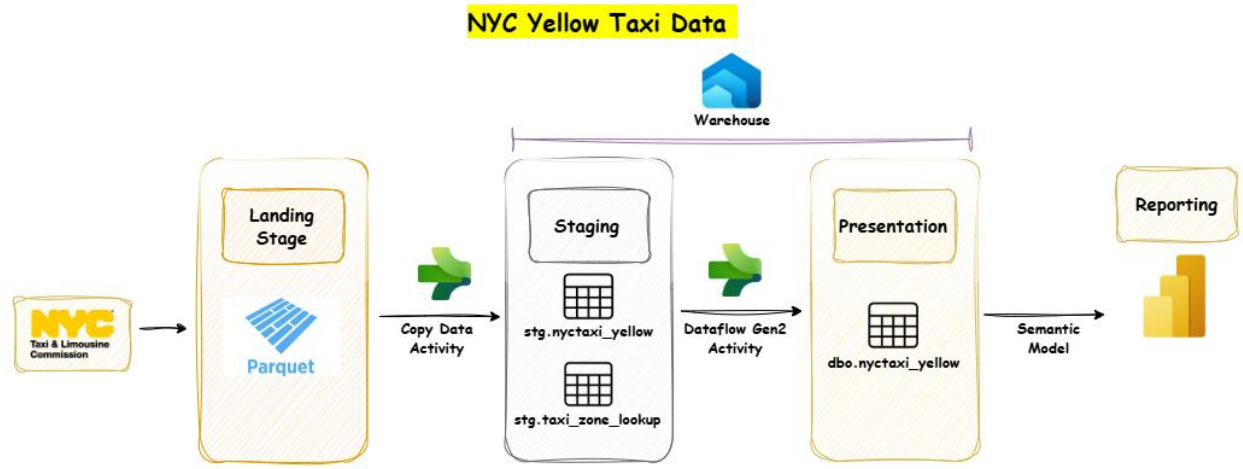
NYC Taxi Data Project Using Microsoft Fabric

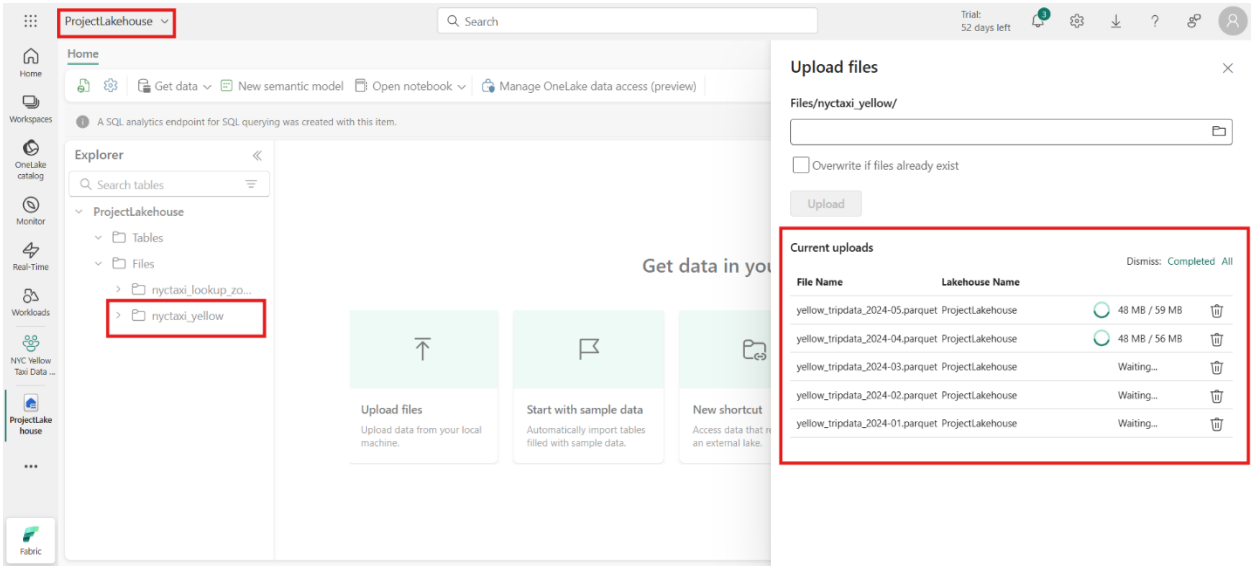
Overall Project Architecture,

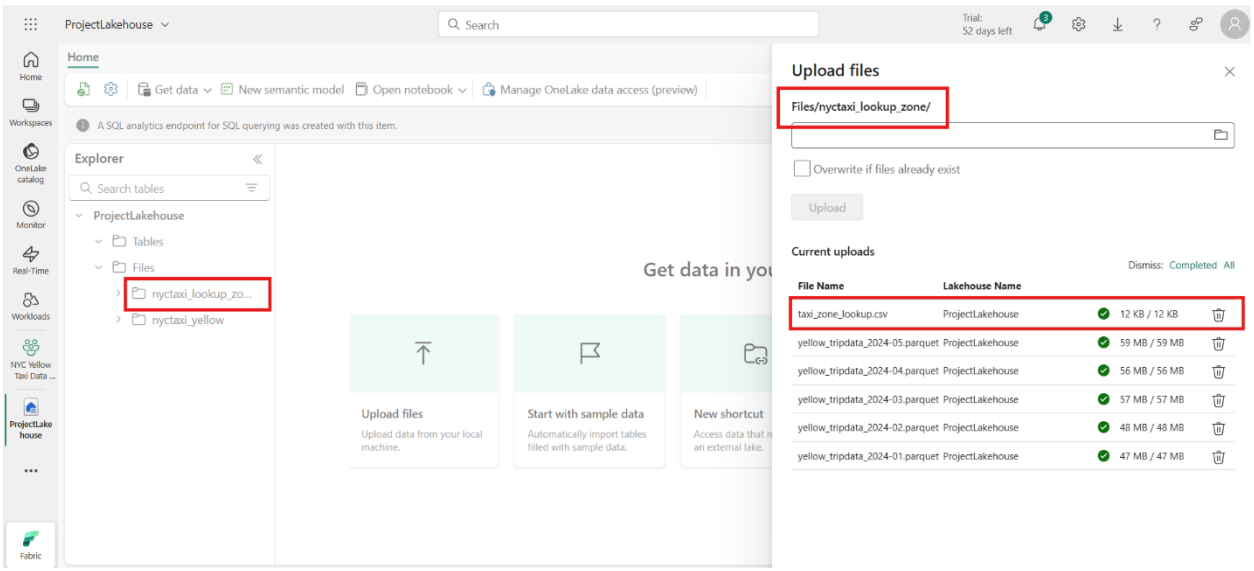


Create Fabric Workspace by providing relevant name for the workspace.

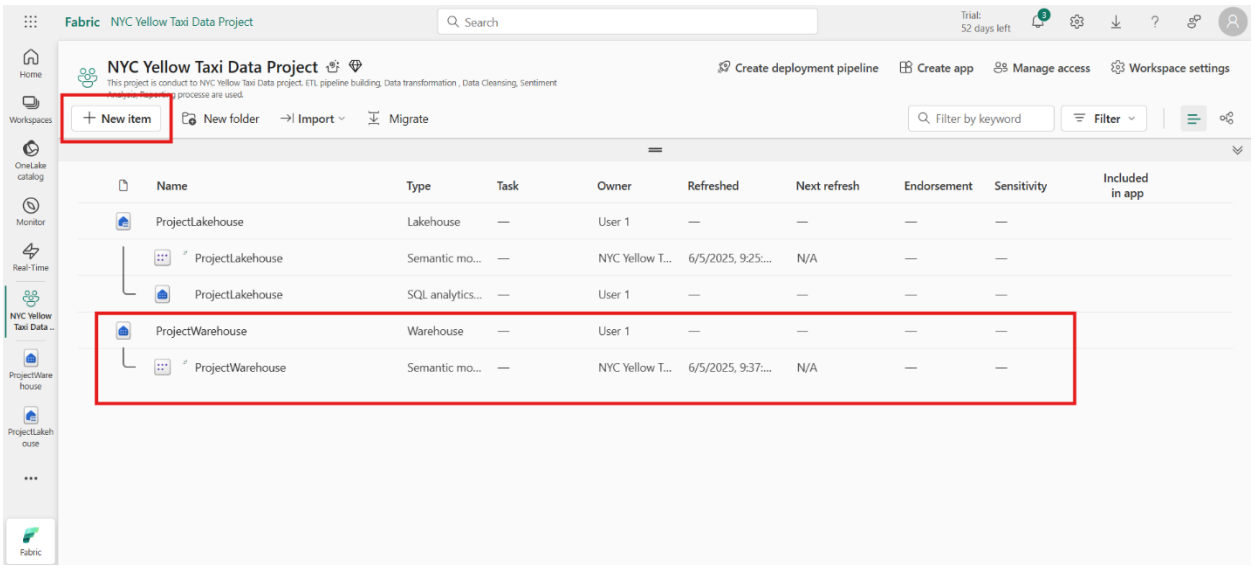
Create Lakehouse as ‘ProjectLakehouse’.

Go to the Lake house and Create sub folders under the Files folder as ‘nyctaxi_lookup_zone’ and ‘nyctaxi_yellow’. Upload Yellow trip data Parquet files to the ‘nyctaxi_yellow’ folder and Upload taxi_zone_lookup.csv file to the ‘nyctaxi_lookup_zone’ folder.

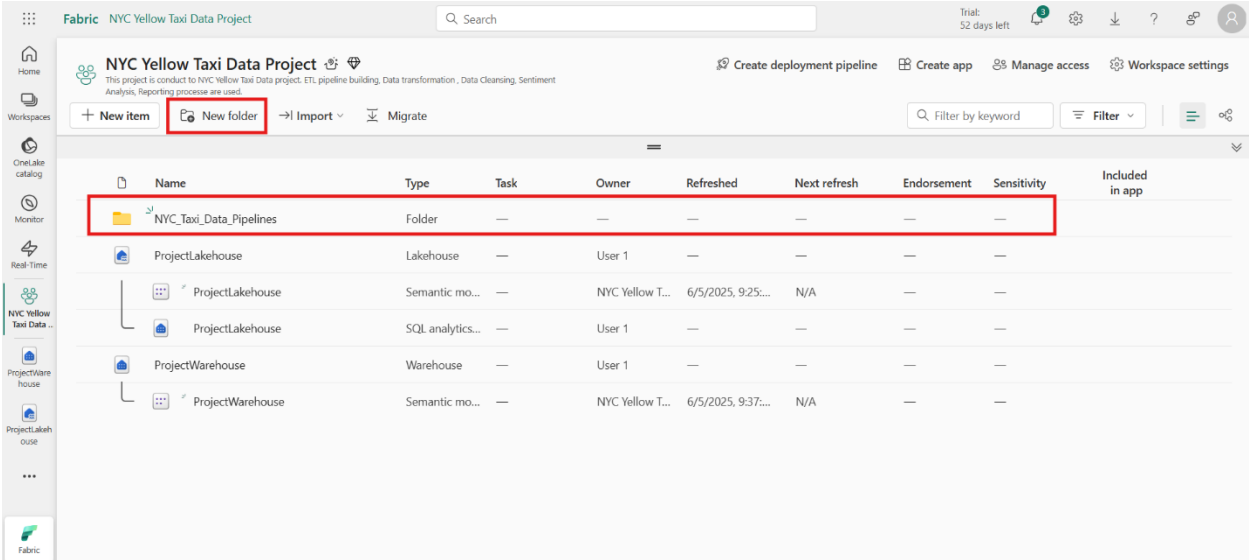




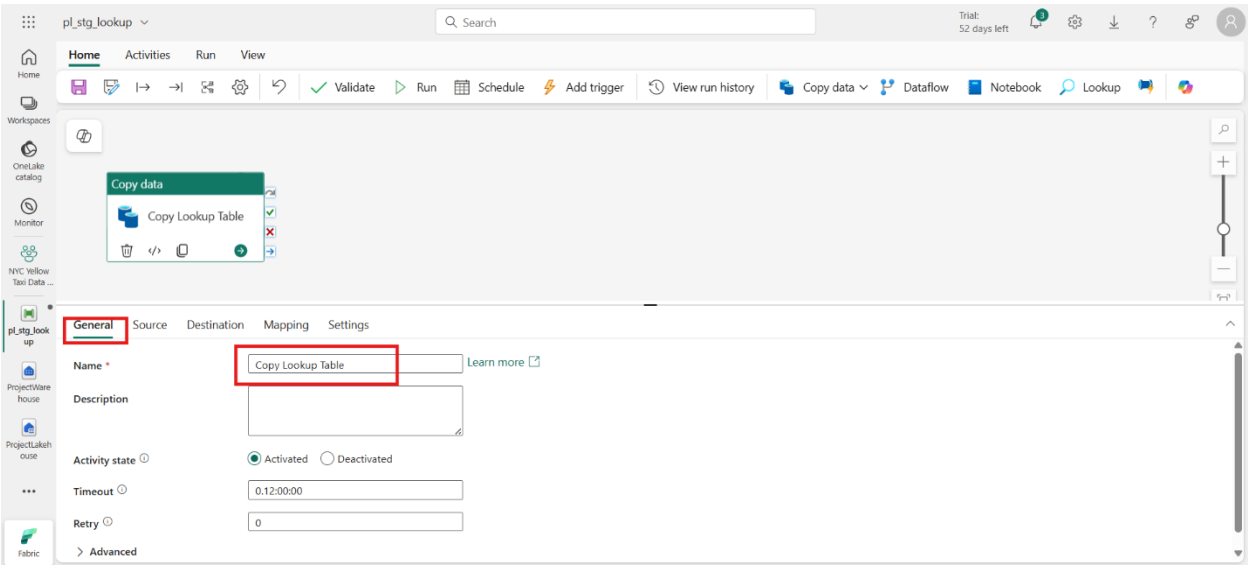
Create Warehouse as ‘ProjectWarehouse’ in workspace.



