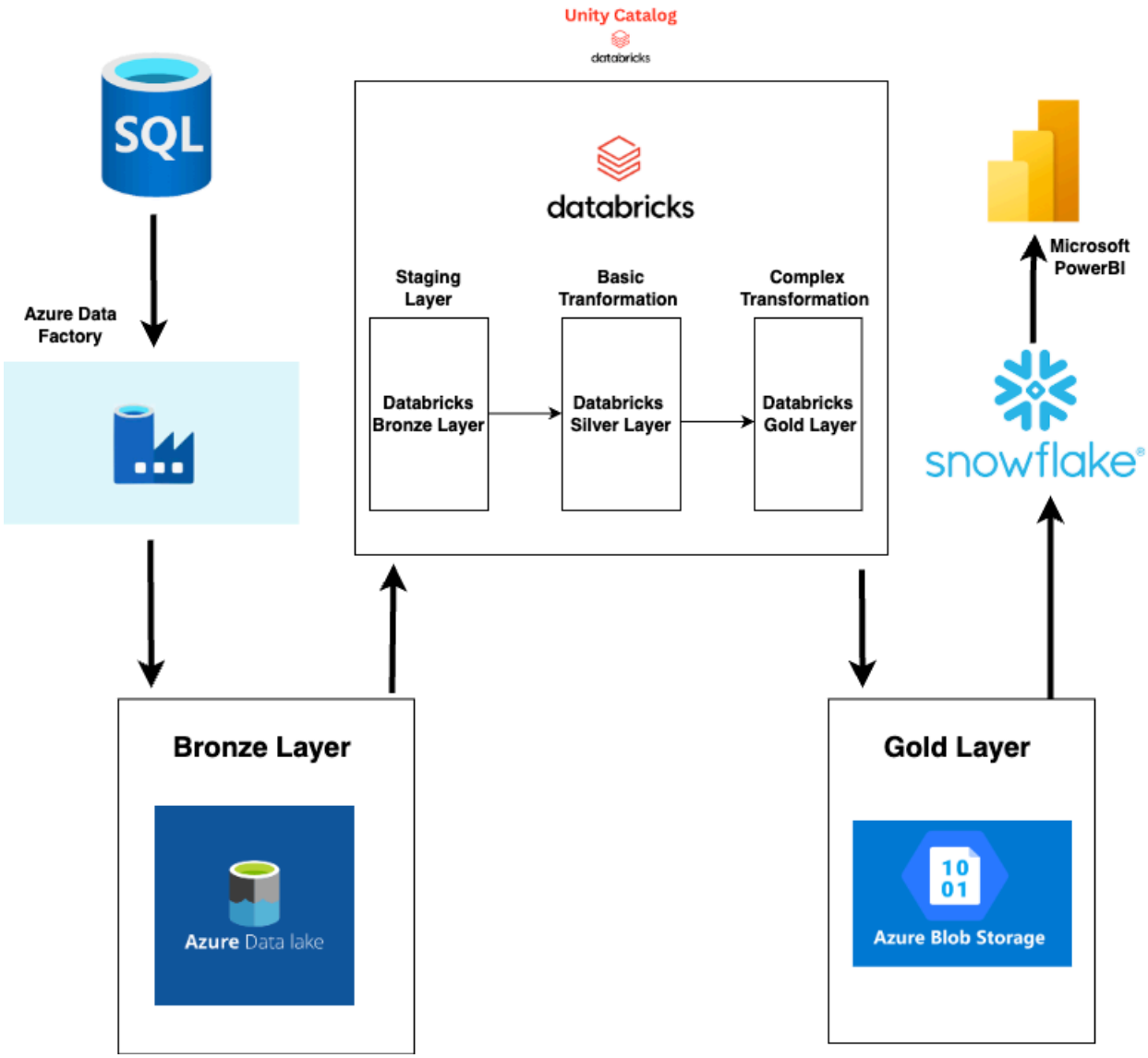
Real Time Sales Analytics

Basic Architecture:

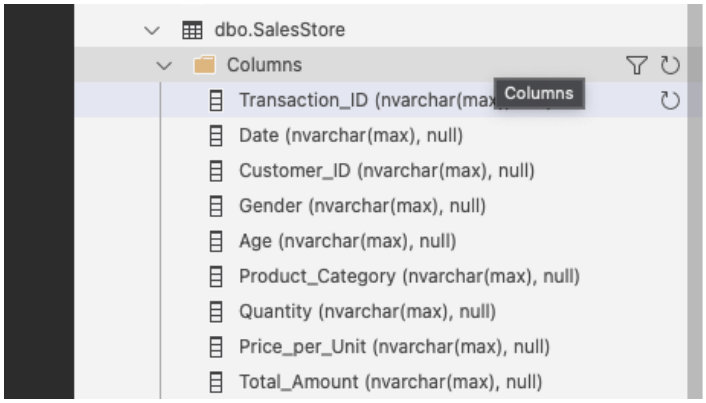


Data Integration:

