Databricks Data Engineering Assignment

- SHERIN L

Table of Contents

[1. Overview 3](#_Toc217477414)

[2. Architecture Summary 3](#_Toc217477415)

[3. Tasks Overview 3](#_Toc217477416)

[A. Configure Unity Catalog with metastore, catalog, schema, external locations. 4](#_Toc217477417)

[B. Lakeflow Connect - Raw Data Ingestion 8](#_Toc217477418)

[B1. Use Lakeflow Connect to ingest raw/orders data into Bronze 8](#_Toc217477419)

[B2. Use Auto Loader to ingest raw/customers data. 9](#_Toc217477420)

[C. Medallion Architecture 10](#_Toc217477421)

[C1. Build Silver transformations with cleansing & type corrections. 10](#_Toc217477422)

[C2. Build Gold aggregations such as daily sales summary. 12](#_Toc217477423)

[D. Delta Live Tables Pipeline 13](#_Toc217477424)

[D1. Create a Delta Live Tables pipeline with expectations. 13](#_Toc217477425)

[E. Data Quality & Data Type Checks 14](#_Toc217477426)

[E1. Implement DQ checks: null, datatype mismatch, pattern, outliers. 14](#_Toc217477427)

[F. Lakebridge Analyzer & Transpiler 16](#_Toc217477428)

[F1. Run Lakebridge Analyzer on Oracle scripts 16](#_Toc217477429)

[F2. Run Lakebridge Transpiler & integrate output. 17](#_Toc217477430)

# 1. Overview

This document outlines the end-to-end design and implementation plan for a modern data engineering pipeline using Databricksand to design, build, and operate a scalable data ingestion and transformation pipeline on Databricks including Lakeflow Connect, Medallion Architecture, Auto Loader, Delta Live Tables, Unity Catalog, and Lakebridge tools.

# 2. Architecture Summary

The platform is designed with the following layers:

* **Ingestion Layer**: Raw data ingestion using Lakeflow Connect and Auto Loader
* **Storage Layer**: Delta Lake with Bronze, Silver, and Gold layers
* **Processing Layer**: Spark & Delta Live Tables (DLT)
* **Governance Layer**: Unity Catalog for access control and metadata
* **Quality Layer**: Data Quality (DQ) checks and expectations
* **Migration Utilities**: Lakebridge Analyzer and Transpiler

# 3. Tasks Overview

A. Unity Catalog Setup  
B. Ingestion (Lakeflow + Autoloader)  
C. Medallion Architecture  
D. DLT Pipeline  
E. Data Quality & Data Type Checks  
F. Lakebridge Analyzer & Transpiler  
G. Final Delivery