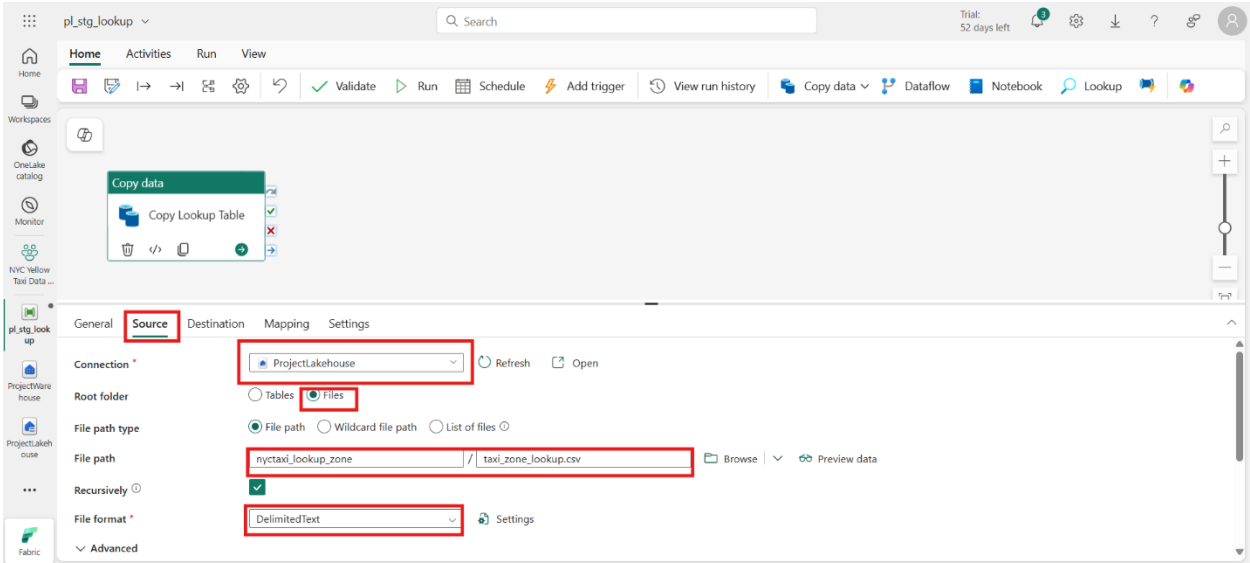
Create Folder for Pipelines as ‘NYC_Taxi_Data_Pipelines’.



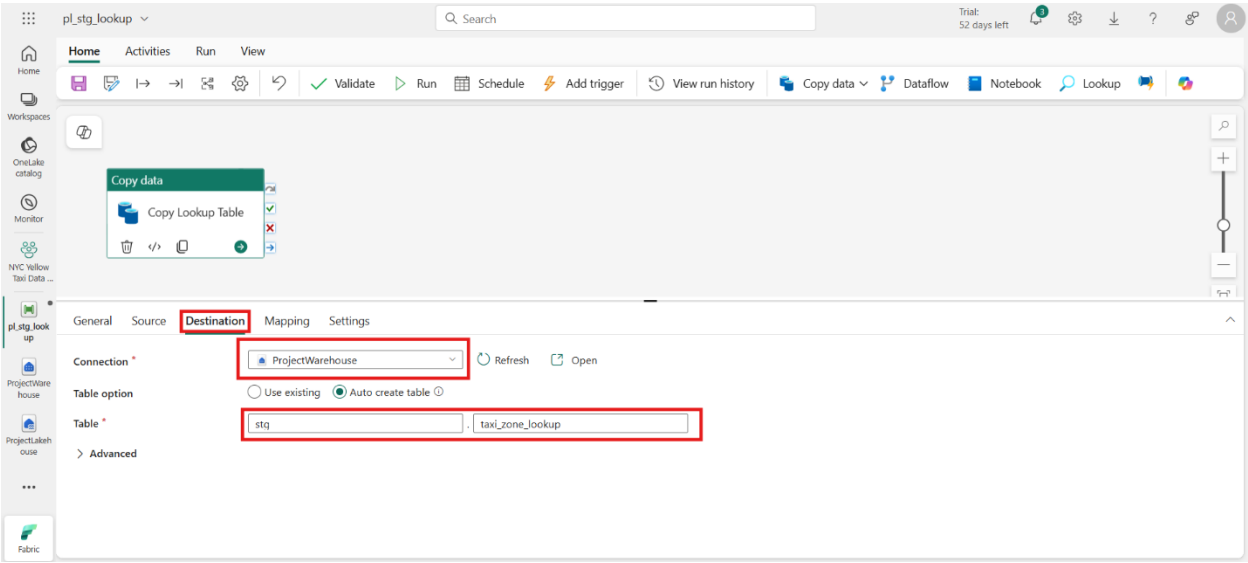
Inside the created folder, Create Pipeline named as ‘pl_stg_lookop’. This pieline is building purpose is Copy all the Data include in the ‘nyctaxi_lookup_zone’ f



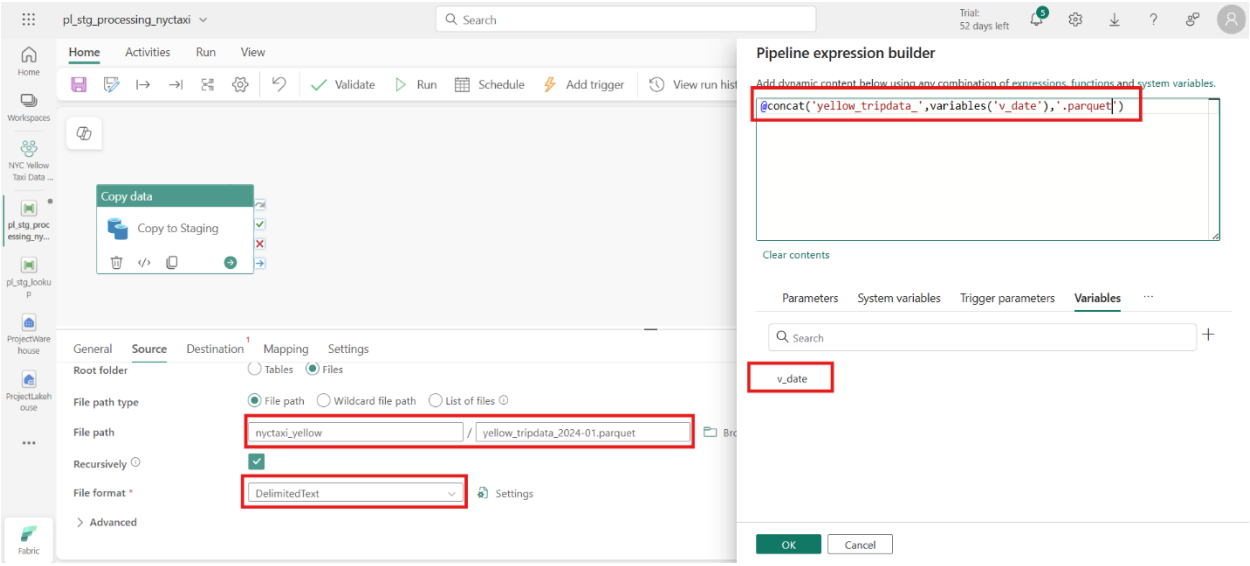
Enter relevant details to the copy data activity source tab.



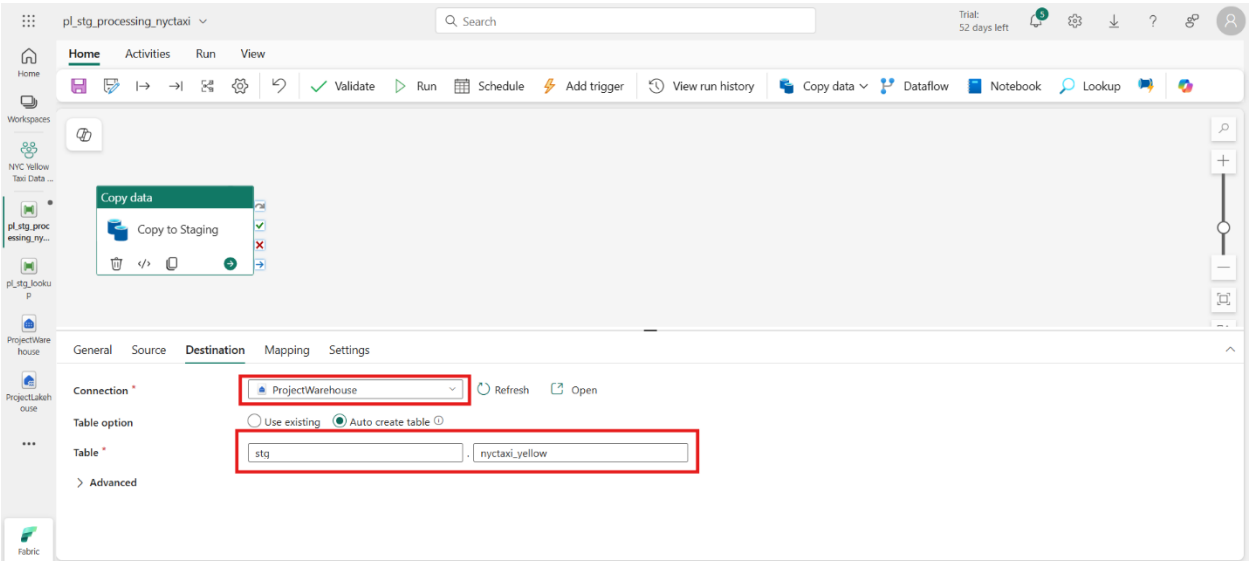
Enter relevant destination details to the destination tab as shown in the below figure.



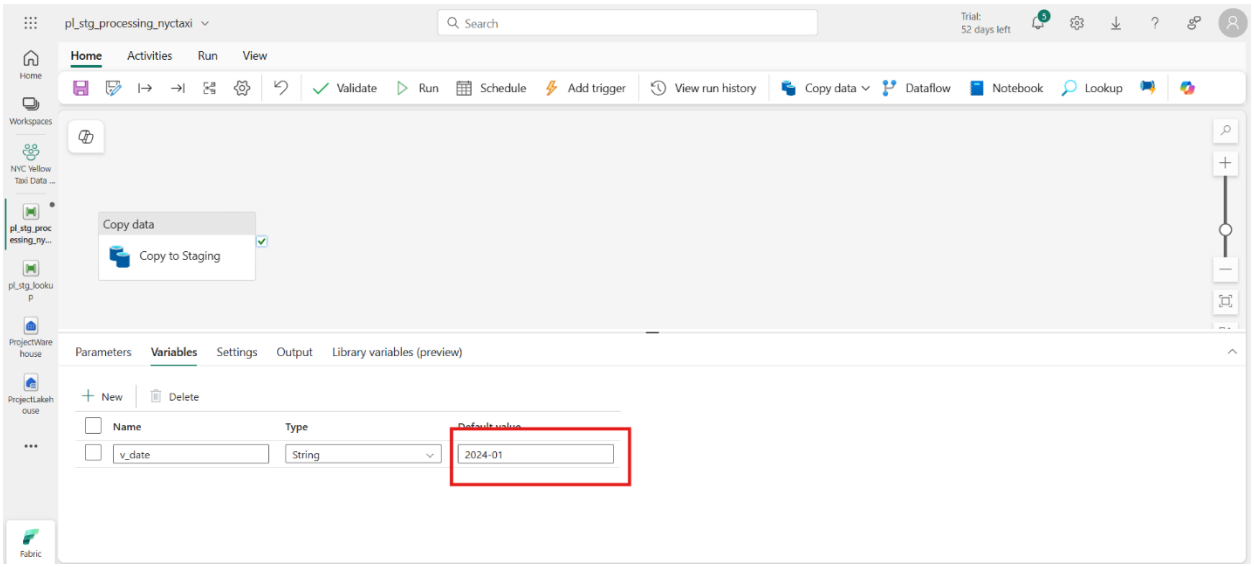
Create another pipeline Inside the ‘NYC_Taxi_Data_Pipelines’ folder. Give pipeline name as ‘pl_stg_processing_nyctaxi’. This pipeline is used to copy ‘Yellow_Tripdata_2024-01.parquet’ file to data warehouse.