Azure SQL Database :

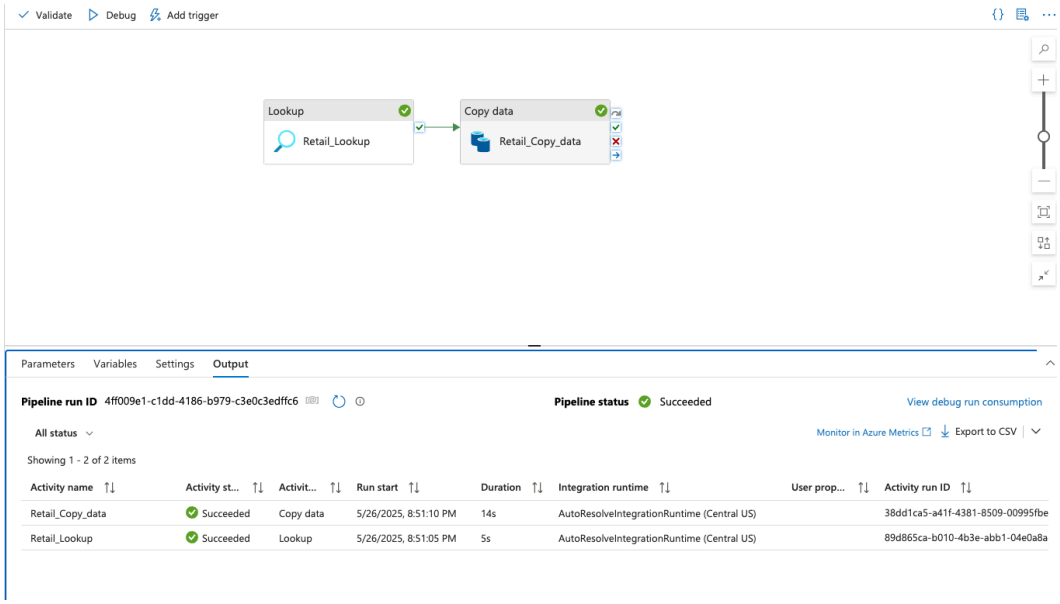
Retail Sales POS System data was extracted from Kaggle and ingested into Azure SQL database using Azure Data Factory

Used Azure Data Studio to view and create table inside SQL database.

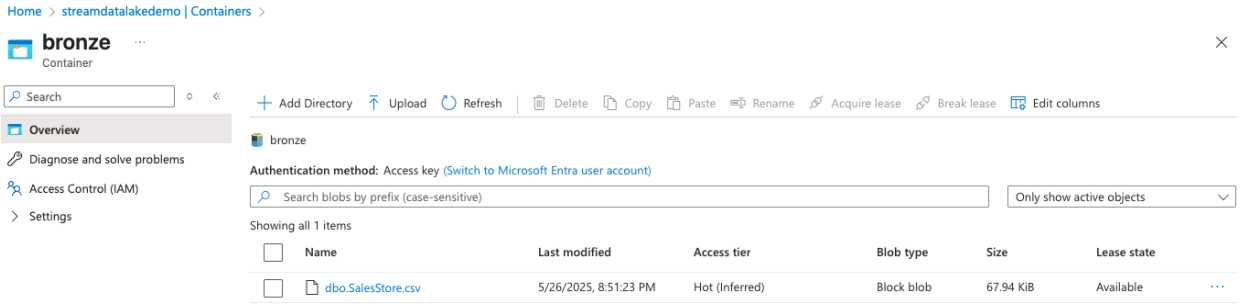


Retail Sales Data was copied from Azure SQL database to Azure Data Lake Gen2 Storage account using Azure Data Factory

Look up activity is chosen to read the data from the data source before copying it to the Data Lake Gen2 (Bronze layer) as part of the Medallion Architecture