# A. Configure Unity Catalog with metastore, catalog, schema, external locations.

Figure 1 creating unity catalog

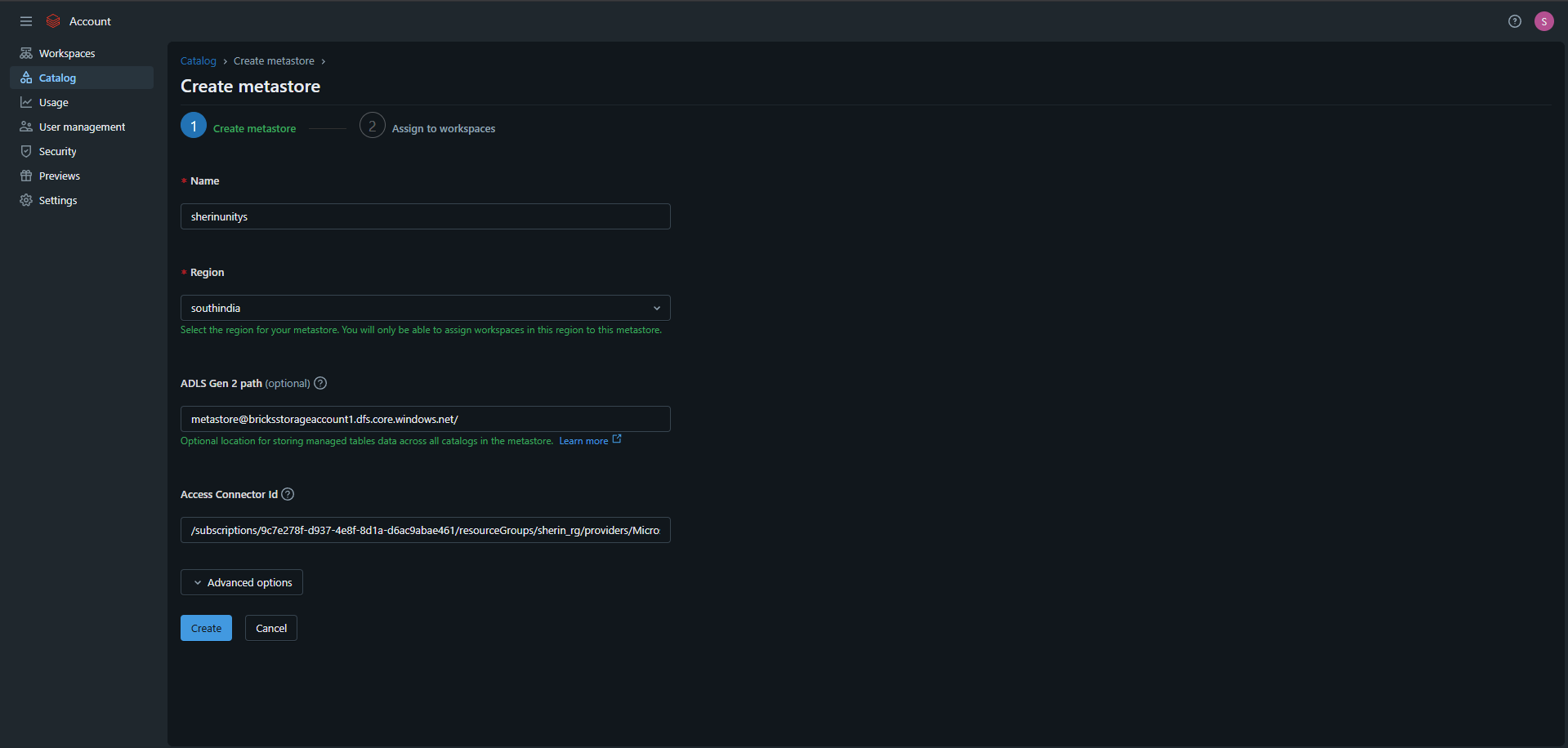


Figure 2 created unity catalog

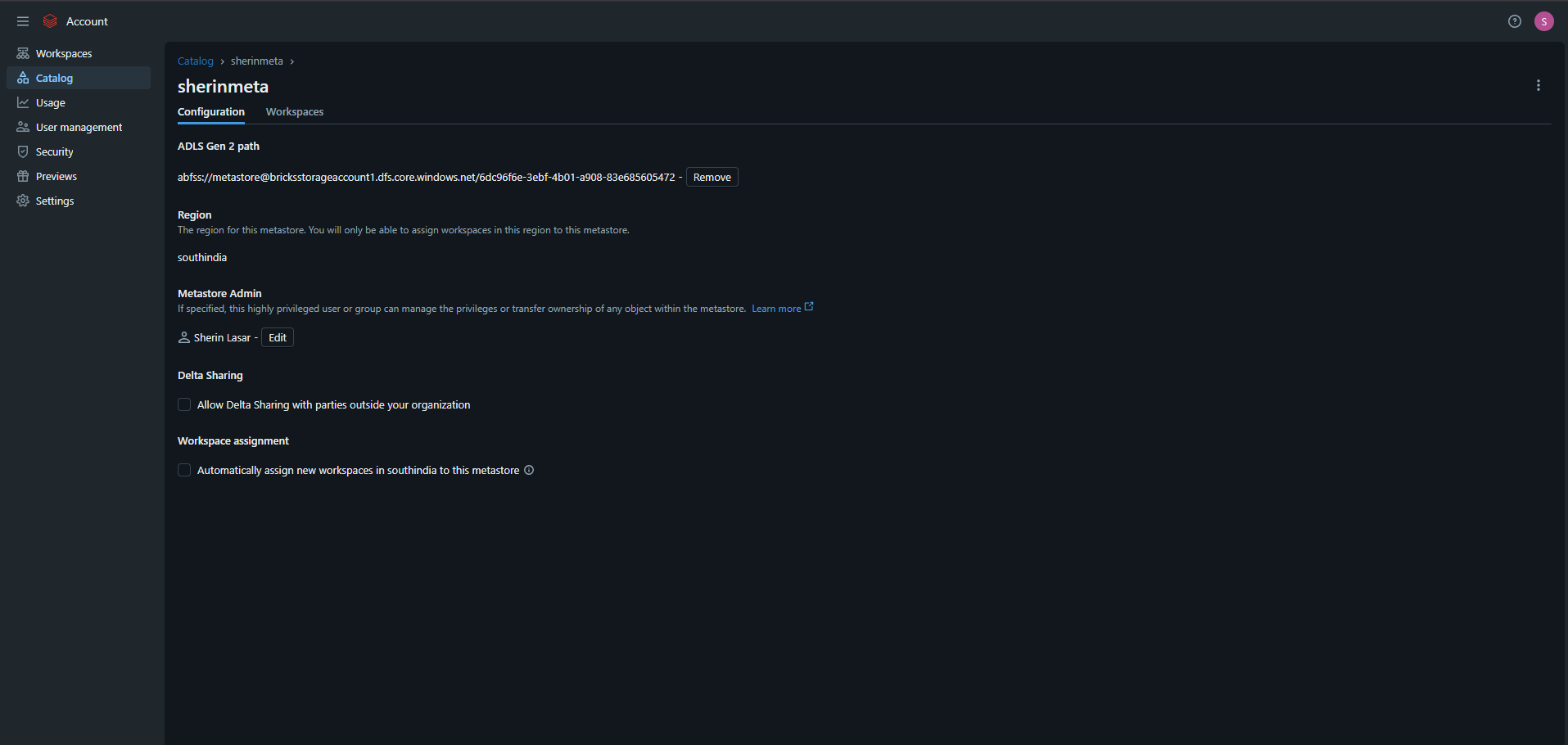


Figure 3 creating catalog



Figure 4 creating external location

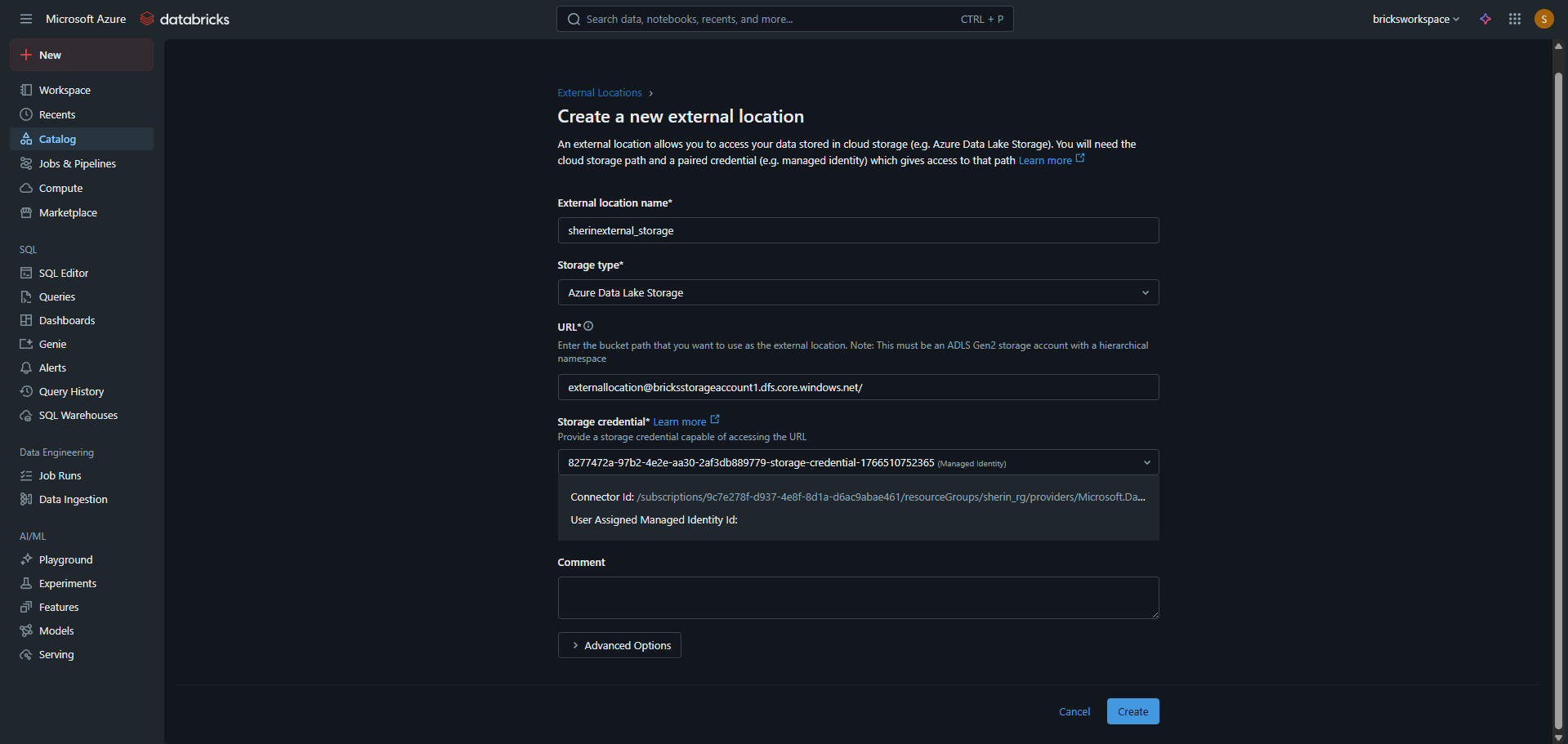


Figure 5 external location created

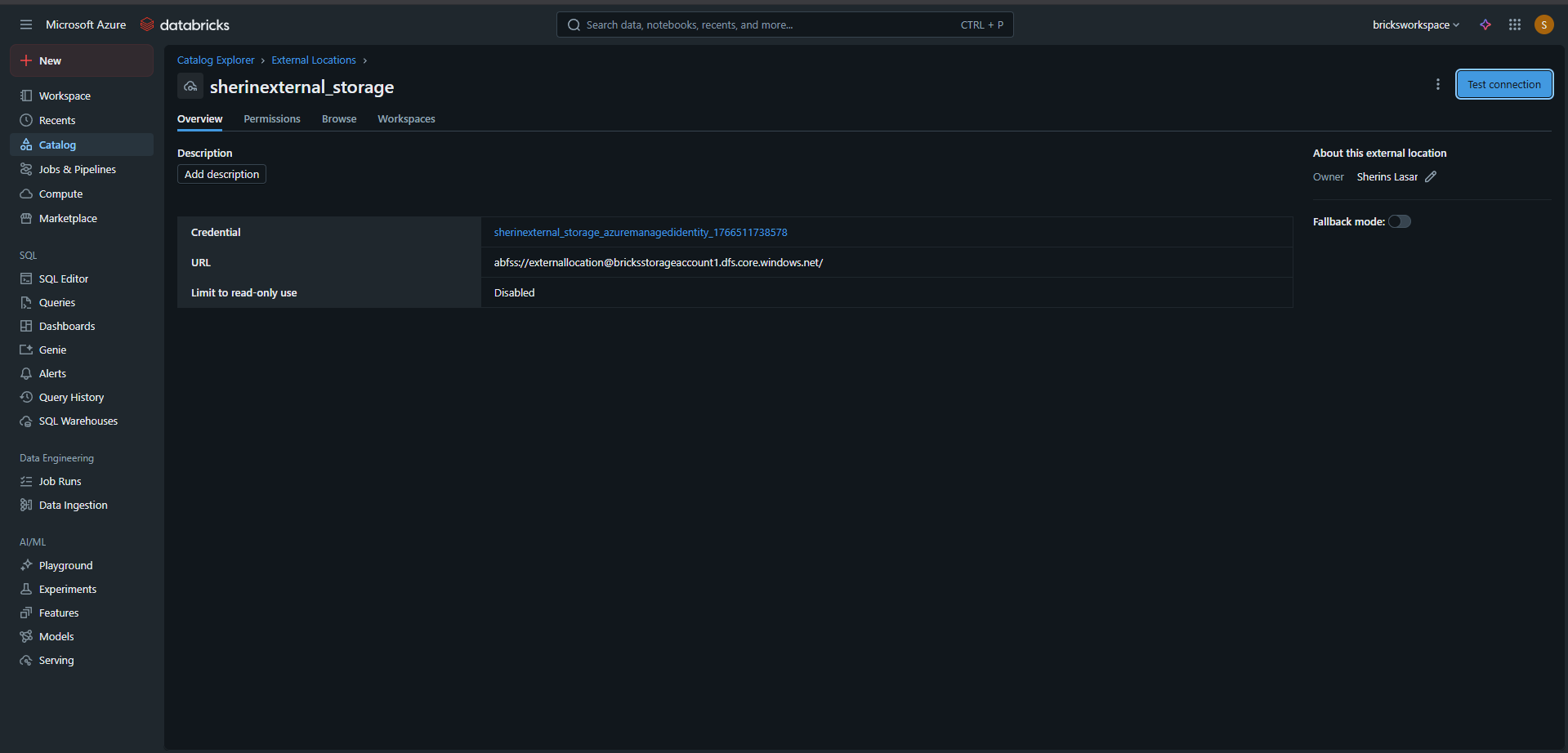


Figure 6 creating schemas-BRONZE

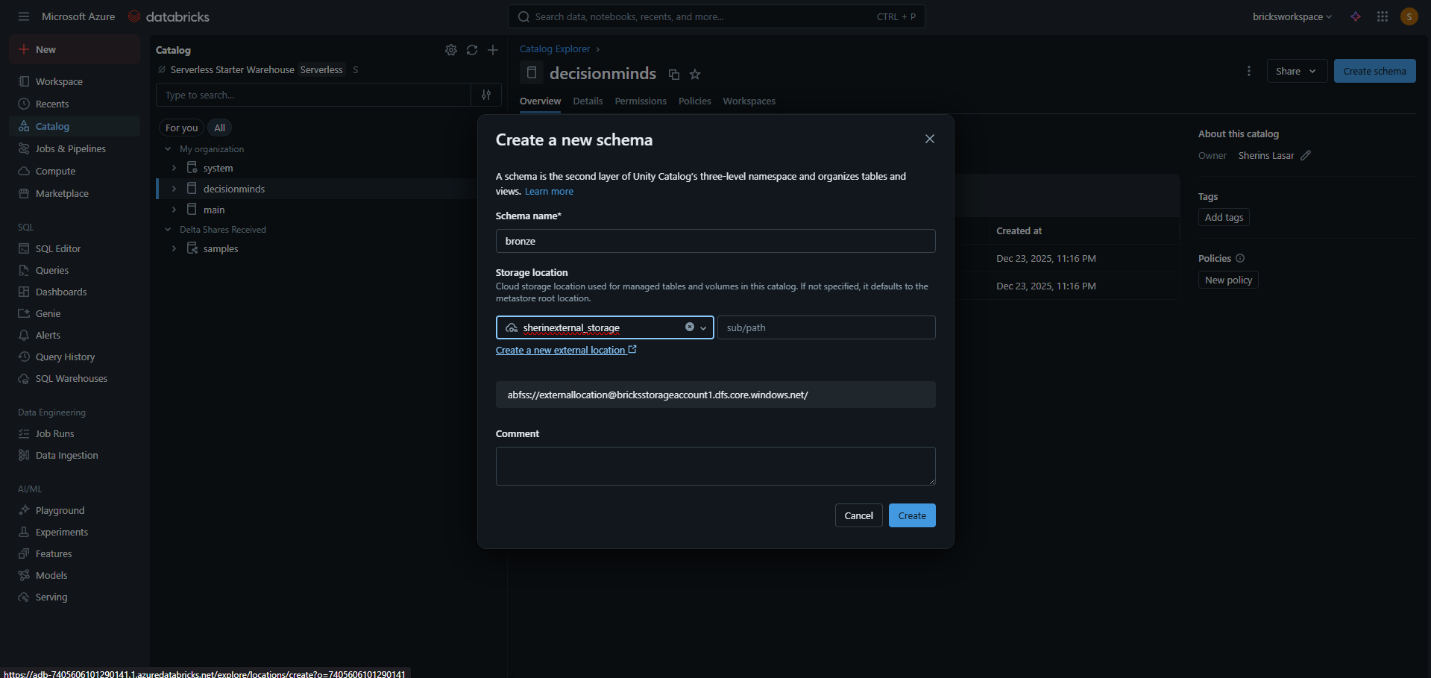


Figure 7 creating schema-SILVER

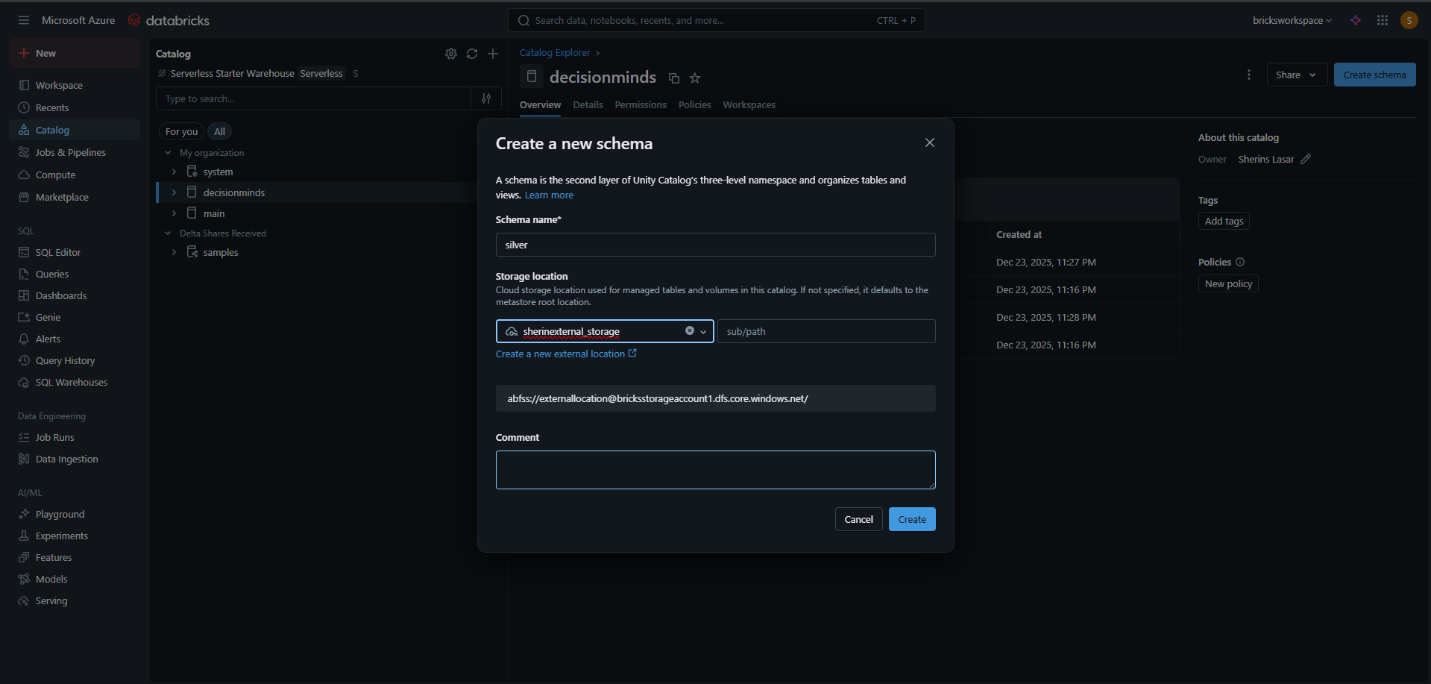
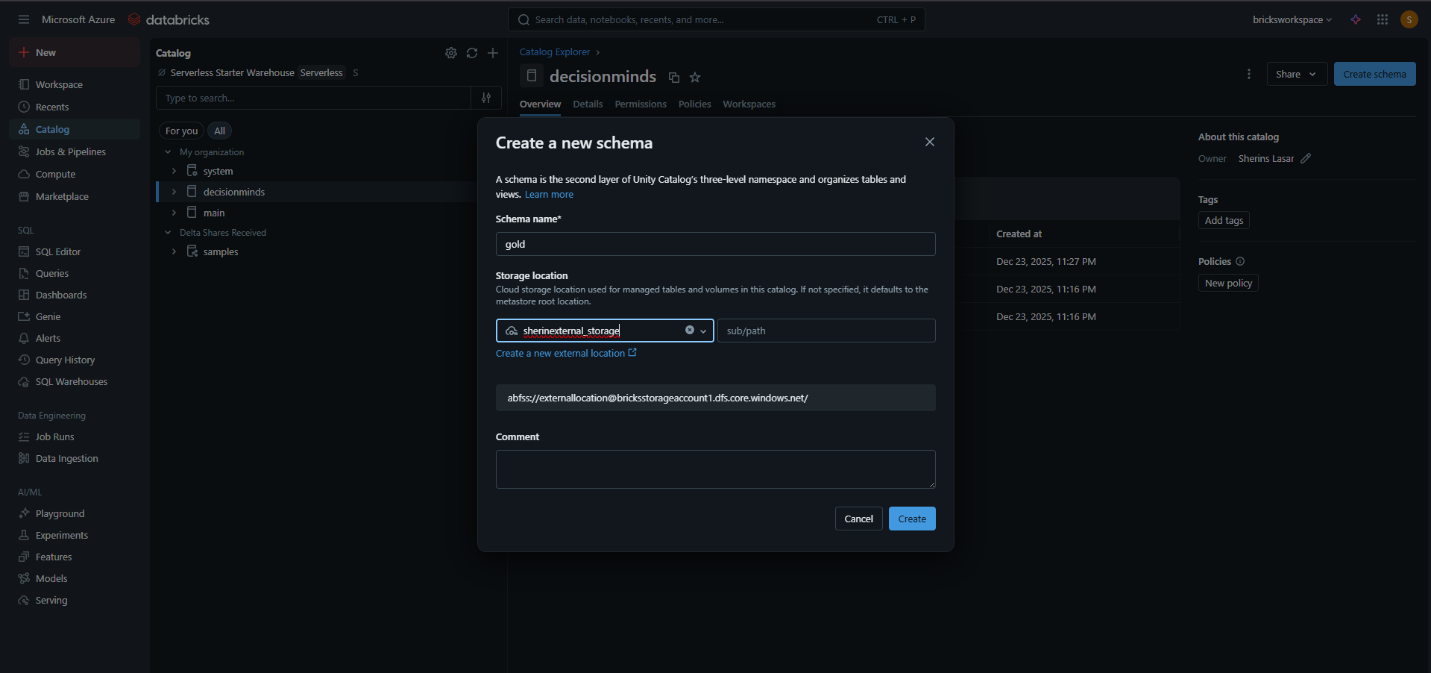