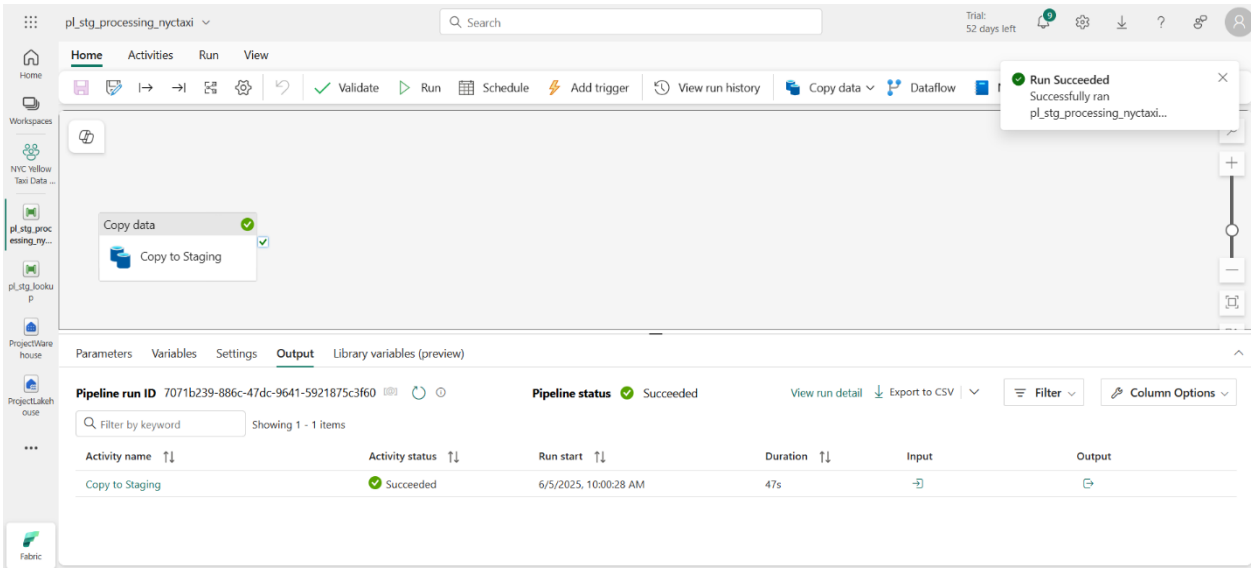
Give relevant details to the destination tab.



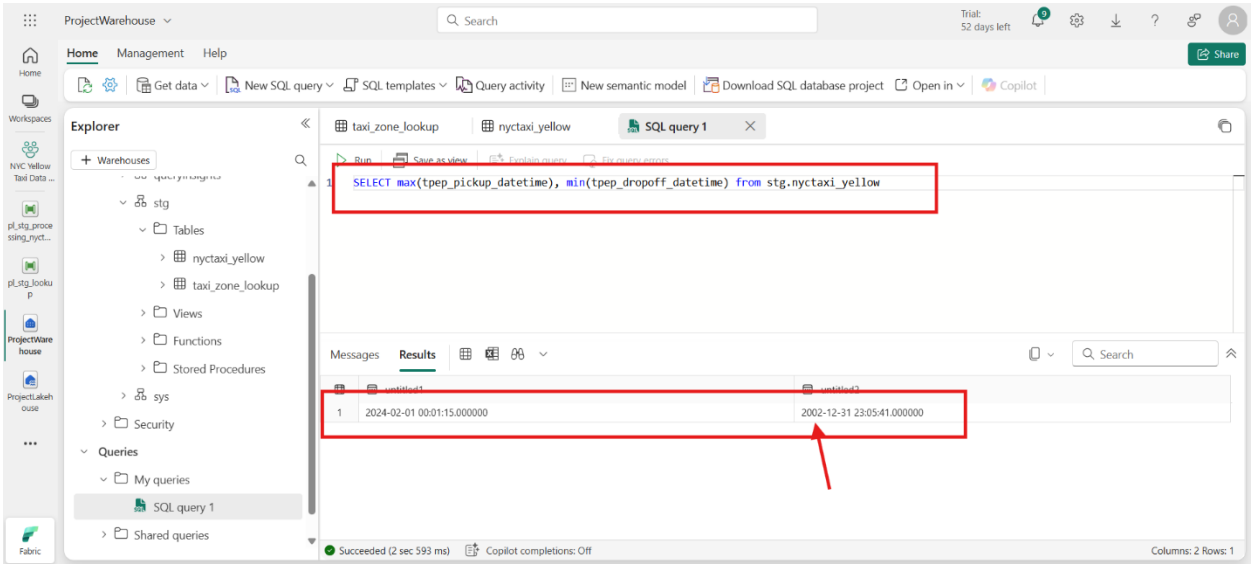
Set value for the v_date variable manually.



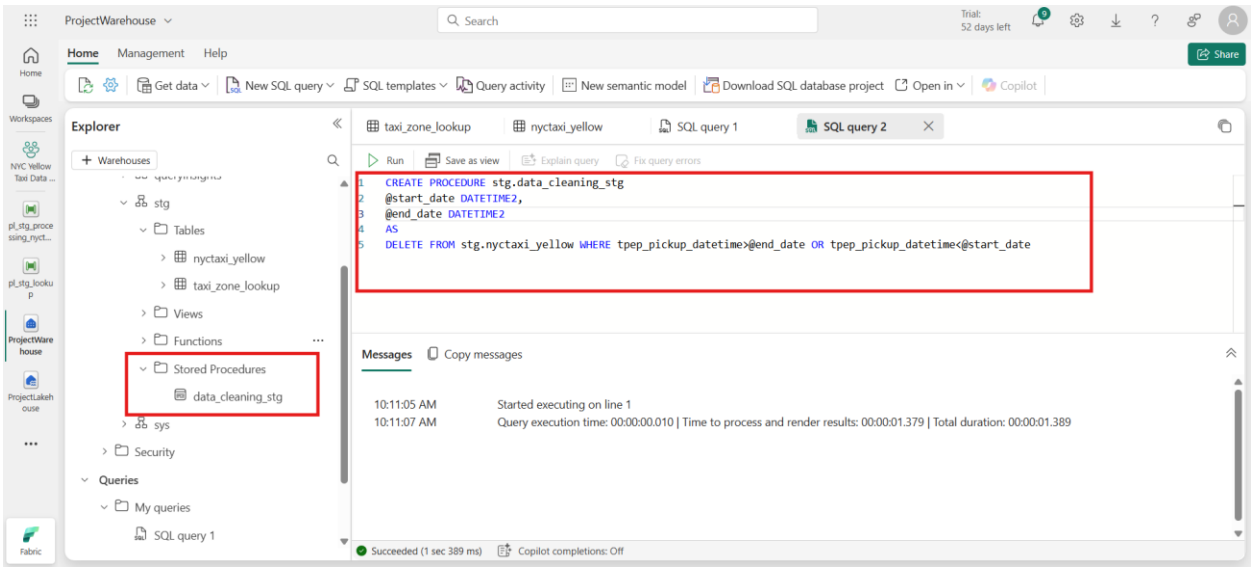
When run the pipeline, if it run successfully, its look like this.



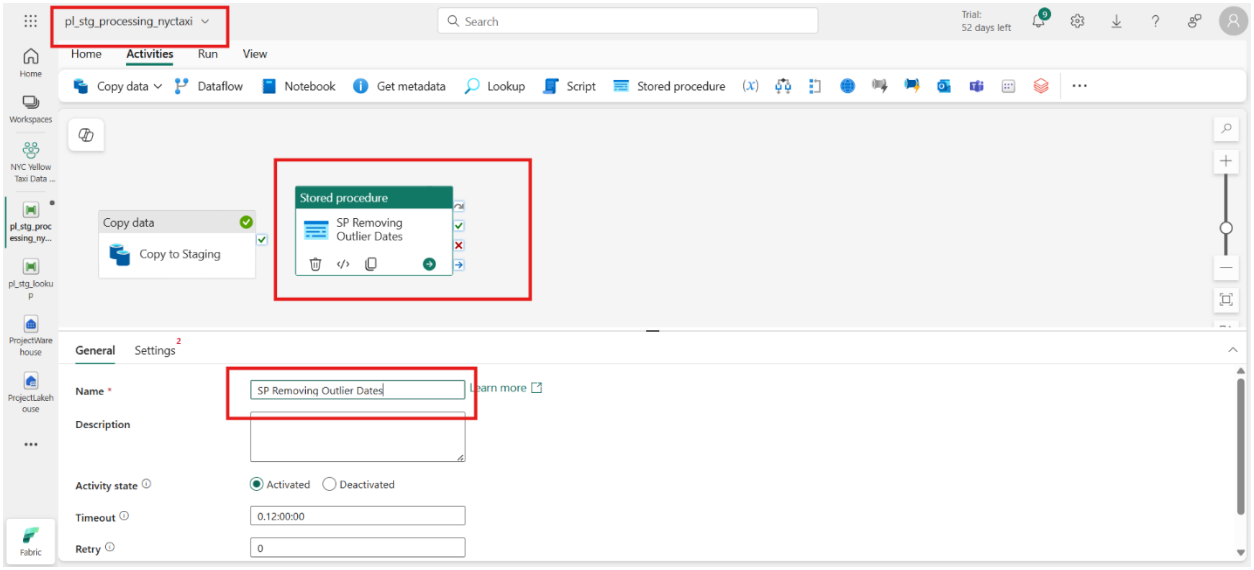
Run this SQL query to identify if there exist any irrelevant data file. As shown in the figure, there exist data relevant to the 2002.



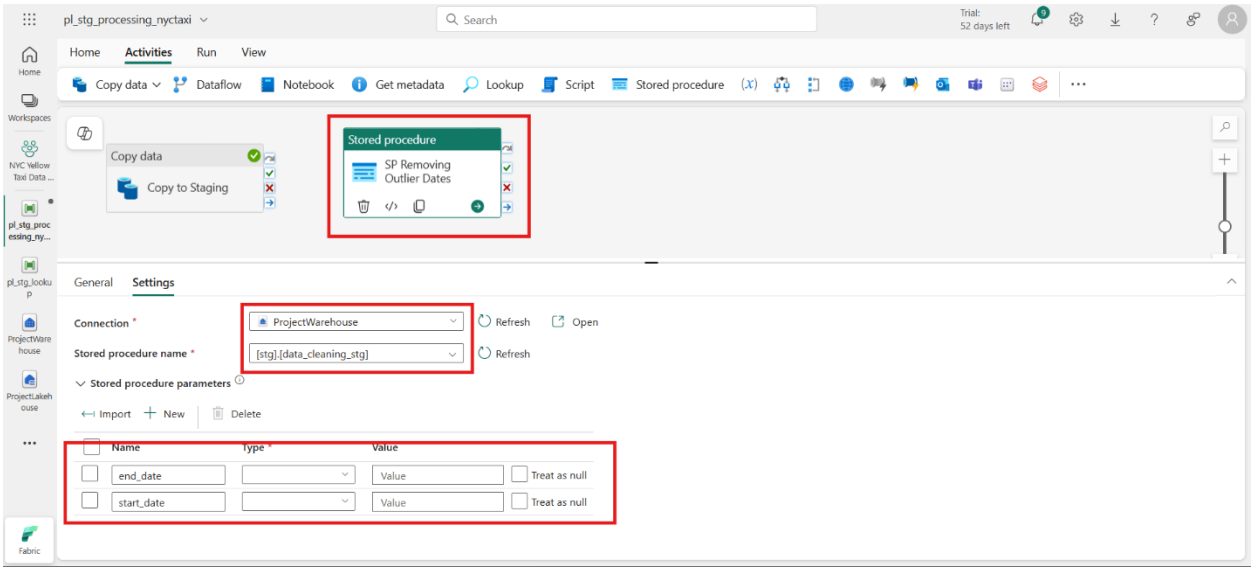
To remove irrelevant data run this query in your warehouse query space. Here create stored procedure as ‘data_cleansing_stg’.



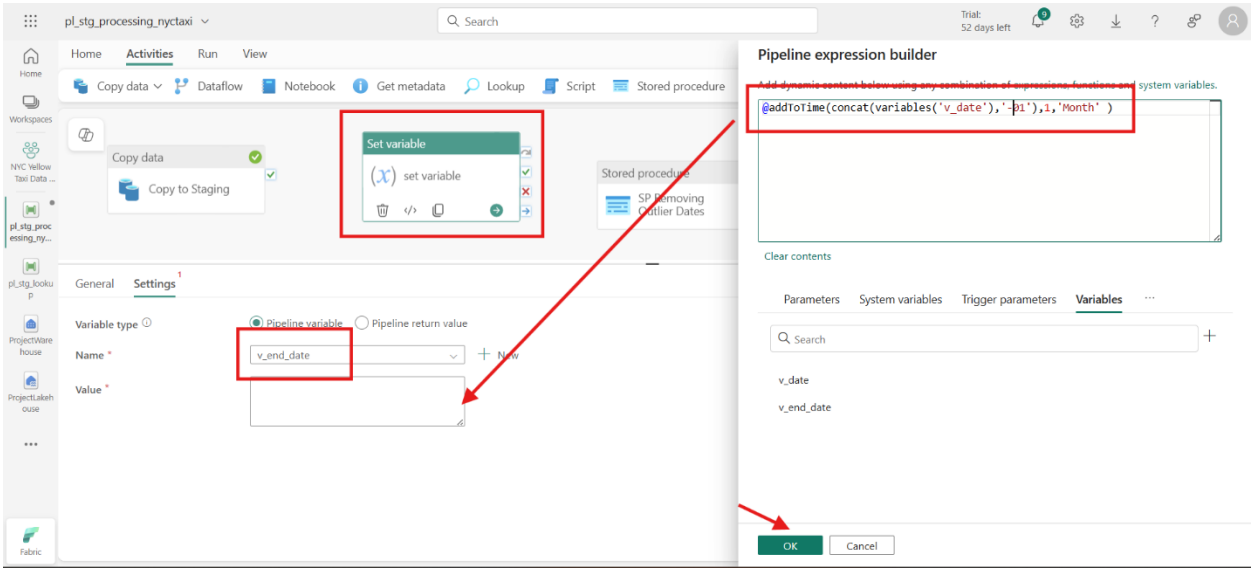
Back to the pl_stg_processing_nyctaxi pipeline and create stored procedure activity named as ‘SP_Removing_Outlier_Dates’.



