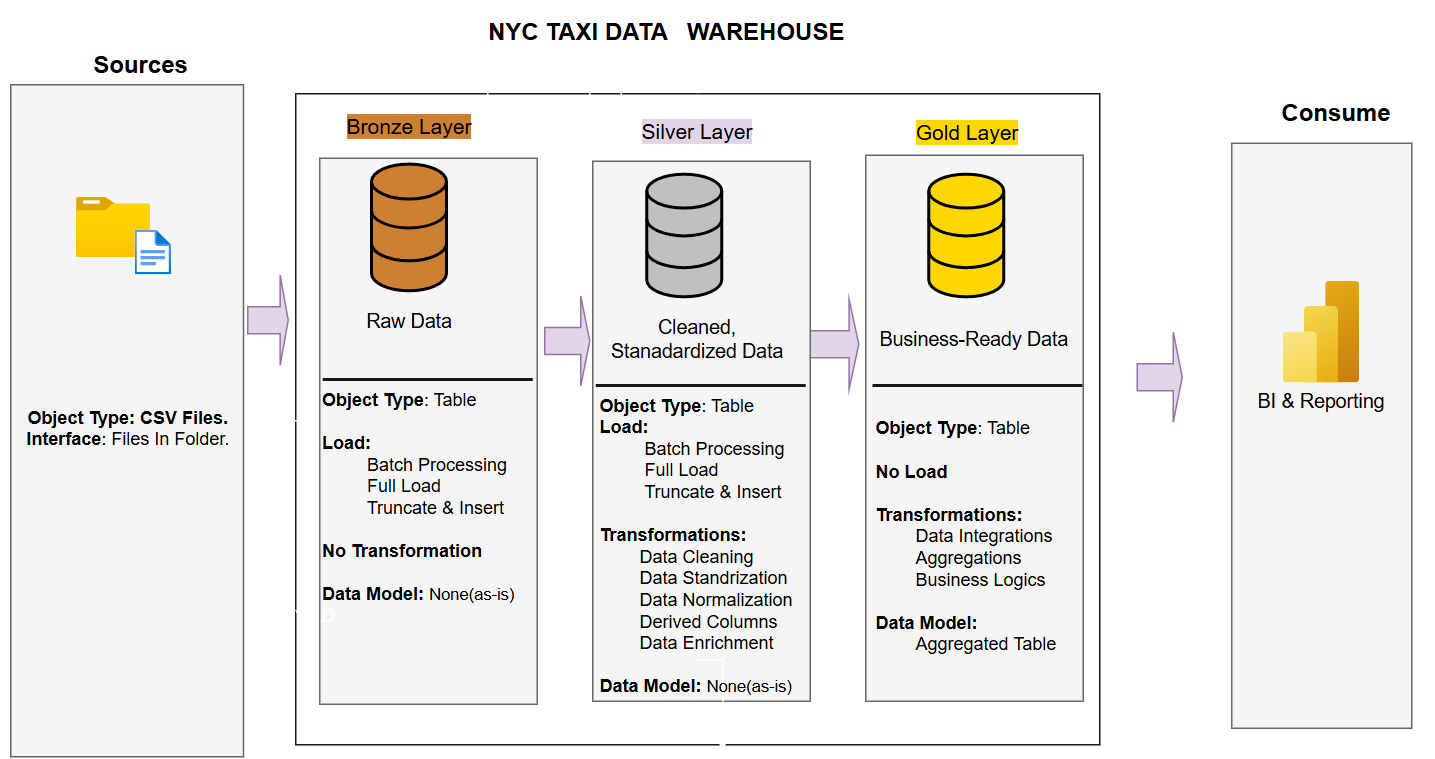
1. 
2. The sql code is available in the github. (https://github.com/Jaamaldeen/NYC-TAXI-DATA-PIPELINE/tree/main)
   1. DESIGN RATIONALE

The Bronze layer stored the raw data without any changes, the silver is where the cleaning, transformation, standardization and deriving new columns that will be useful for the business is done and the gold layer is where the aggregation and summarizing of data is done.

* 1. FULL VS INCREMENTAL LOAD

The full refresh has a direct solution, I create the table for the bronze layer, and upload all the csv directly with COPY FROM for my folder, i use CTAS(Create Table As) to perform the cleaning, transformation, derived new columns and inserting the data to the silver layer and CTAS also in creating 5 different table for data aggregation and summary.

The incremental load is a little tricky, after much brainstorming, I decided to use triggers on my layers, starting from the bronze layer, and I also create a metadata schema, which contain a table that will store the current time when there is any change in my bronze, silver and gold layer, the save time will be used in determine which data should come into the silver, although natural we supposed to use a natural column from the raw data, but we can’t because we are working with a past data that is why a add a new column new time\_uploaded, which is used to compare with the last time that was an updating in the bronze table, so no matter the volume of data in the bronze table, the silver will ingested data that is new i.e the time\_uploaded is greater than the last time we updated our bronze table.

So this is how, my code on incremental loading work, I create a trigger on the bronze layer that anytime a new data is inserted into my bronze layer, the silver layer will check the data, if it a new data and the silver layer will do all the necessary transformation, cleaning and deriving new columns before inserting the data into the silver table and also another trigger is on that silver.taxi table. Anytime new data is appended in the silver layer, the trigger will be triggered and it appends only the new data to the gold layer, during the appending the necessary aggregation and summarization will occur into the 5 different tables that were created for the gold layer.

The only human interaction that will occur is when you want to upload a new data, the silver and the gold layer will automatically happen.

* 1. METADATA MANAGEMENT

The metadata is done by creating the last\_loading\_period that will automatically update any time that is a new update in the bronze tables, silver tables and even gold tables.

* 1. These are some few businesses that are answered with the gold layer.

SELECT trip\_date, total\_revenue

FROM gold.daily\_summary

ORDER BY trip\_date;

SELECT vendor\_name, total\_revenue, avg\_fare

FROM gold.vendor\_summary

ORDER BY total\_revenue DESC

LIMIT 5;

SELECT payment\_description, avg\_tip\_percent

FROM gold.payment\_summary

ORDER BY avg\_tip\_percent DESC;

SELECT month, total\_trips, total\_revenue

FROM gold.monthly\_summary

ORDER BY month;

SELECT PULocationID, pickups, revenue\_from\_pickups

FROM gold.zone\_summary

ORDER BY pickups DESC

LIMIT 10;

1. The pipeline can run multiple times, without duplicates, because there is CONSTRAINT unique in the bronze layer, so the bronze layer will automatically check for any duplicate and reject any duplicates data and once the bronze layer rejects duplicates record, silver layer and gold are safe.