Figure 8 creating schema-GOLD



# B. Lakeflow Connect - Raw Data Ingestion

B1. Use Lakeflow Connect to ingest raw/orders data into Bronze.

Figure 9 lakeflow connect with google drive

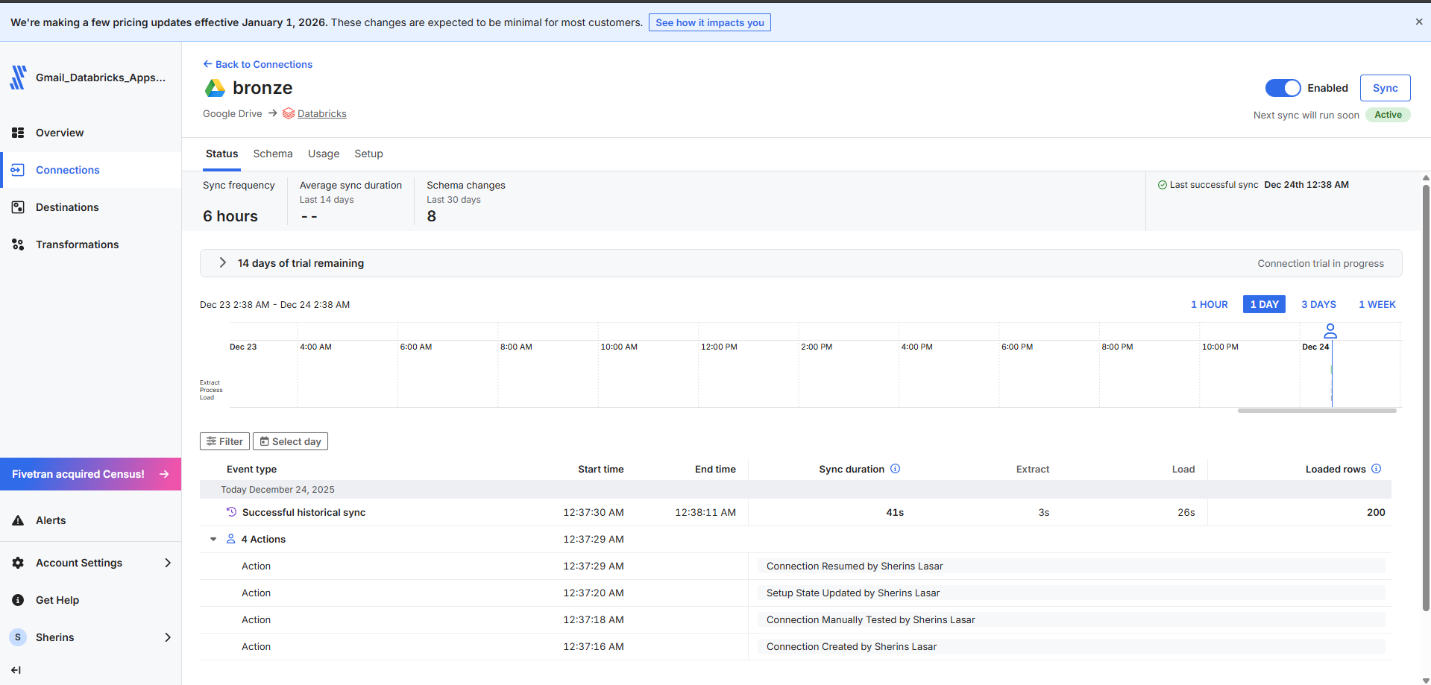
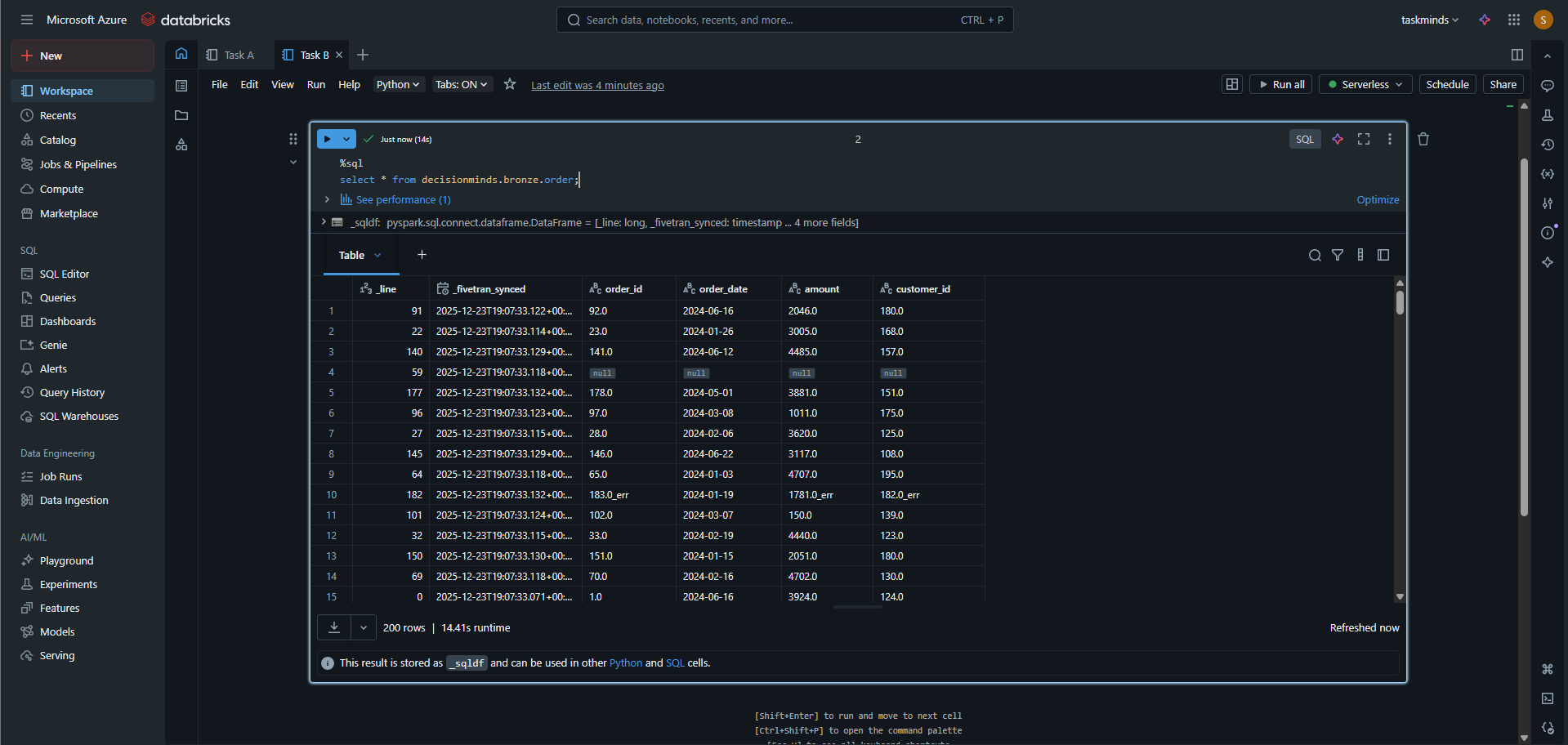