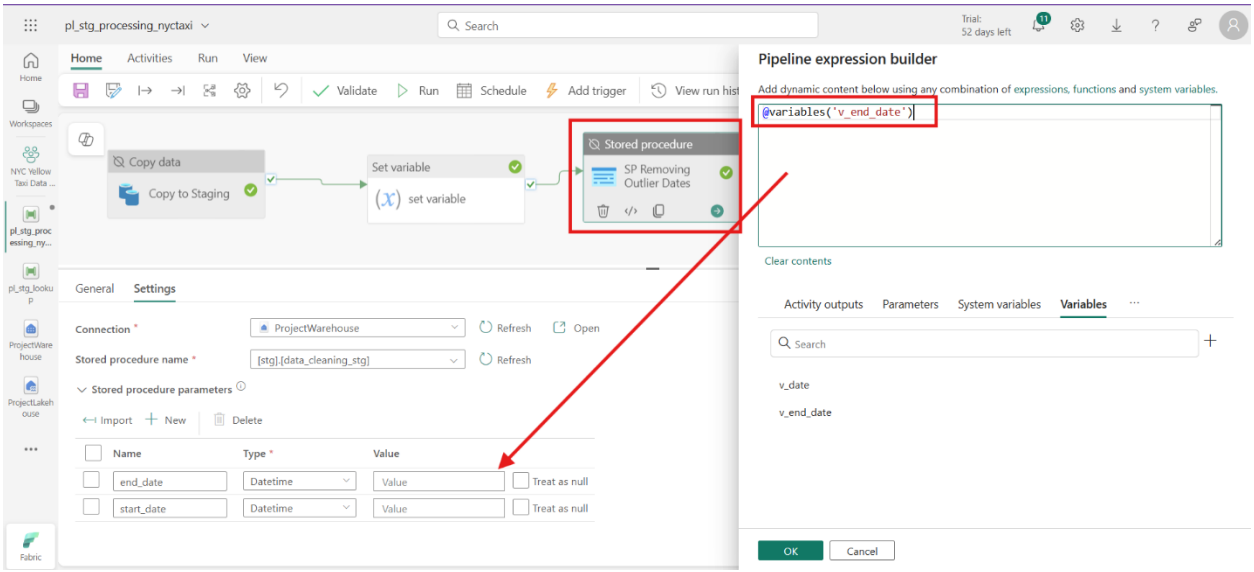
Provide details to the Settings tab as below figure.

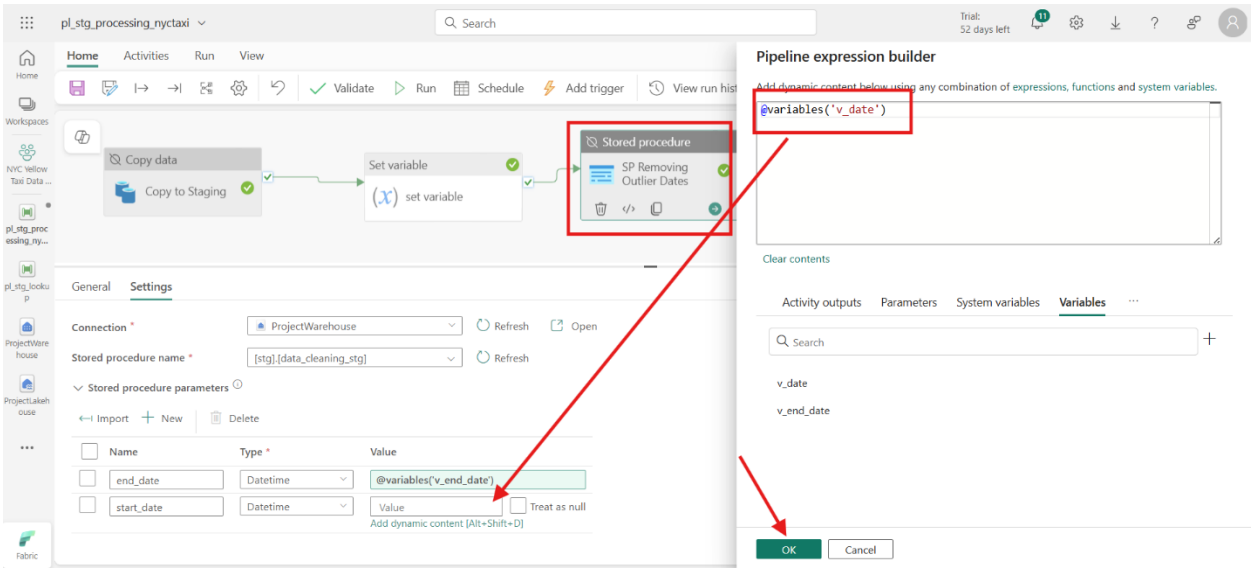


Create set variable activity to store v_end_date value. This variable used to increment month one by one.

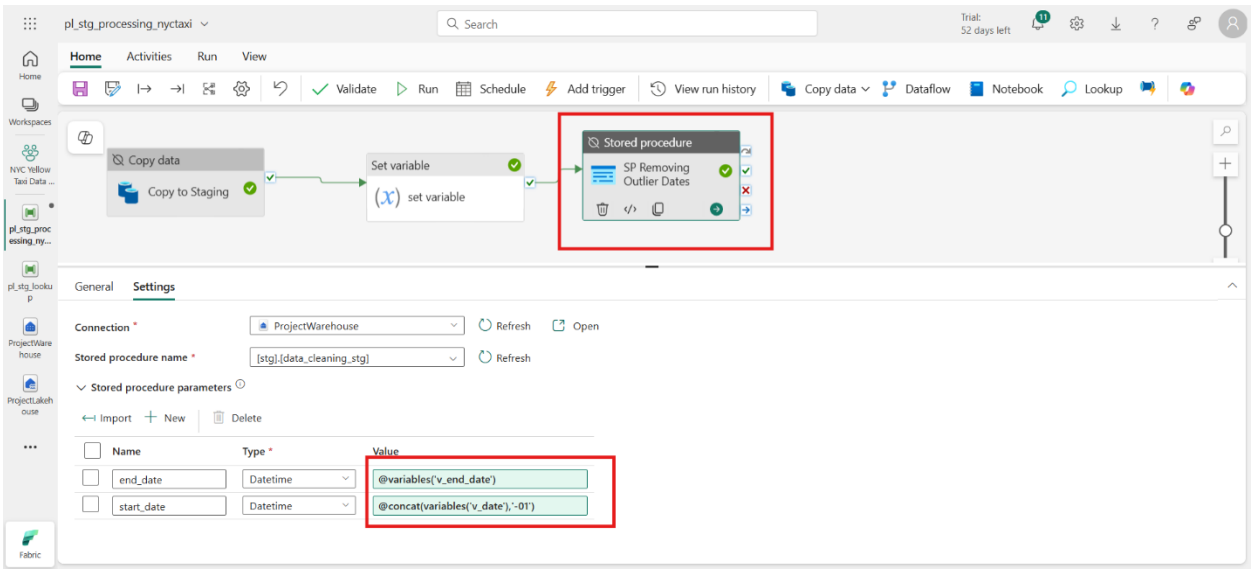


Go to the stored procedure activity and import variables and set values as shown in the figure.

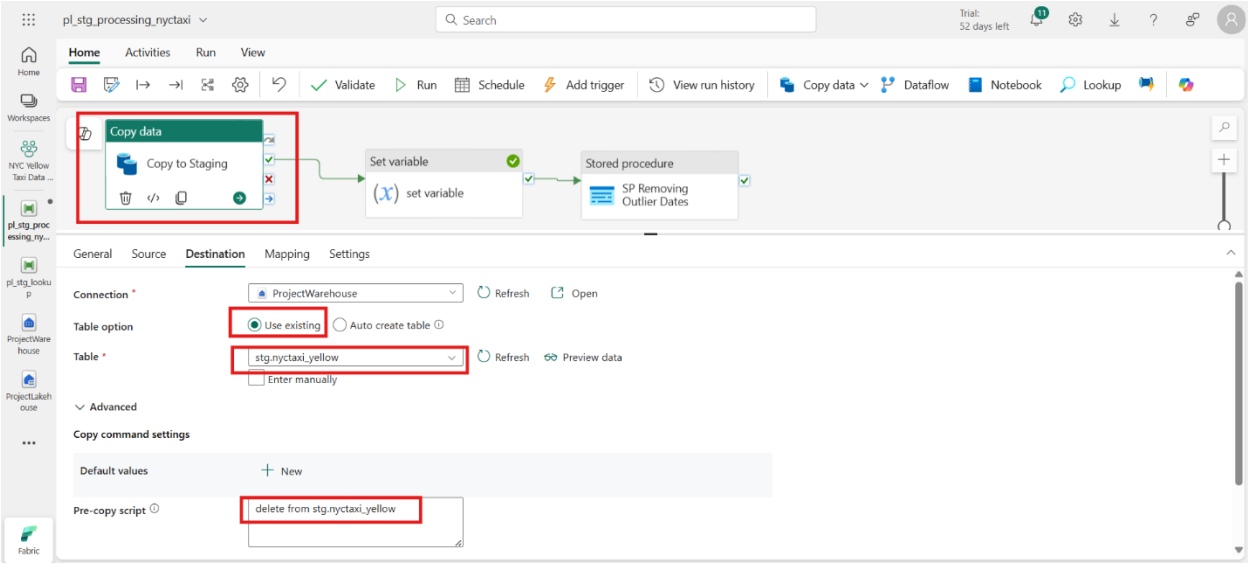




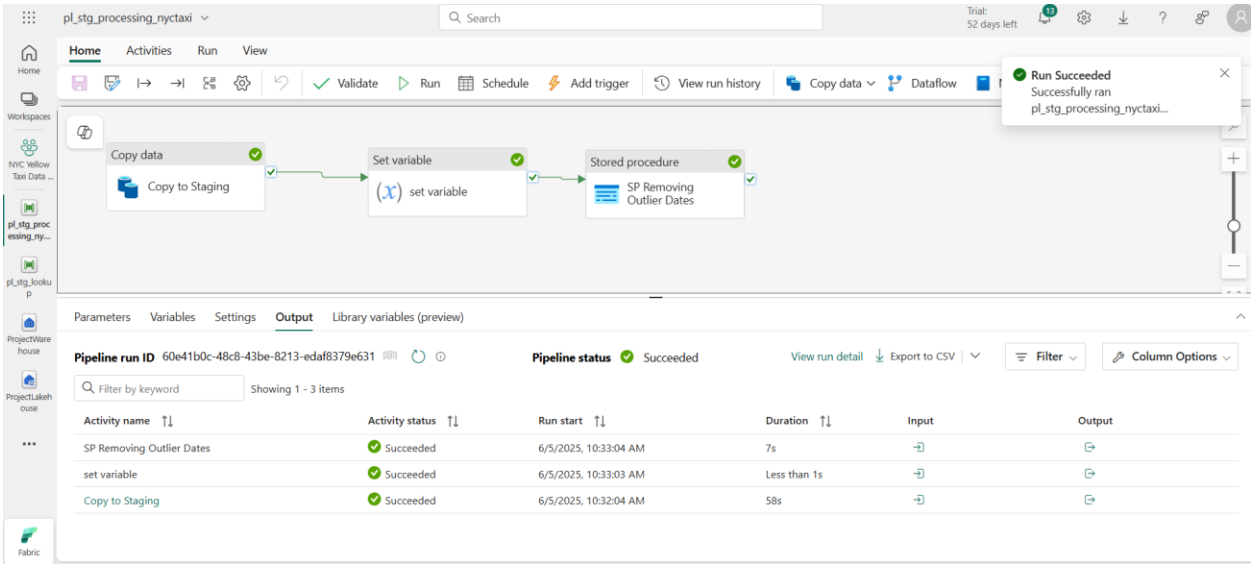
Finally stored procedure activity looks like this.



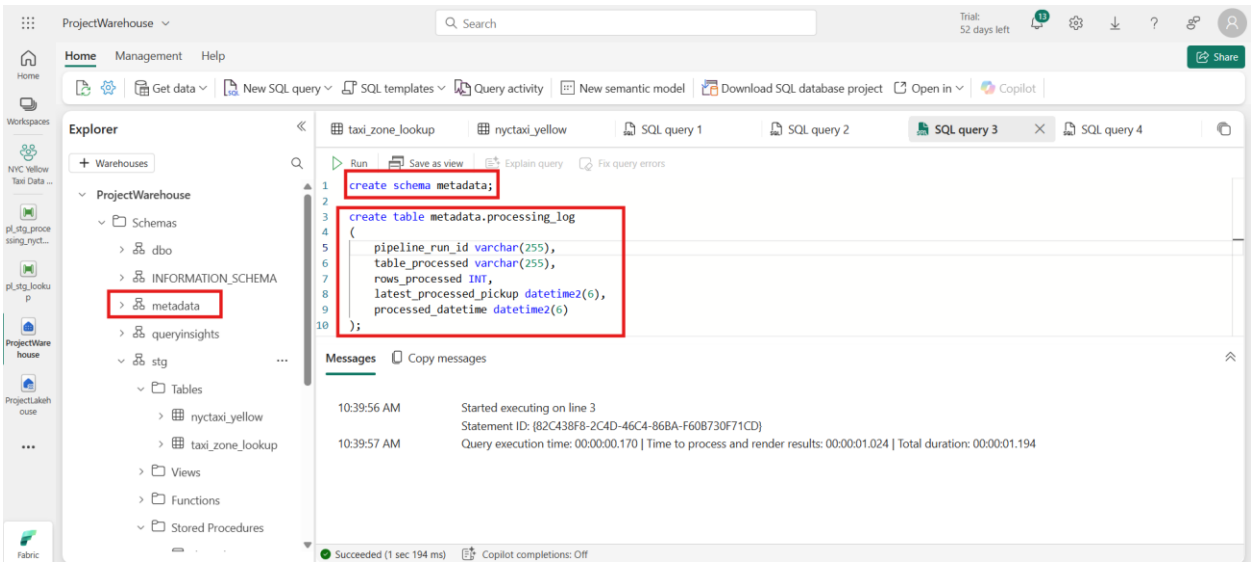
Back to the Copy Data activity change data as shown in the figure. Pre -copy script used to ensure delete previous data from stg.nyctaxi_yellow table. Still fabric doesn't support truncate data function that's why use this script here.



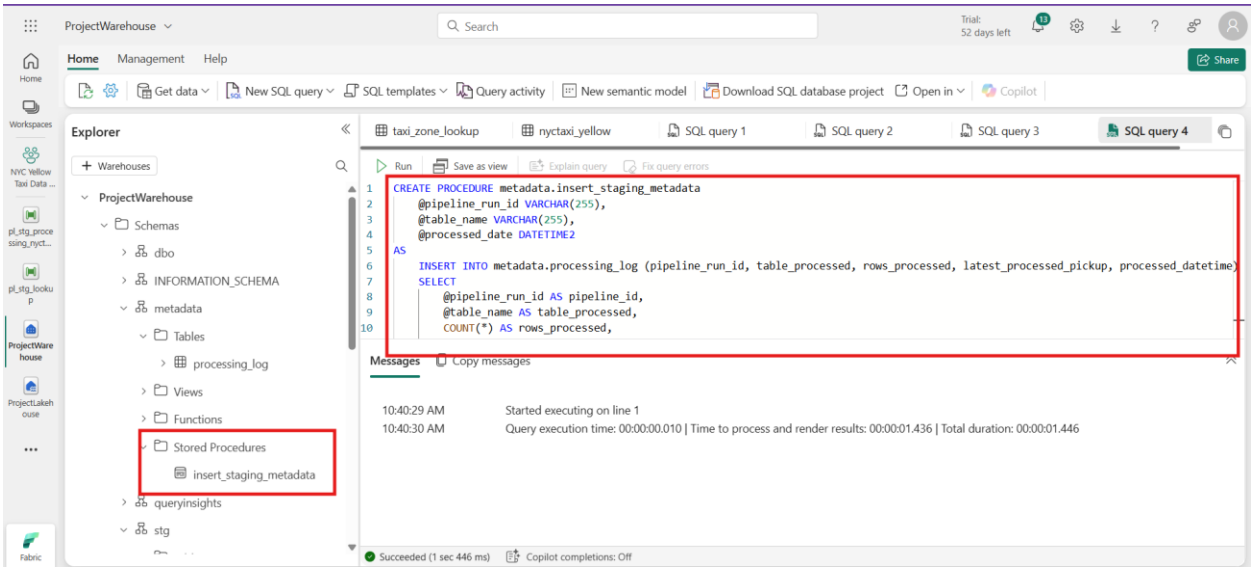
When run the created pipeline it's run successfully as shown inn the figure.



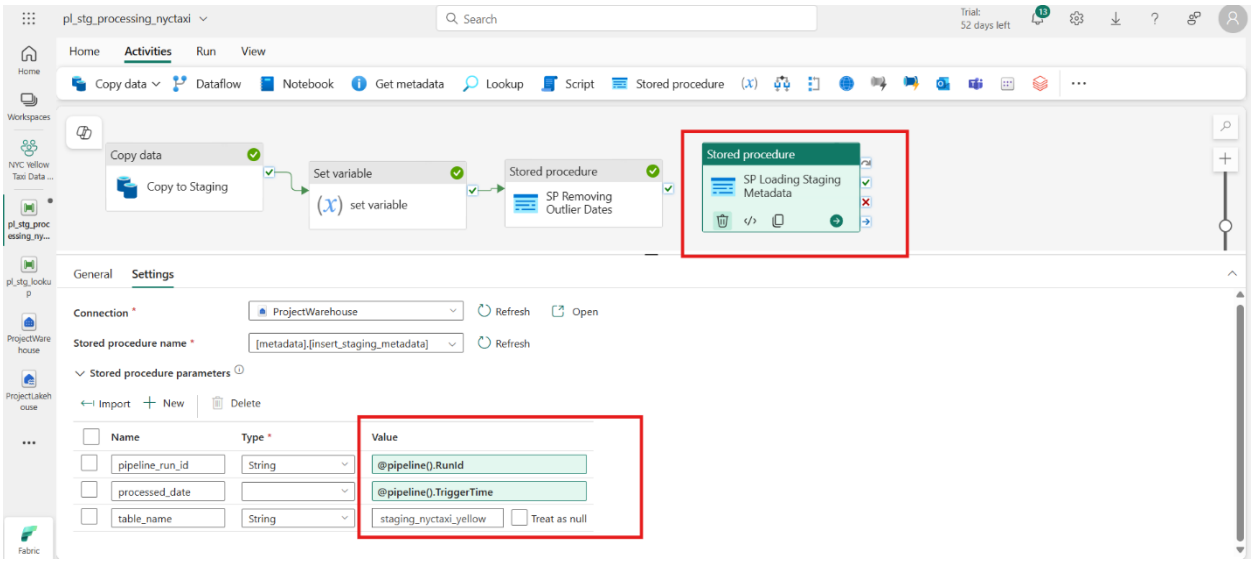
Create sql query to create metadata schema and under the metadata schema create table as metadata.processing_log.



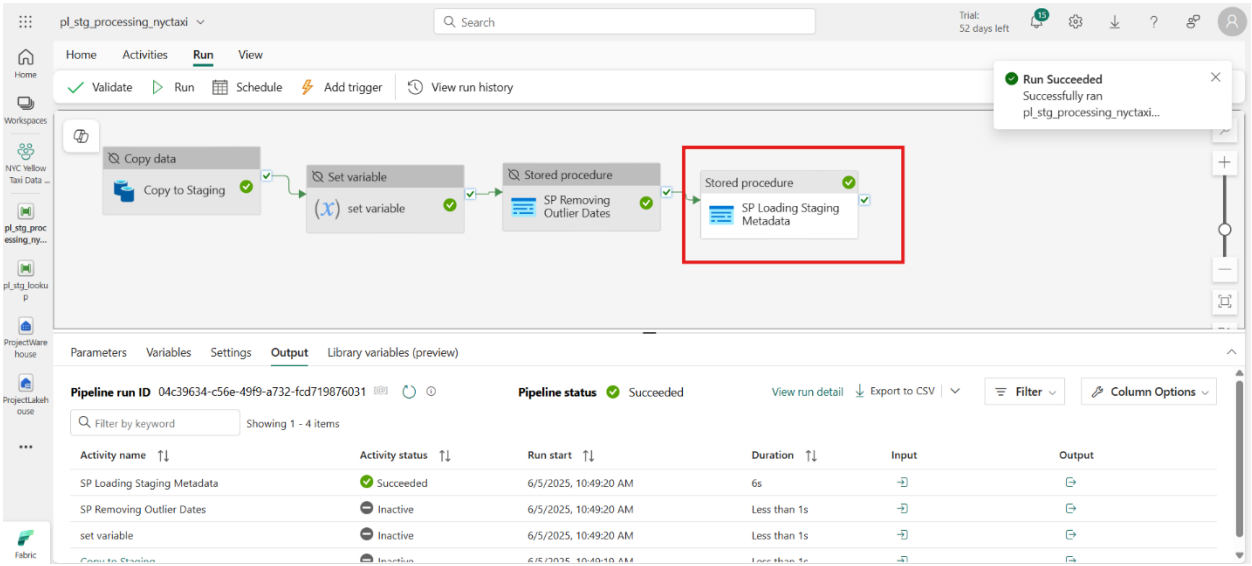
Create stored procedure to store data to metadata.processing_log table.



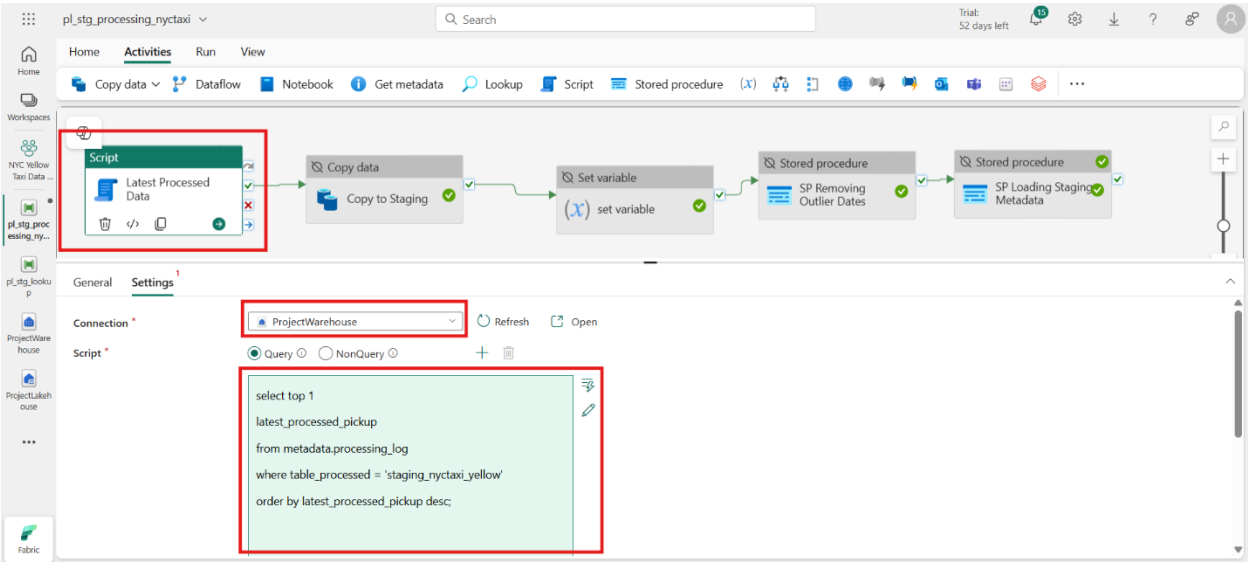
Create another stored procedure activity in pl_stg_processing_nyctaxi ppipeline. Provide relevant details as shown in the below figure.



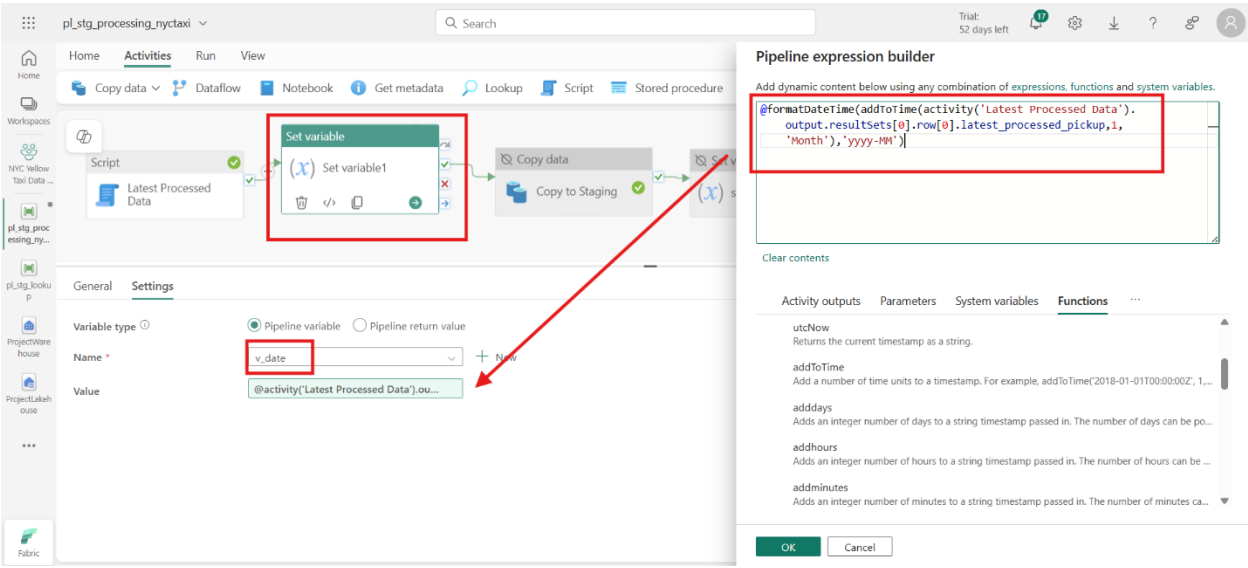
Deactivate remaining activities in the pipeline and run stored procedure activity only. It's run successfully as shown in the figure.