Figure 10 order\_table ingested using lakeflow



## B2. Use Auto Loader to ingest raw/customers data.

Figure 11 autoloader customer table ingestion

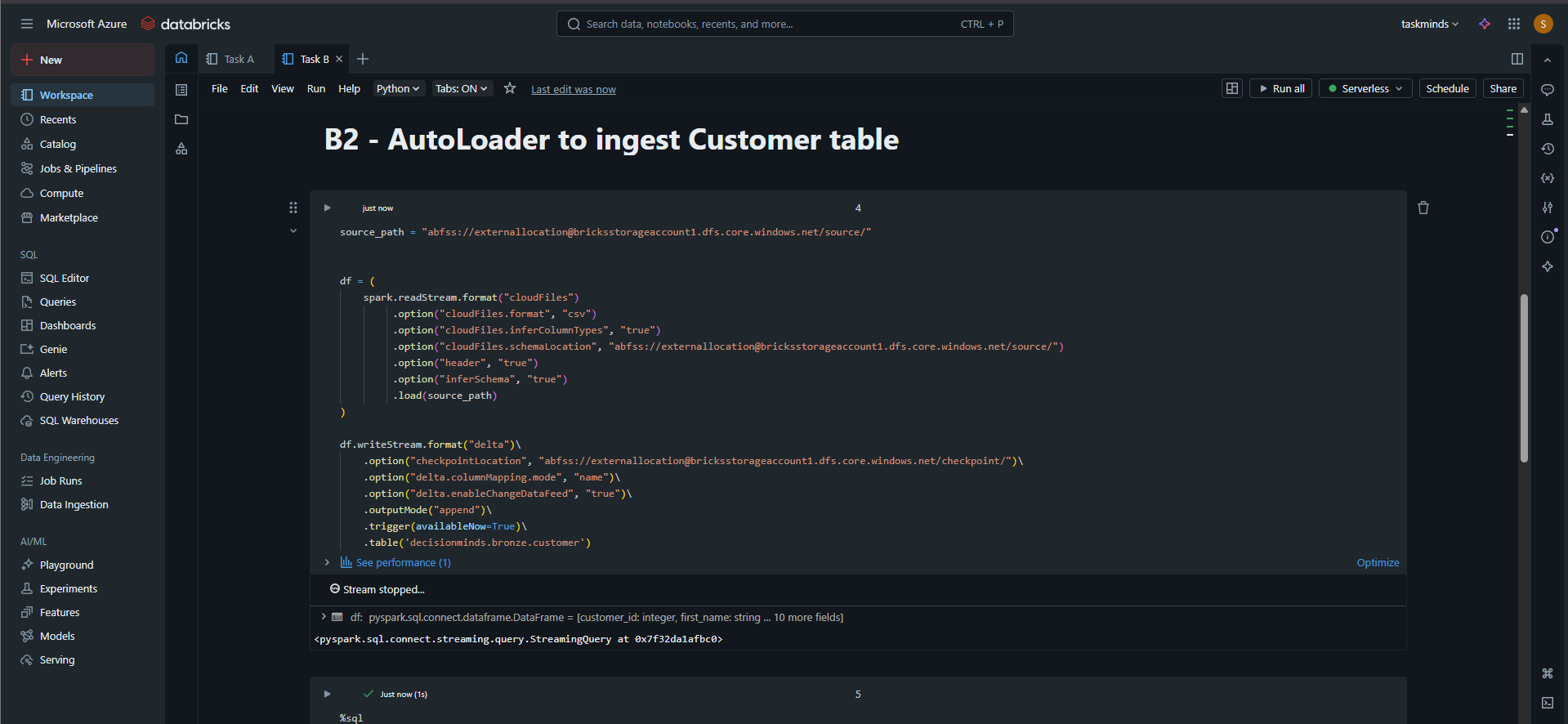
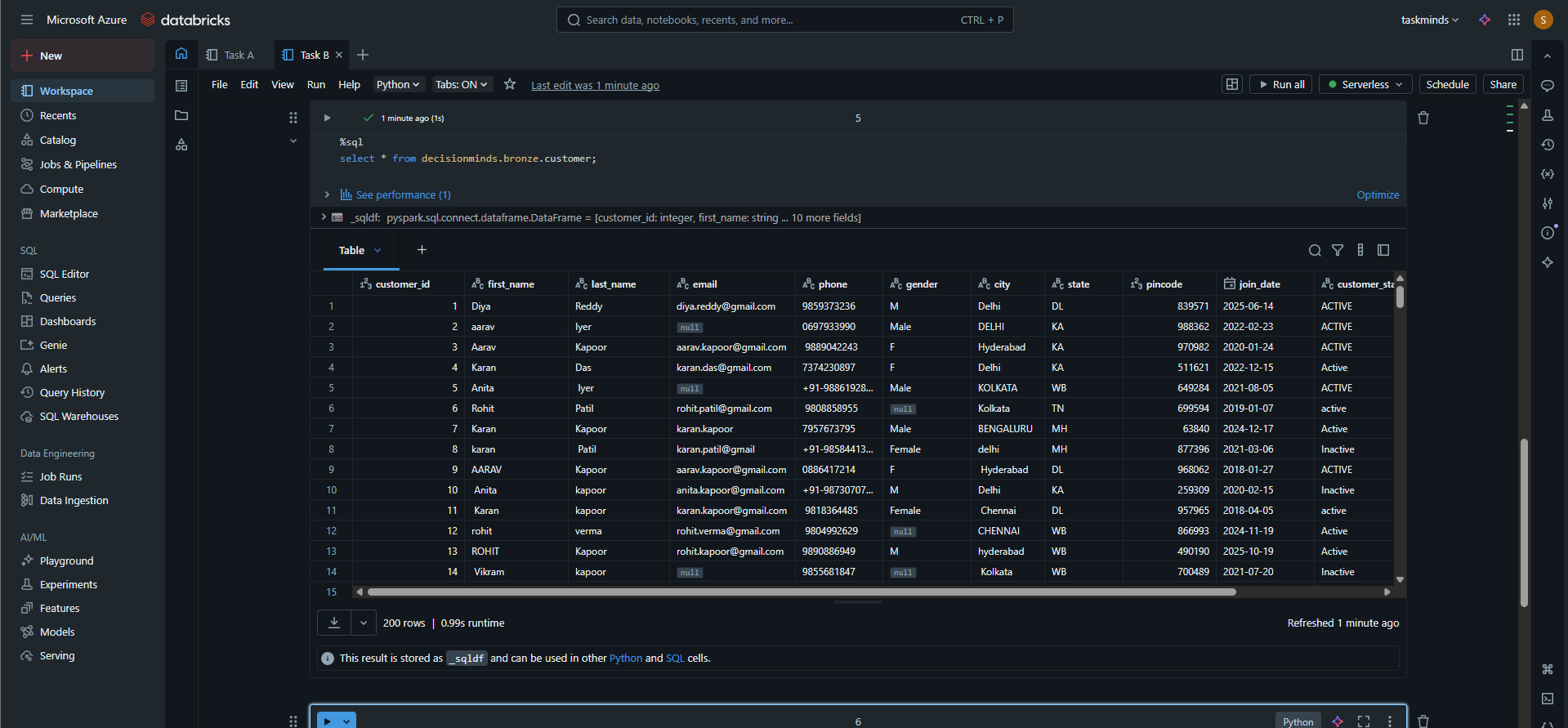