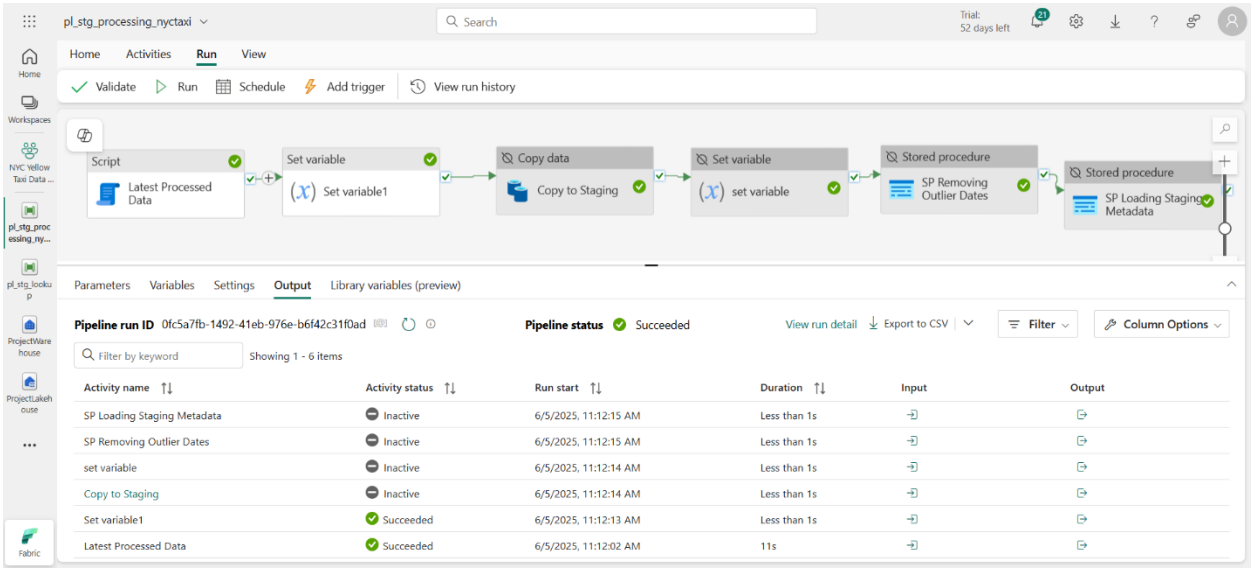
Create Script activity to get latest_processed_pickup top value. We only need the top value to further process.



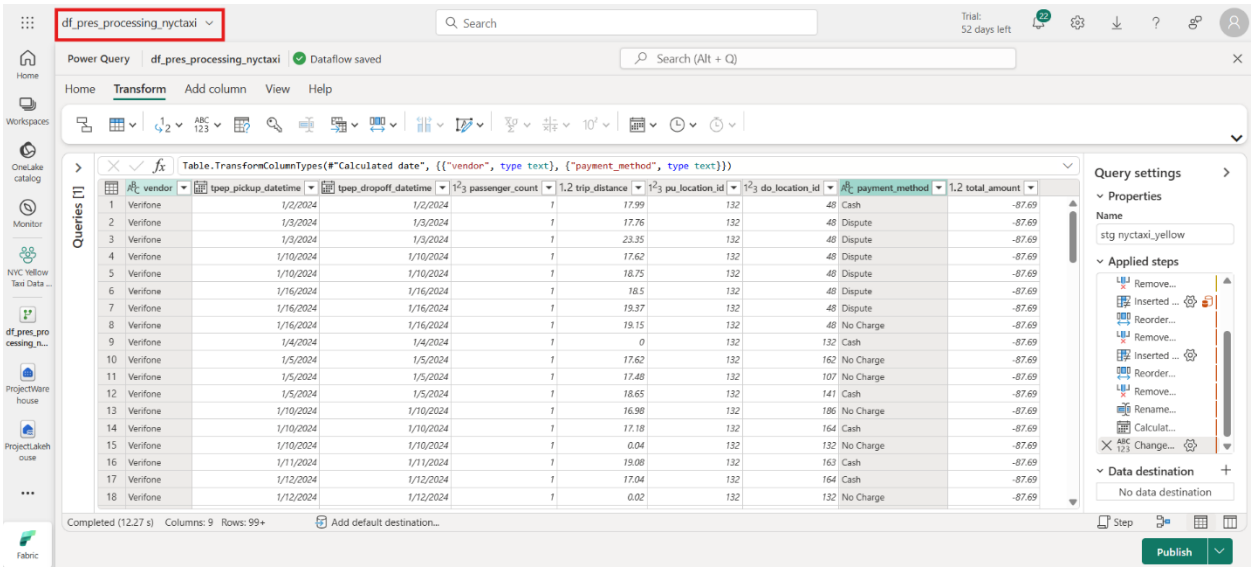
Create another set variable activity after the script activity store that value. Provide details according to the shown figure.



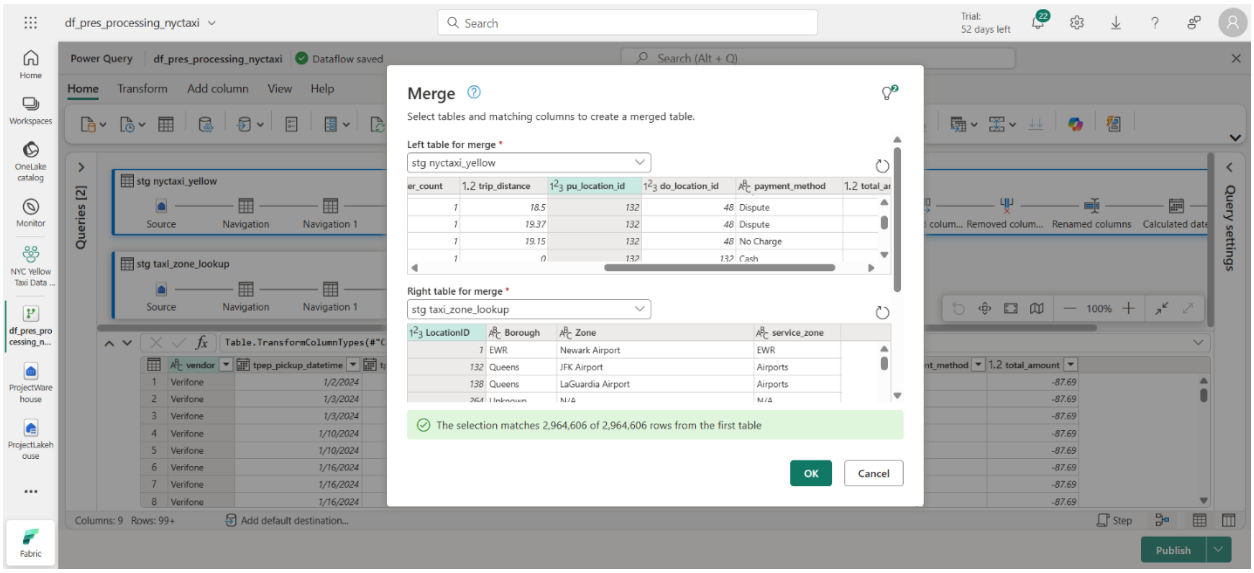
Deactivate all the remaining activities in the pipeline and run only script and set variable activity.



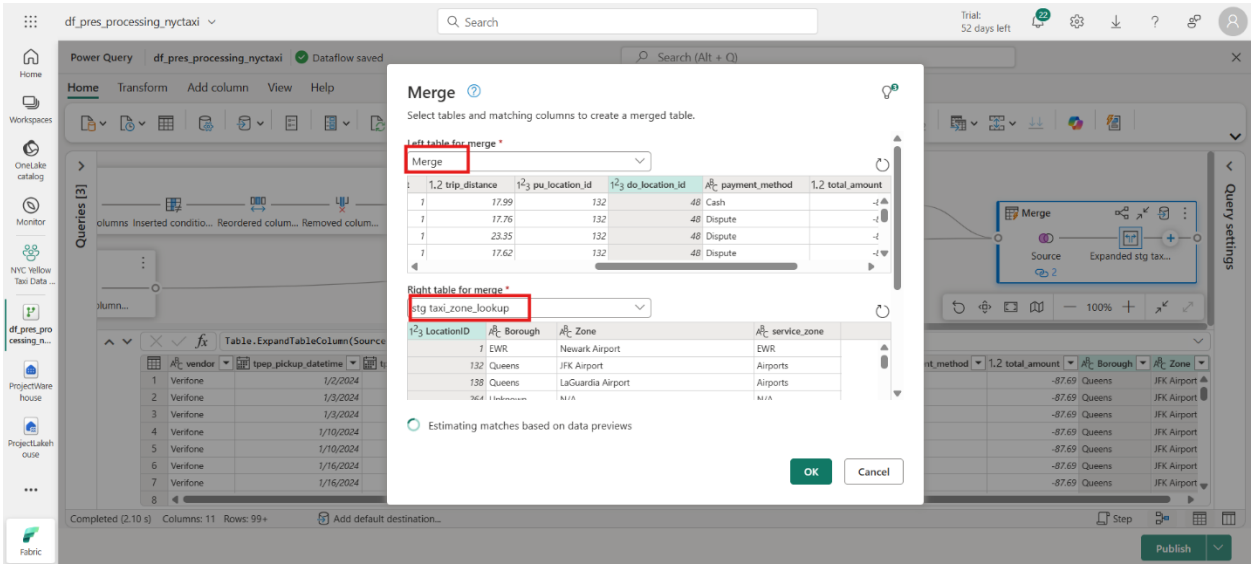
Create Dataflow Gen2 activity to clean, transform and aggregated data.



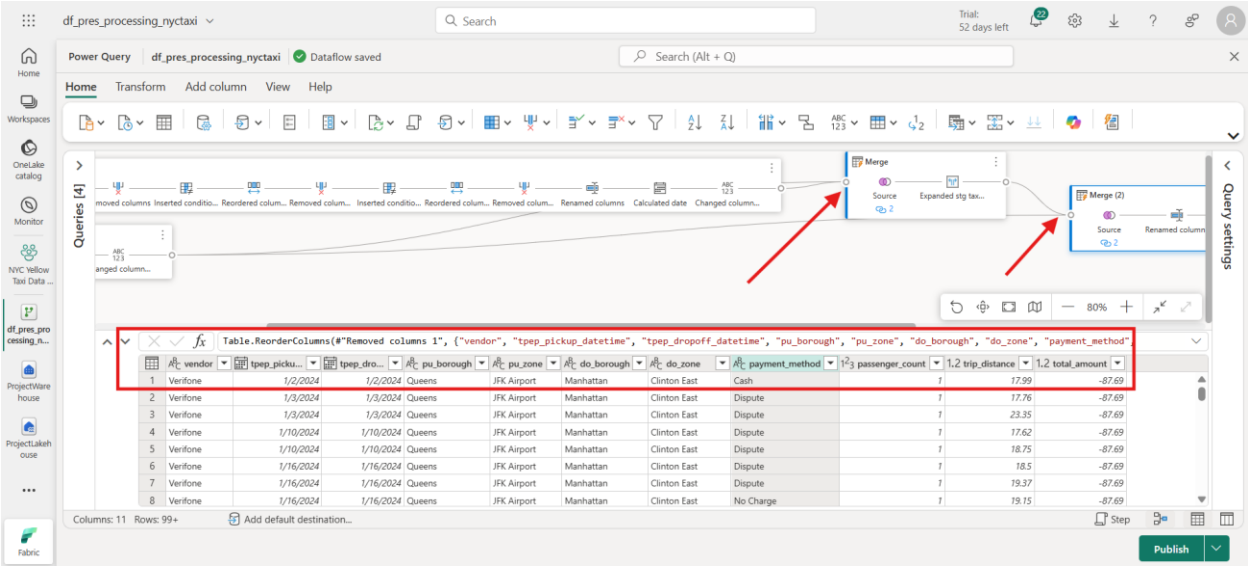
After clean stg.nyctaxi_yellow table and stg.taxi_zone_lookup table merge both tables using left outer join. pu_location_id(stg.nyctaxi_yellow table) with LocationID(stg.taxi_zonne_lookup table).