Figure 12 autoloaded customer table



# C. Medallion Architecture

## C1. Build Silver transformations with cleansing & type corrections.

Figure 13 silver order\_table cleaning

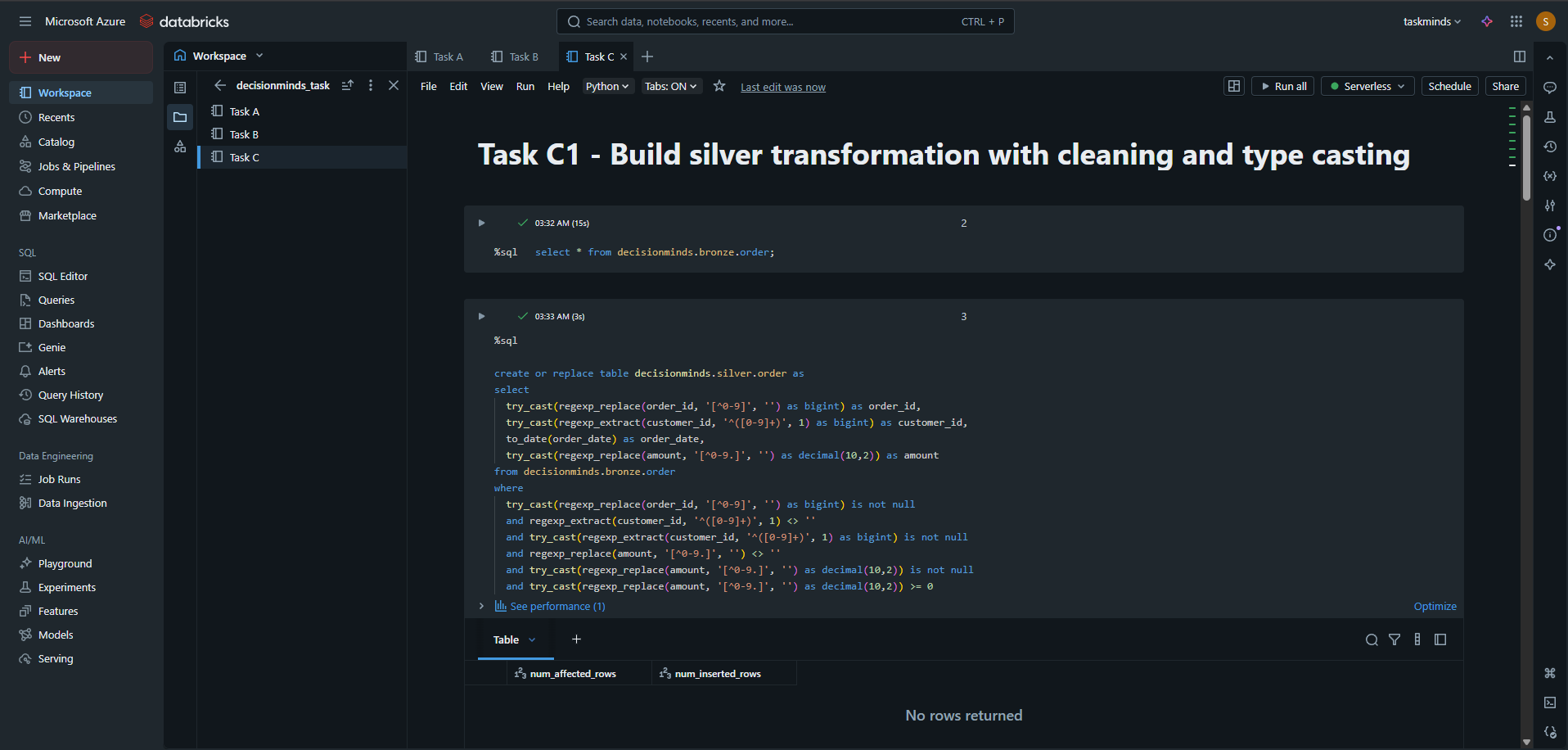


Figure 14 cleaned silver\_table

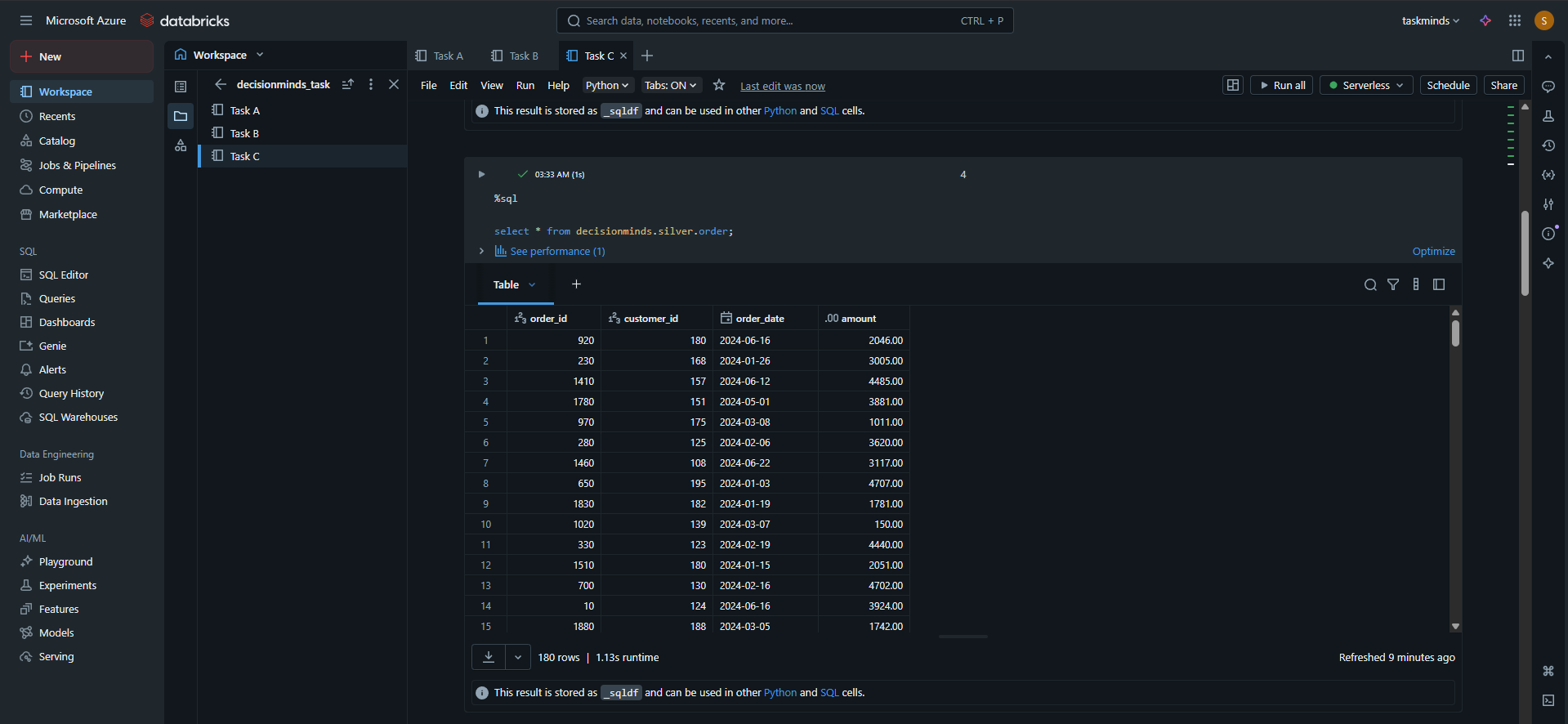


Figure 15 cleaning customer\_table

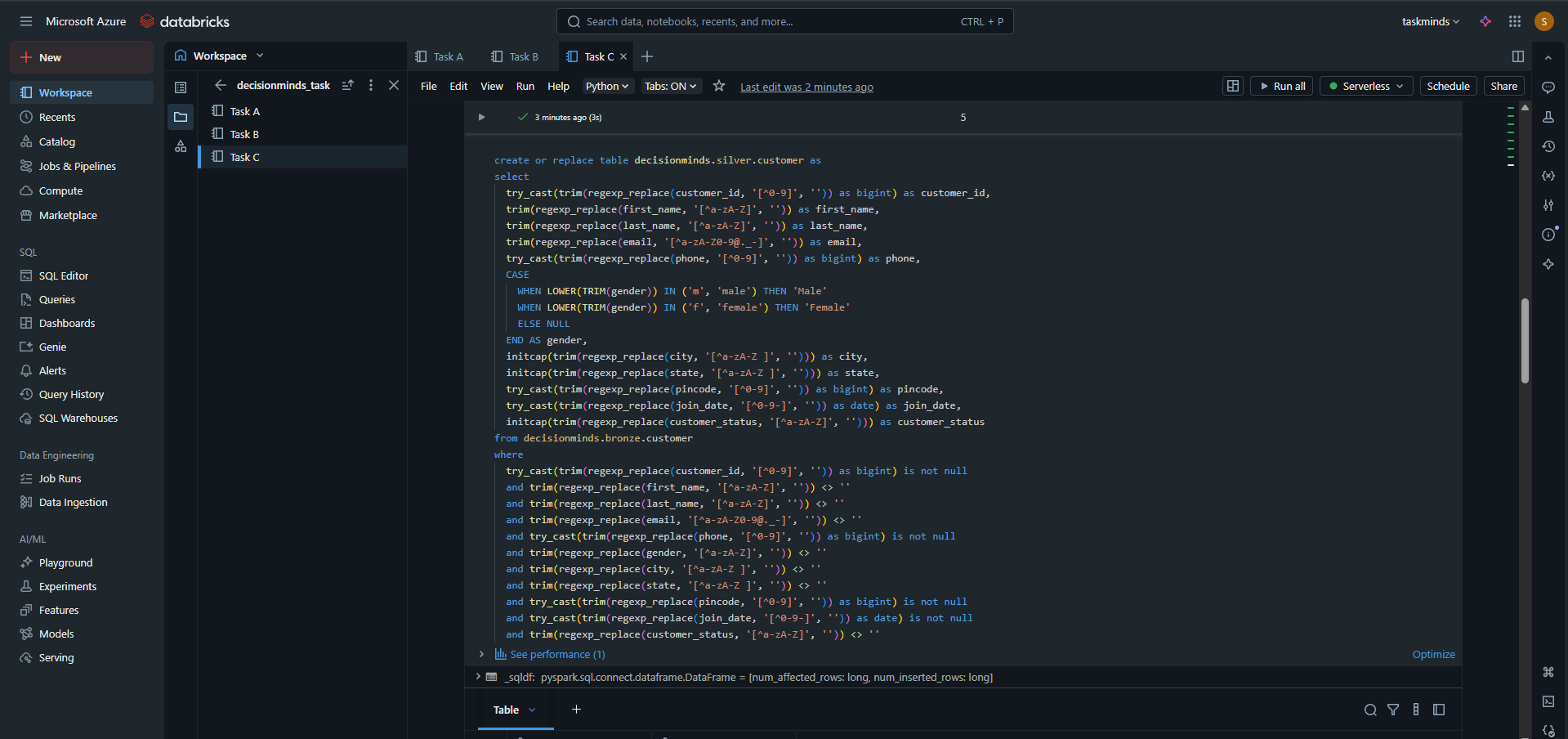
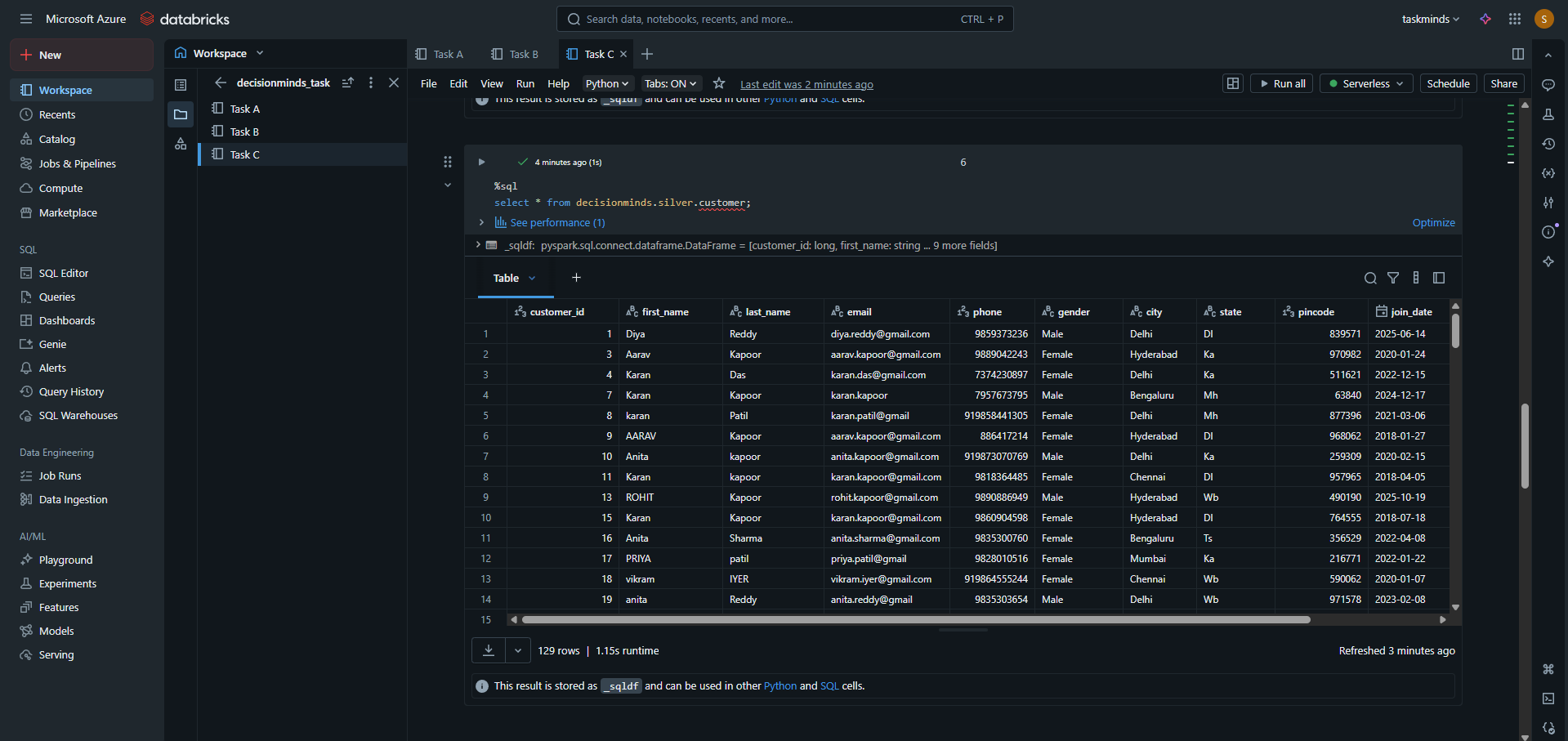


Figure 16 cleaned customer\_table



## C2. Build Gold aggregations such as daily sales summary.

Figure 17 Gold aggregation

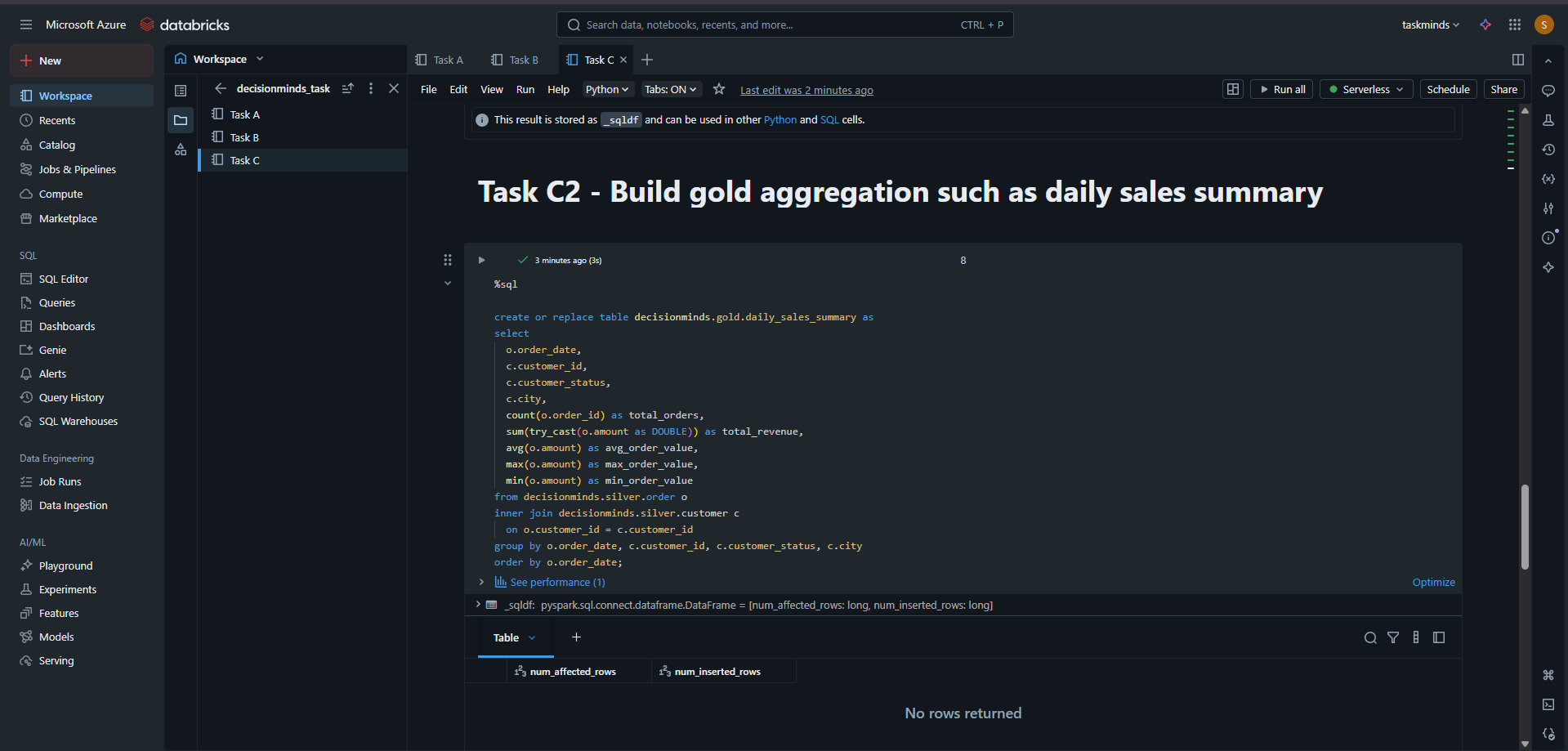
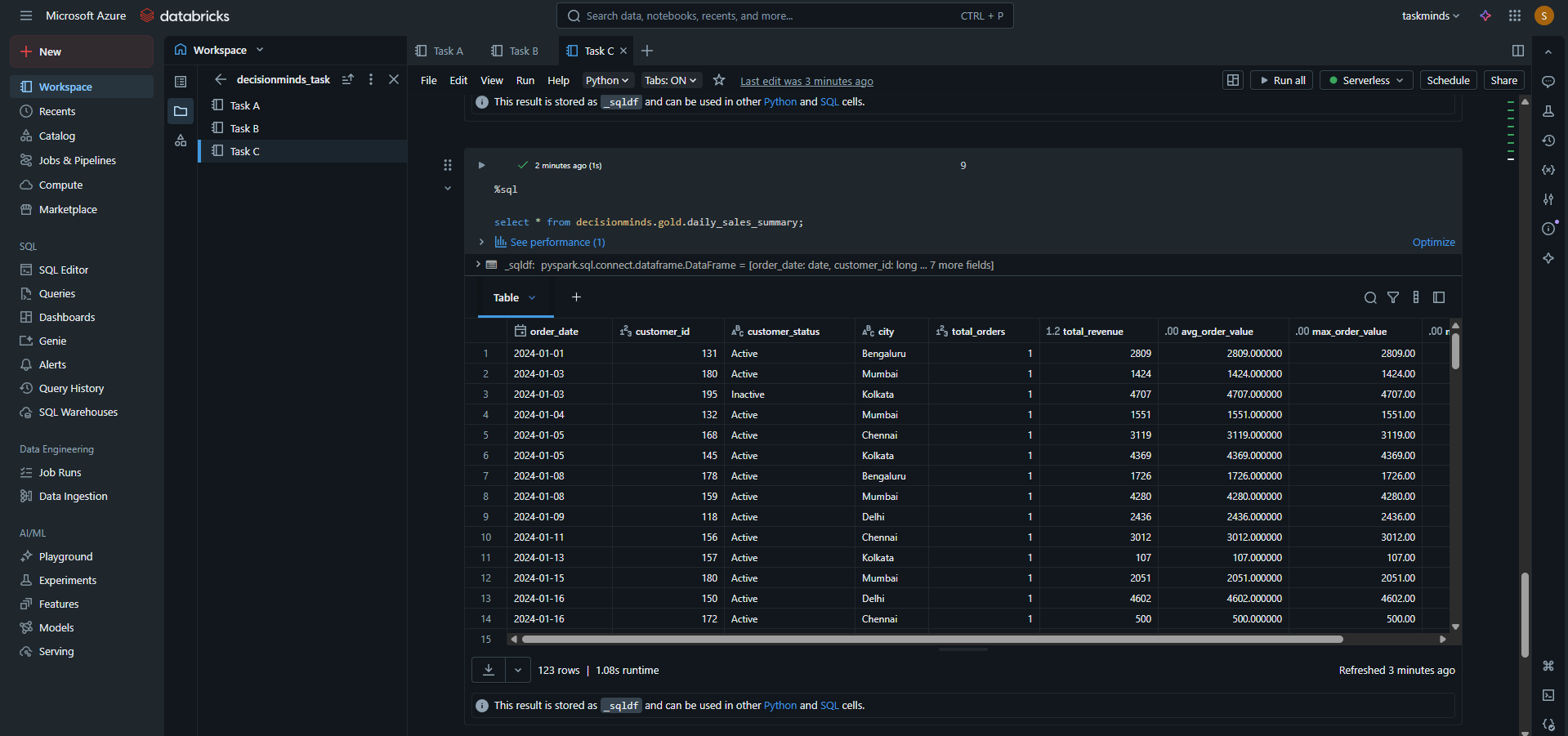