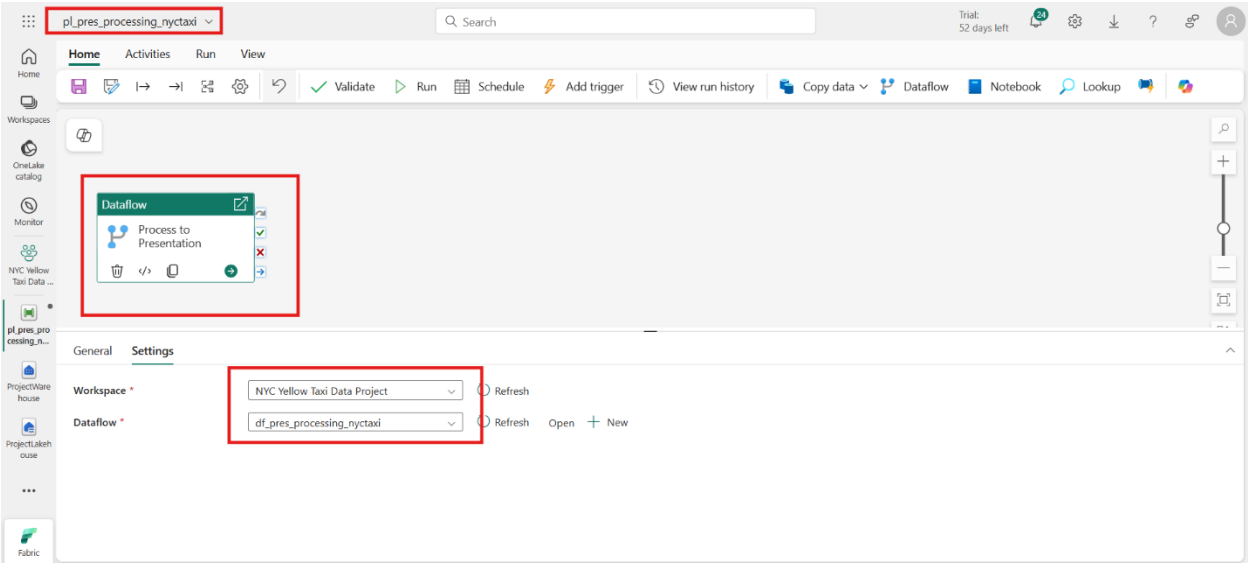
Again, join merged table with stg.taxi_zone_lookup table using left outer join. do_location_id(merged table) with LocationID(stg.taxi_zone_lookup).



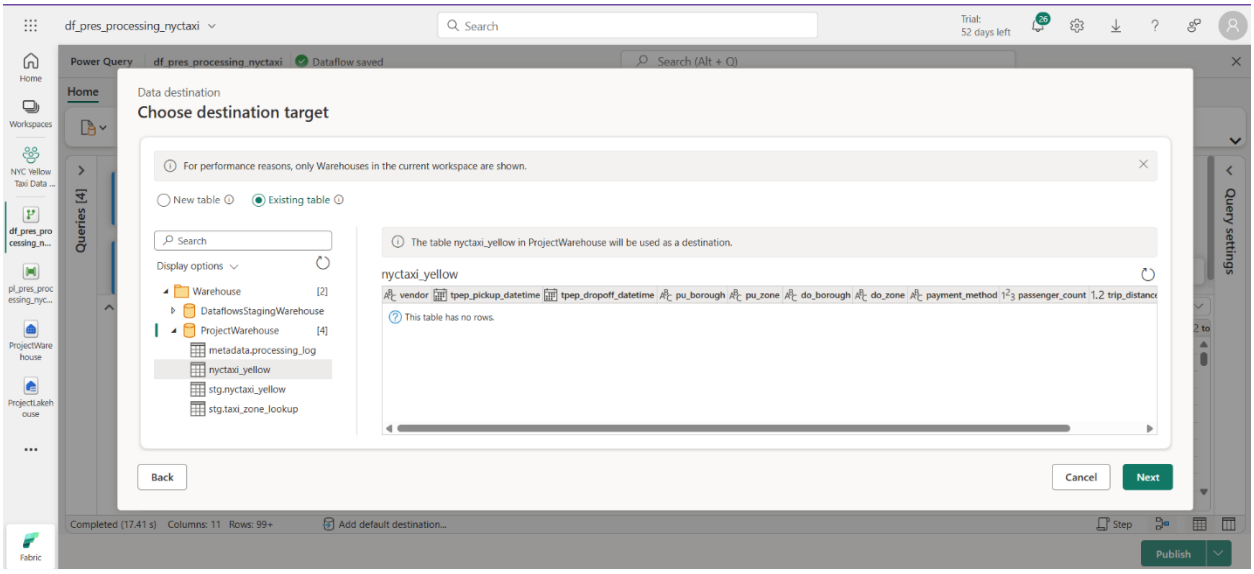
Final look like this,



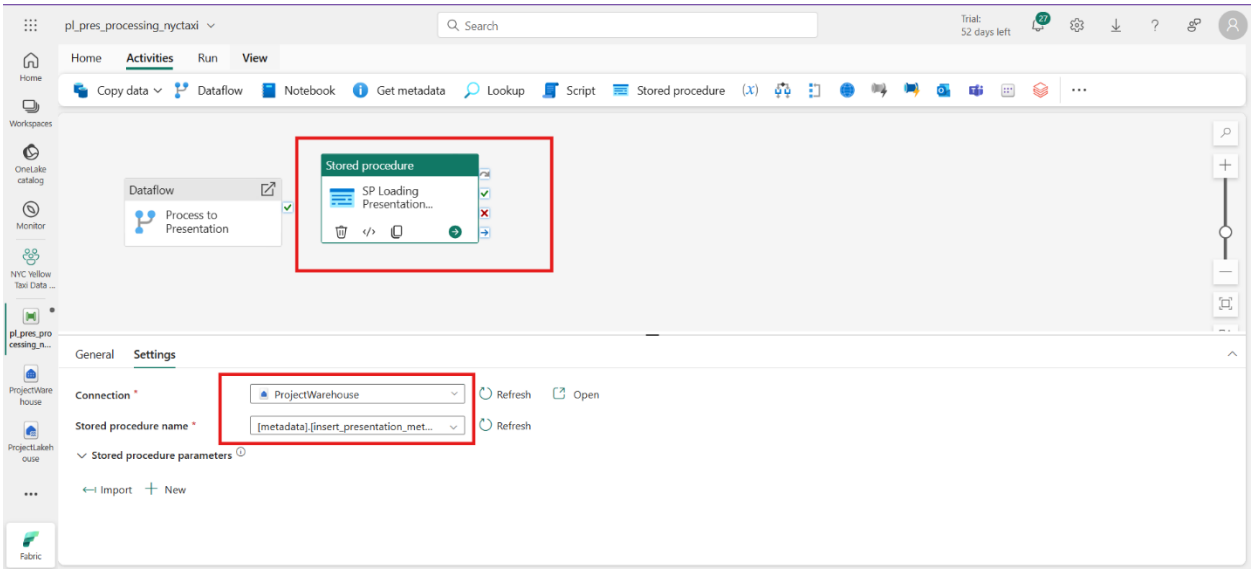
Back to data pipeline folder and create new pipeline named as ‘pl_pres_processing_nyctaxi’. Create Dataflow activity and provide relevant details to the setting pane.



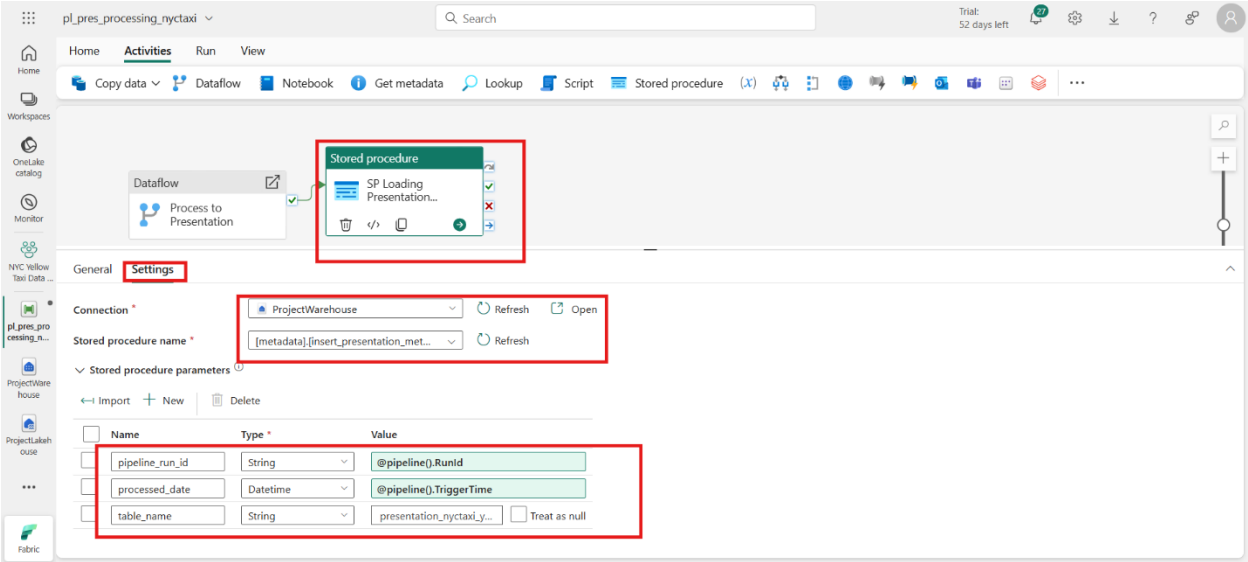
Select destination from data warehouse as nyctaxi_yellow table to store cleansed data.



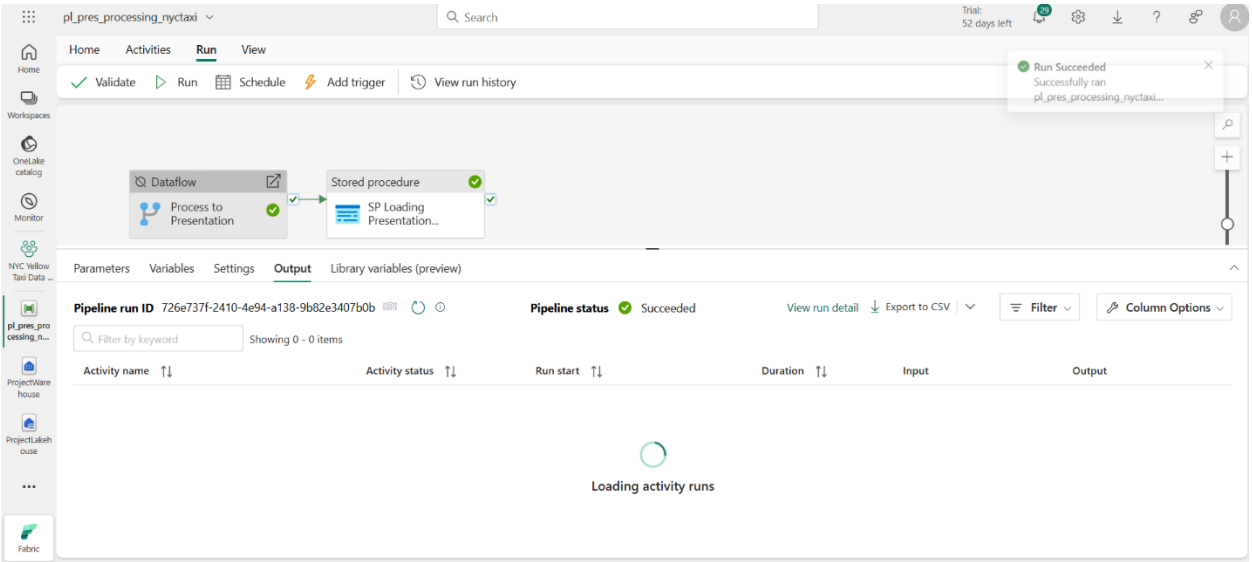
Create Stored procedure activity as shown in the figure.



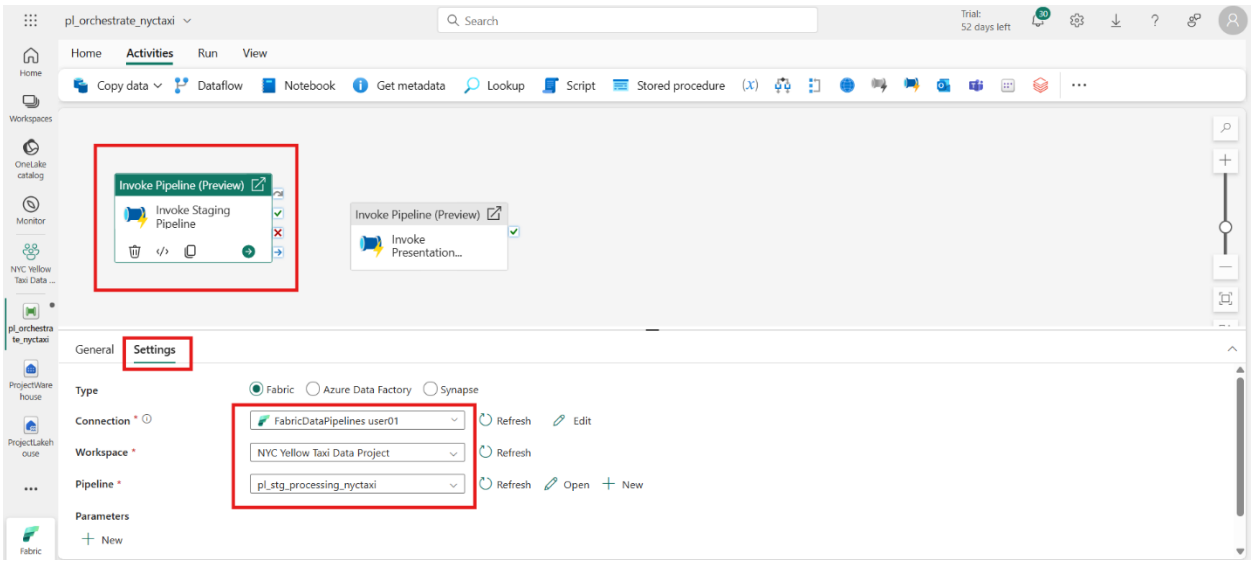
Provide further details to the setting tab as shown in the below figure.