Figure 18 Gold analysis



# D. Delta Live Tables Pipeline

## D1. Create a Delta Live Tables pipeline with expectations.

Figure 19 DLT pipeline

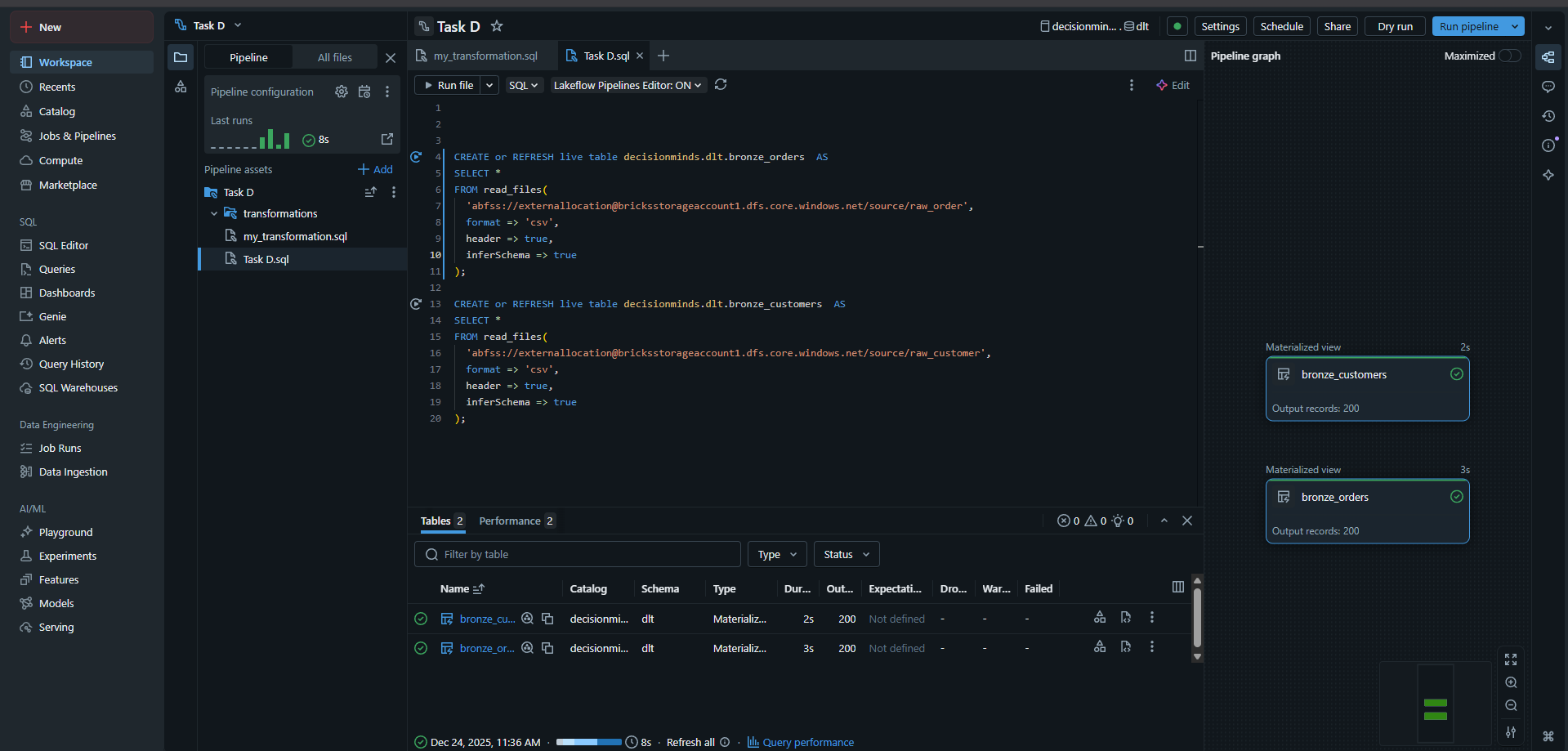


Figure 20 DLT order\_table

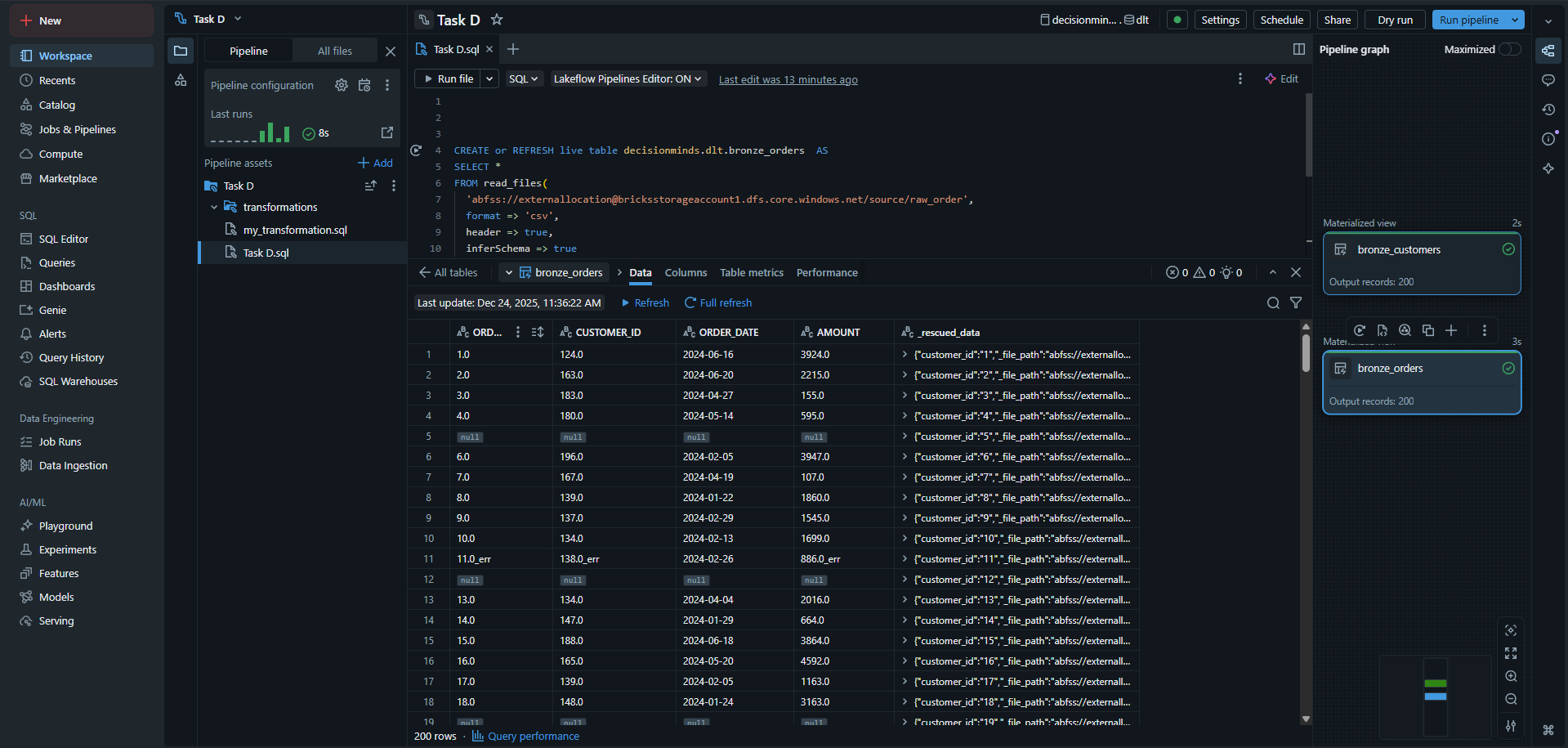
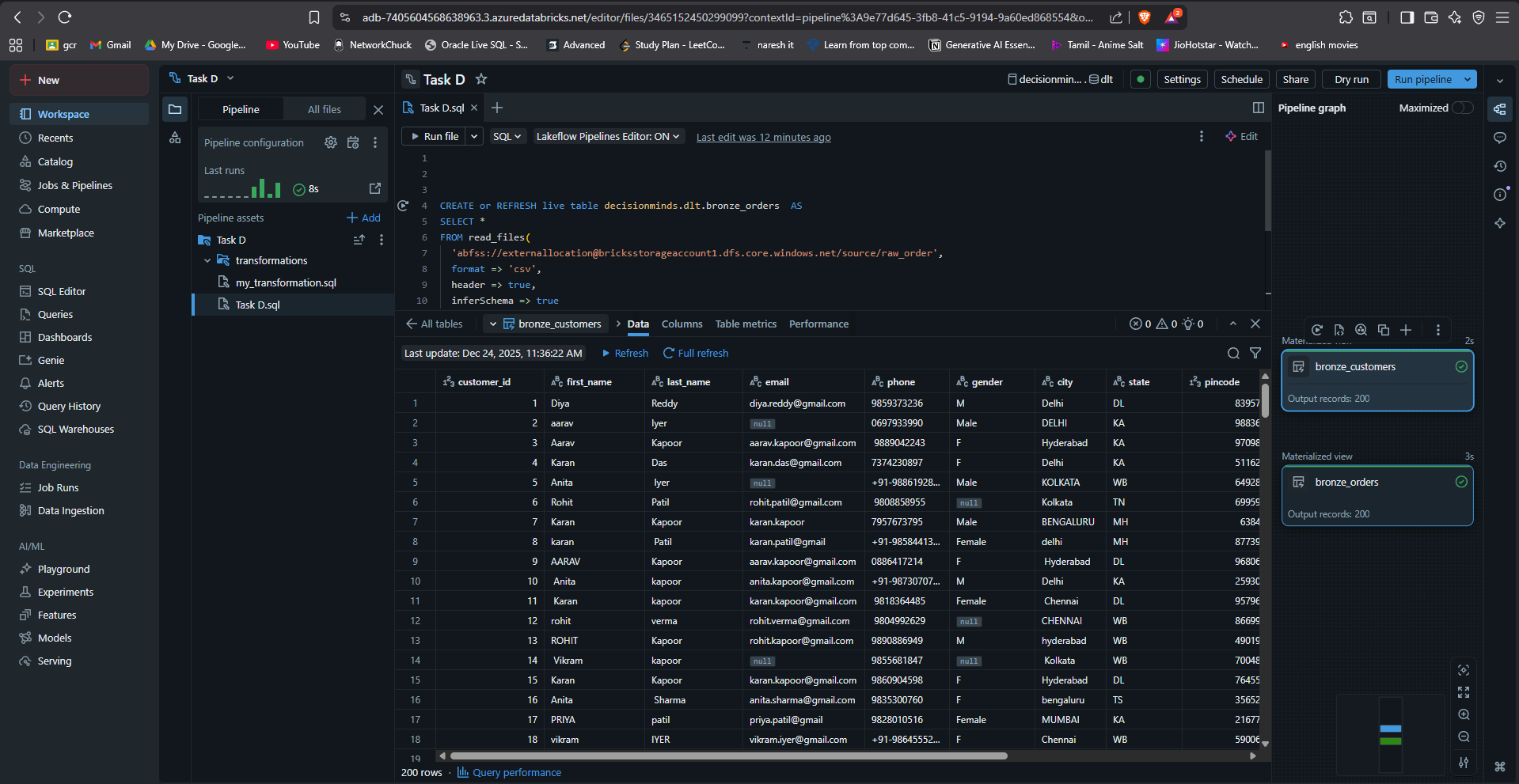