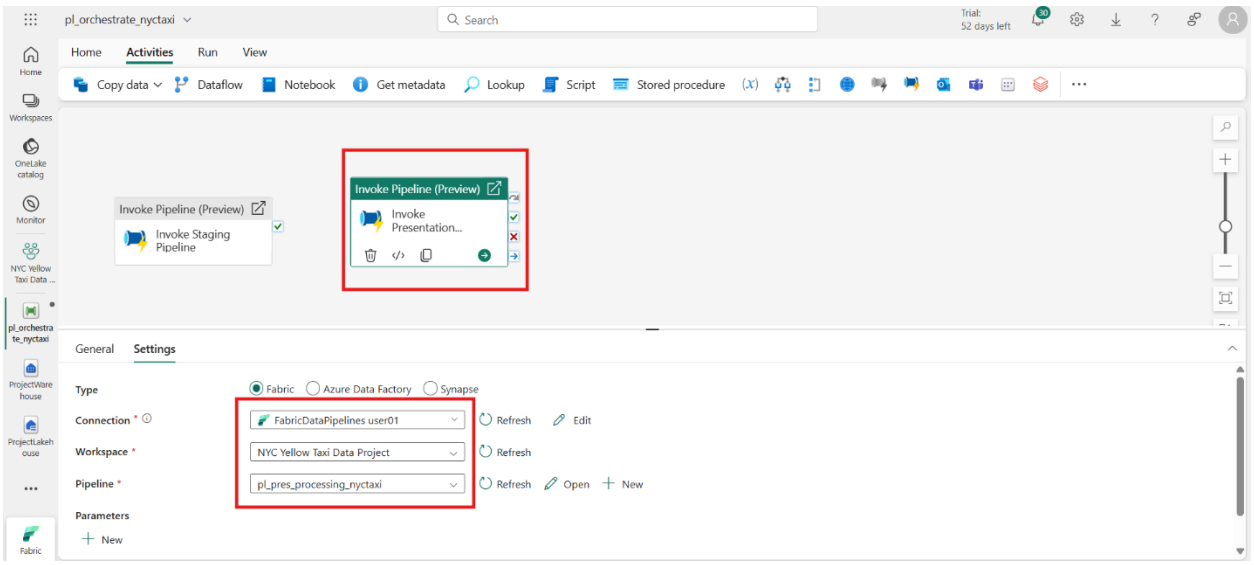


Run this pipeline and get successfully run message.

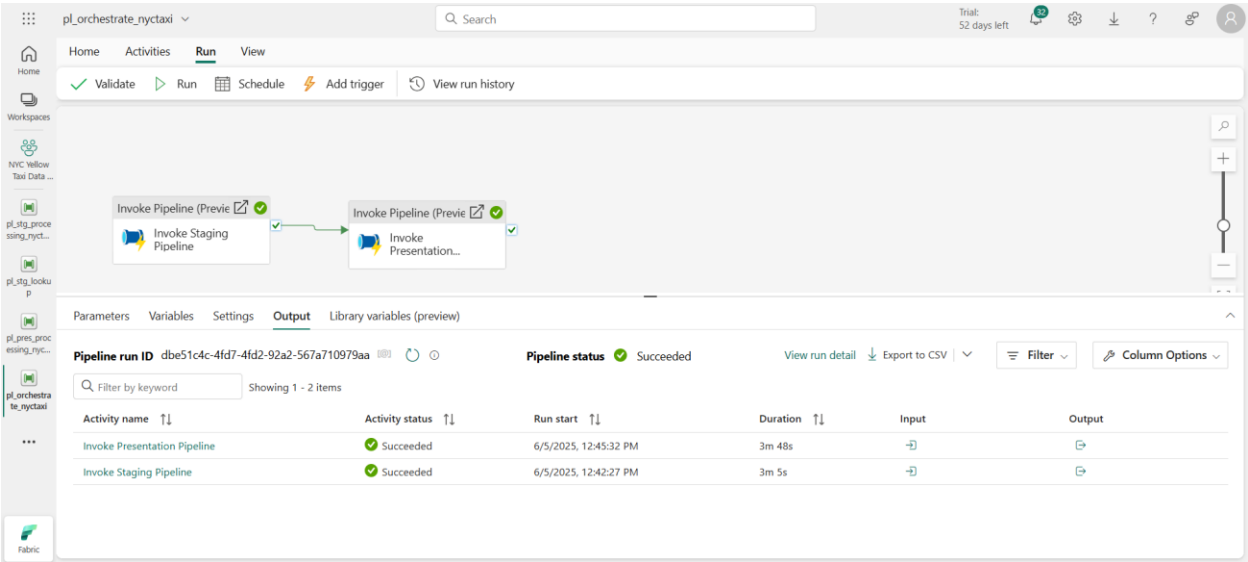


Create final orchestration pipeline to invoke pl_stg_processing_nyctaxi pipeline and pl_pres_processing_nyctaxi pipeline.

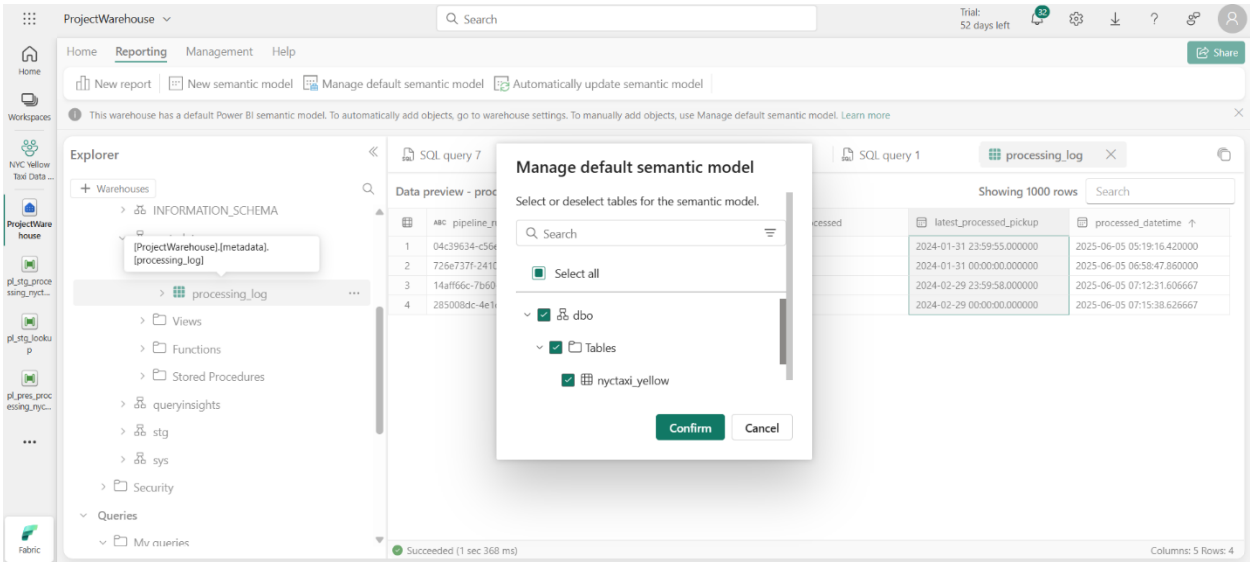




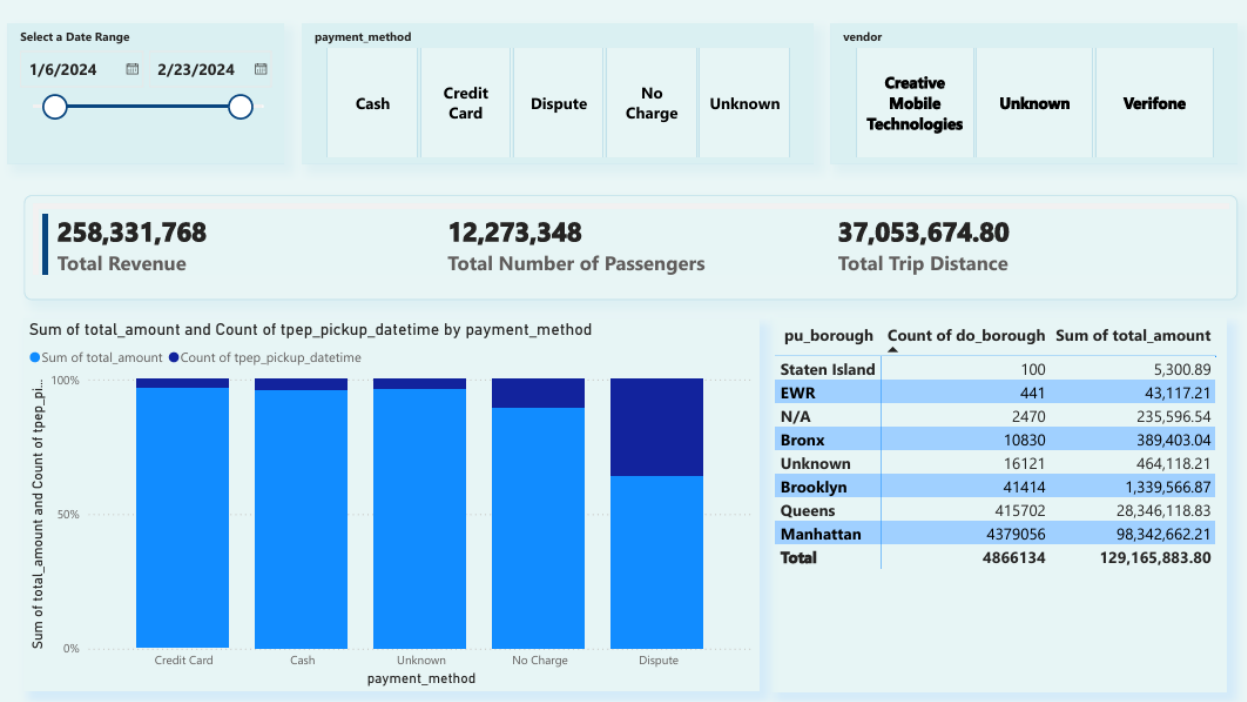
Final orchestration pipeline ran successfully.



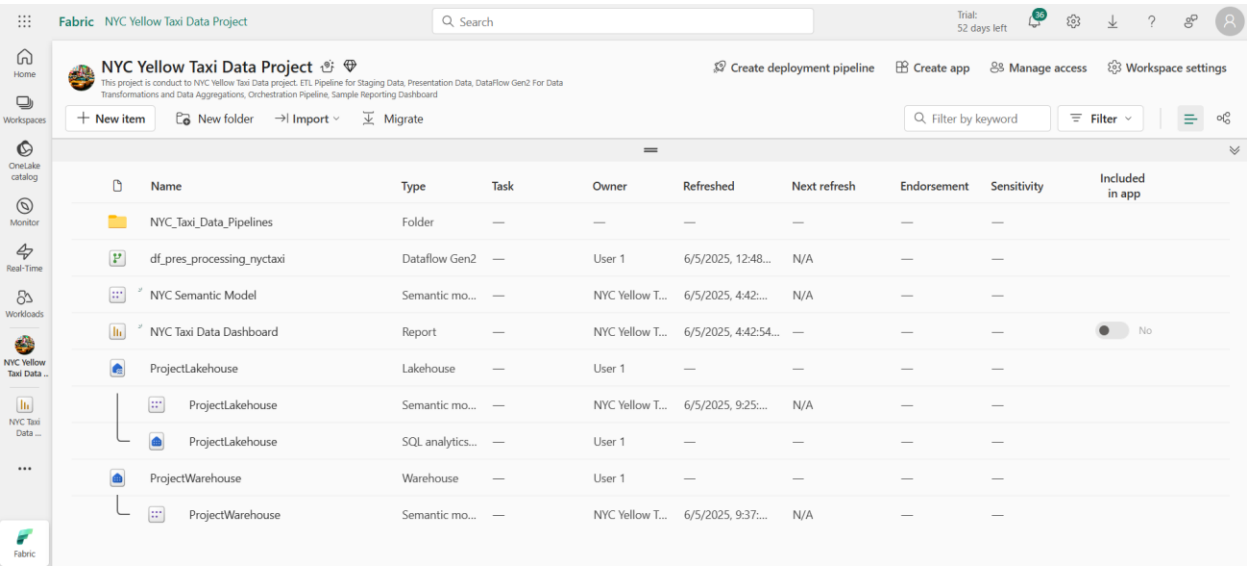
Create New semantic model and create power BI report using dbo.nyctaxi_yellow table.



Created simple dashboard looks like this,



Final fabric workspace looks like this,



End