Figure 21 DLT customer table



# E. Data Quality & Data Type Checks

## E1. Implement DQ checks: null, datatype mismatch, pattern, outliers.

Figure 22 DQ check-pipeline

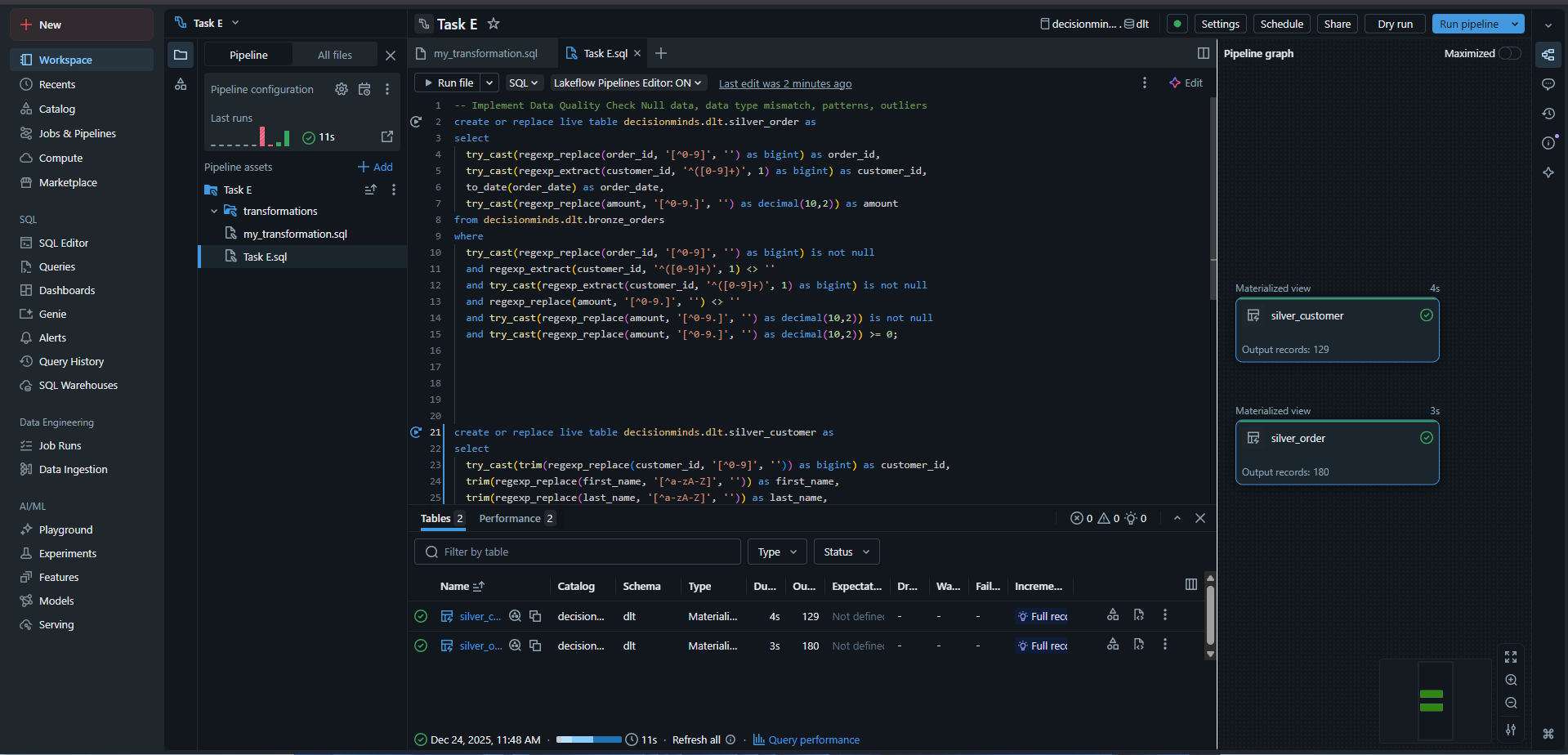
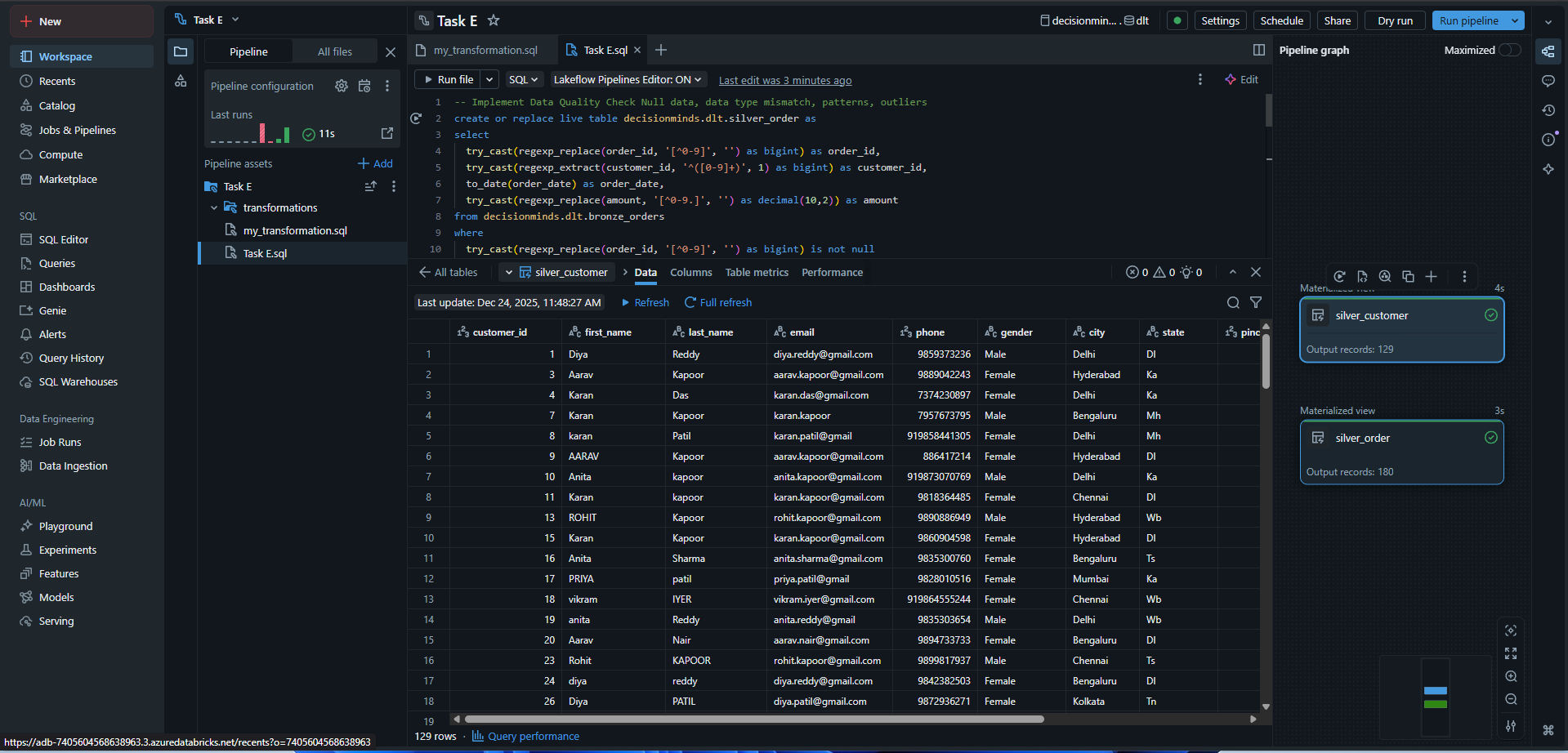


Figure 23 DQ check order\_table

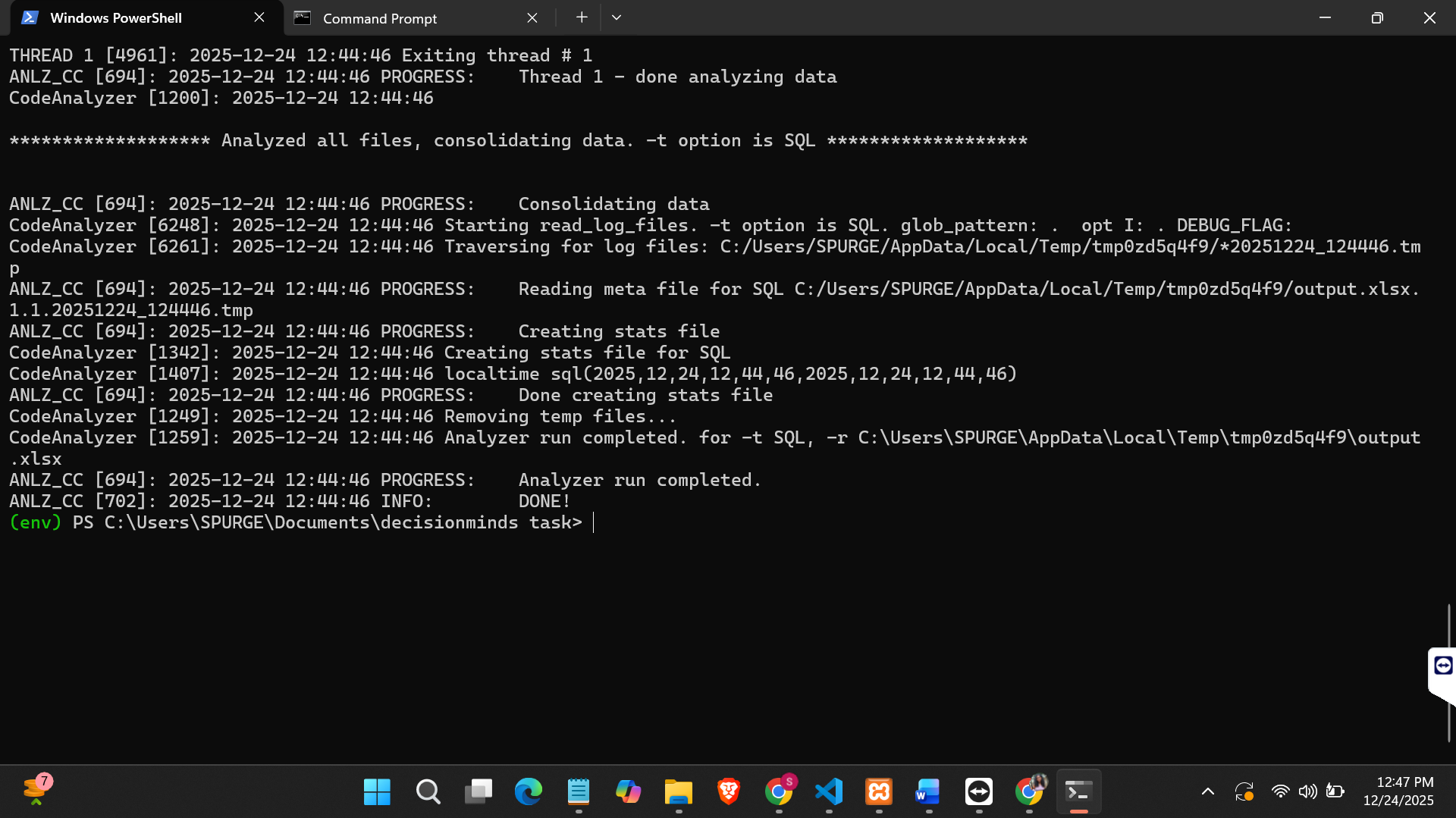
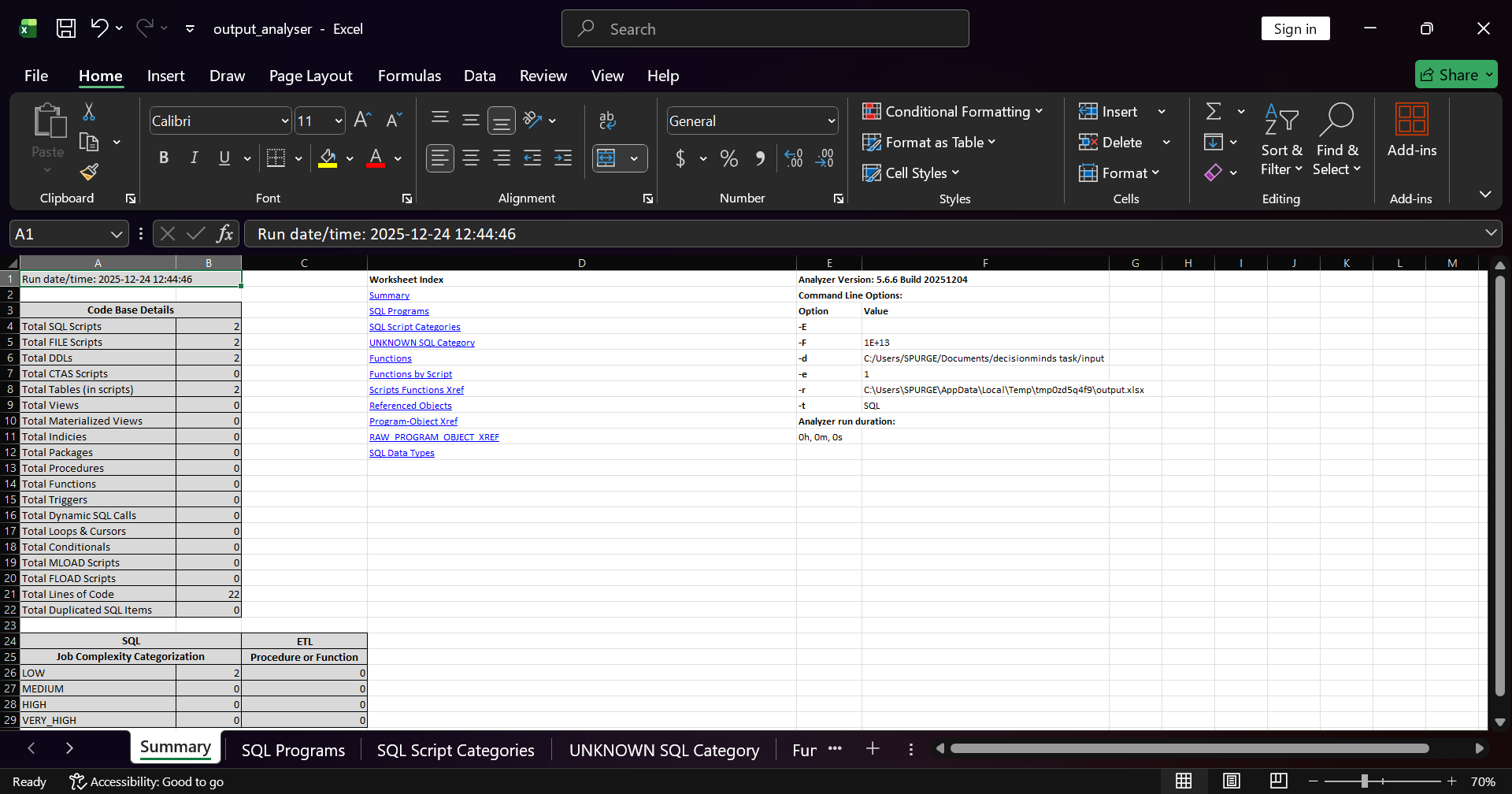


Figure 24 DQ check customer\_table



# F. Lakebridge Analyzer & Transpiler

## F1. Run Lakebridge Analyzer on Oracle scripts

Figure 25 Lakebridge analyzerFigure 26 Analyzer output

## F2. Run Lakebridge Transpiler & integrate output.

Figure 27 Lakebridge transpiler

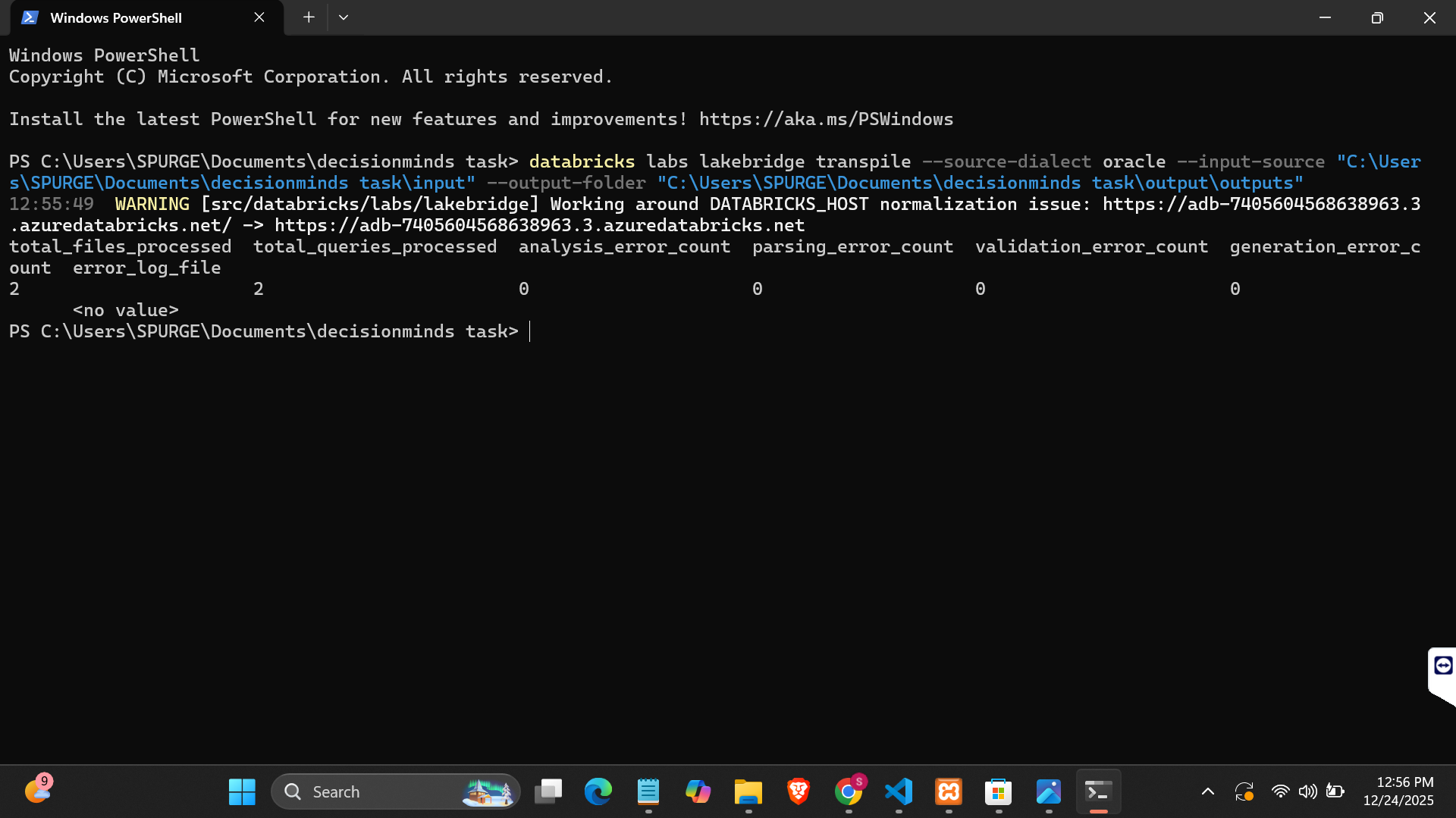


Figure 28 Transplier input-1

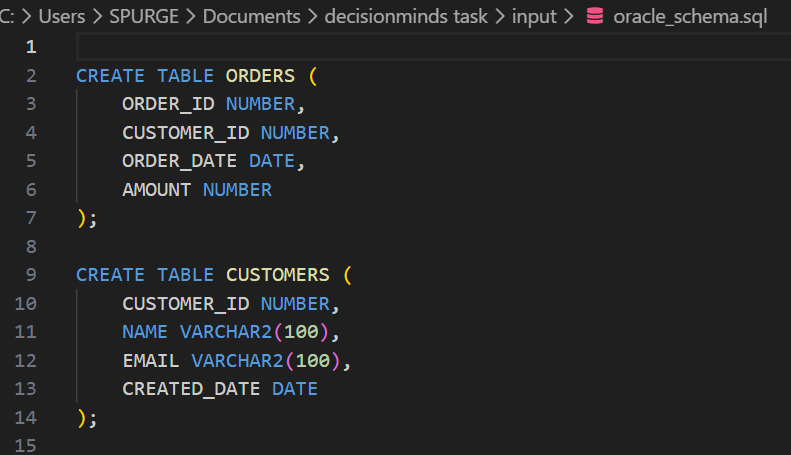


Figure 29 Transplier output-1

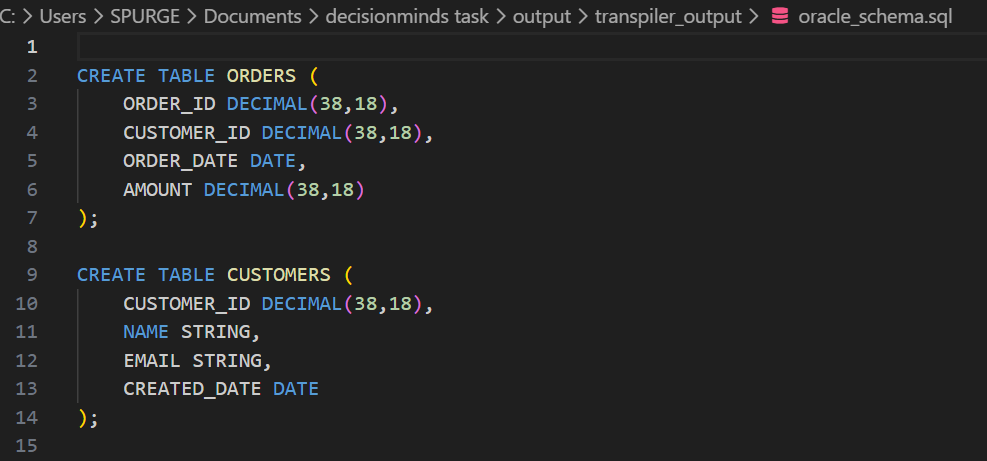


Figure 30 Transplier input-2

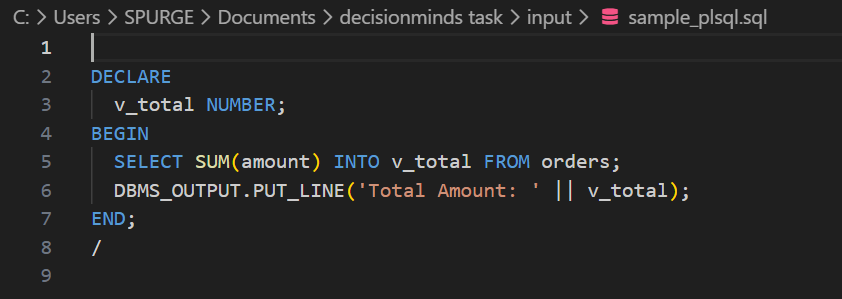


Figure 31 Transplier output-2