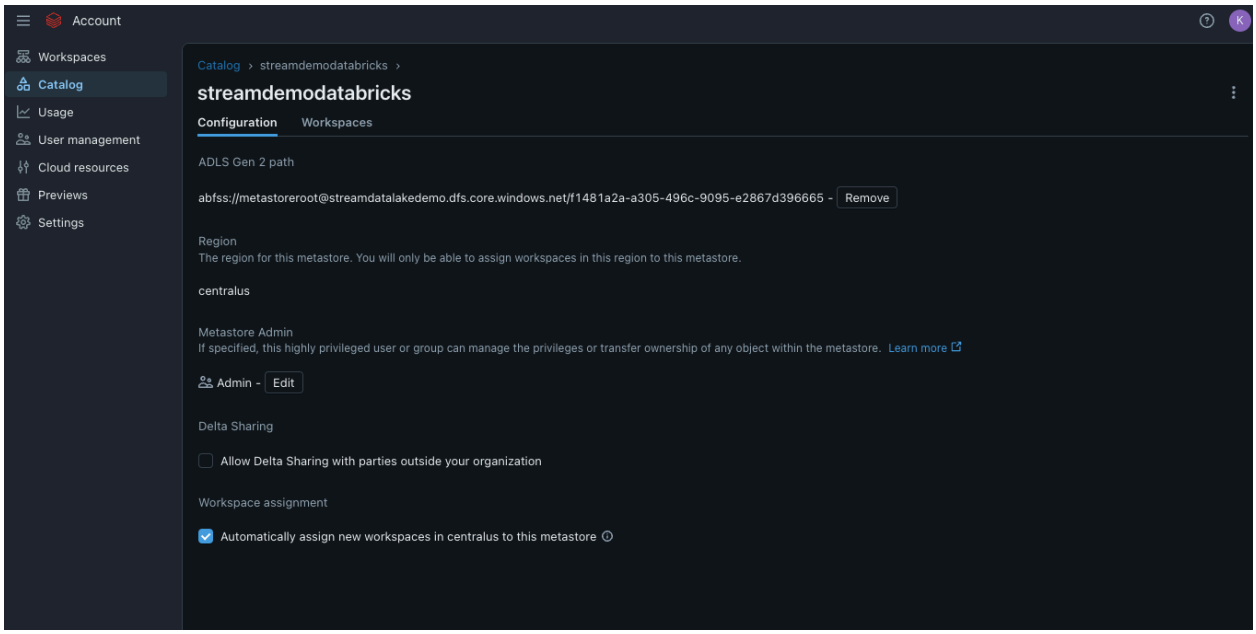


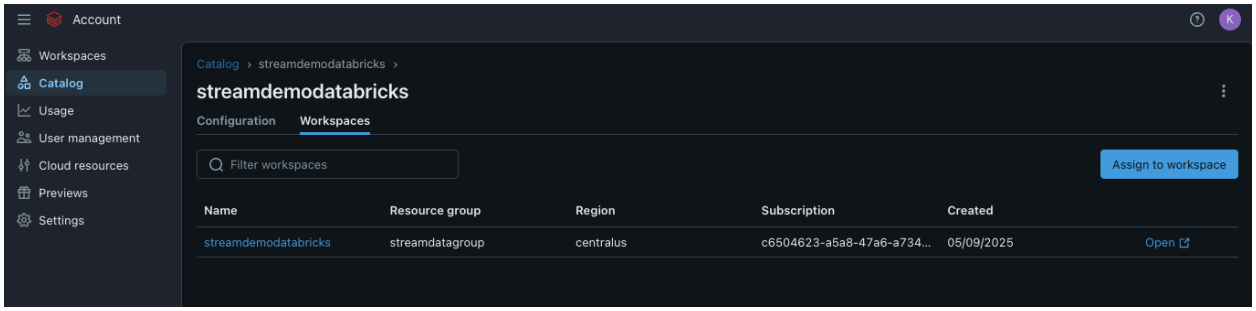
Data lake Gen 2 output:



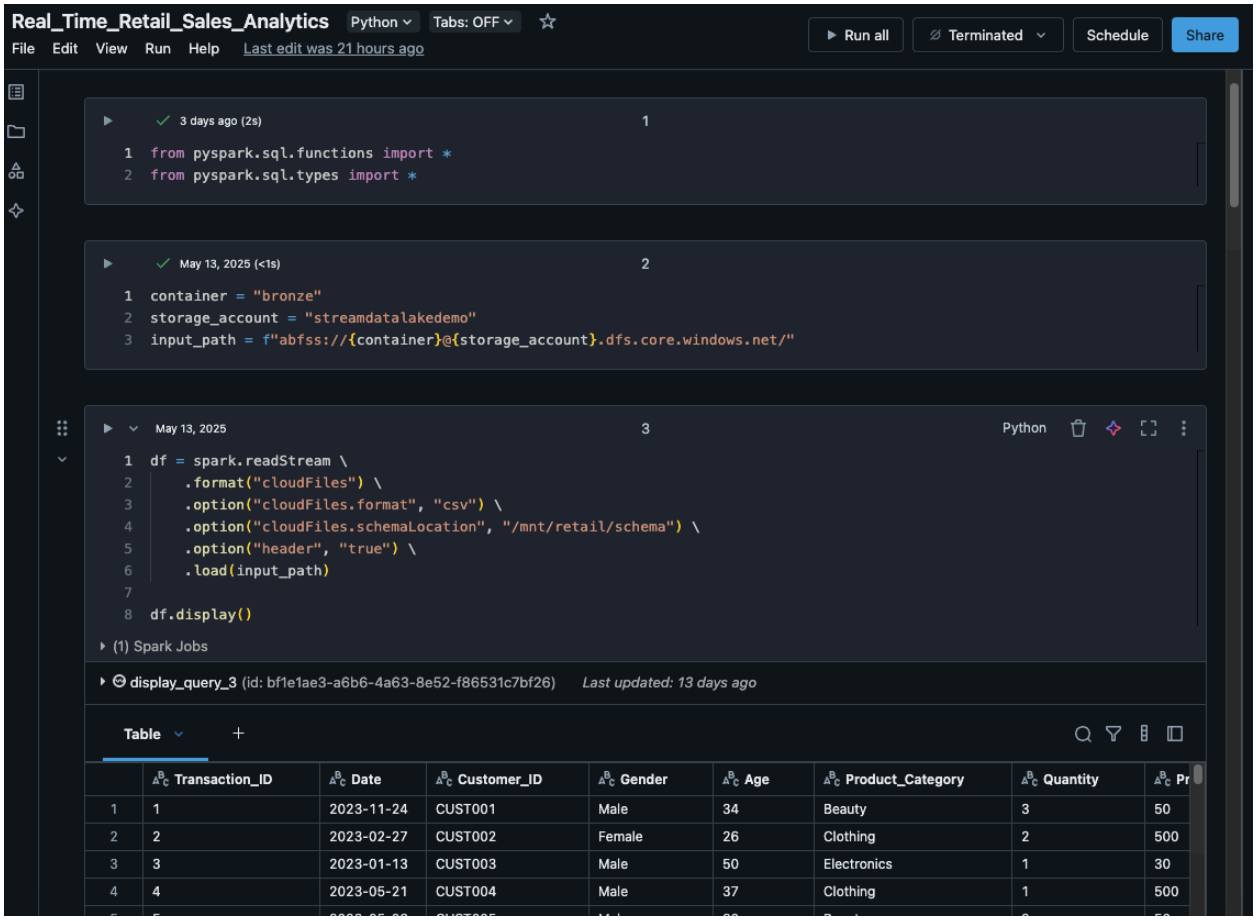
Performed data governance and connection from Data lake Gen2 Storage account to Azure Databricks using Unity Catalog

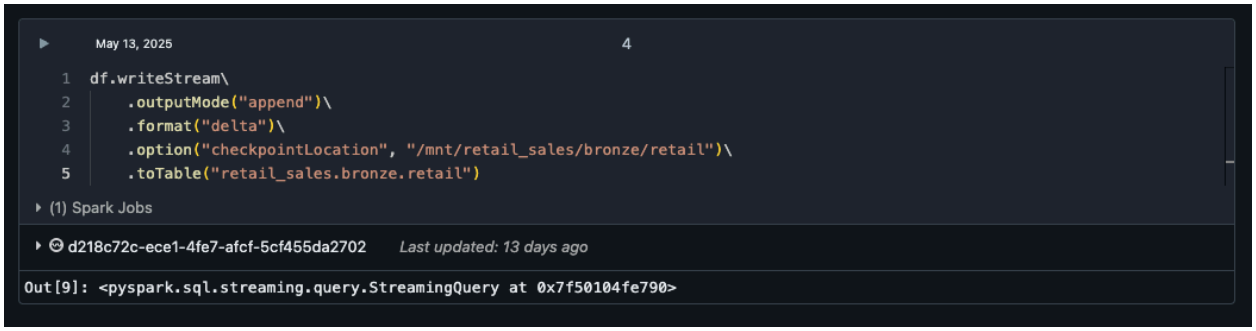
Unity catalog is a centralized data governance platform which helps handle RBAC, data lineage & auditability across workspaces while boosting cross-team collaboration.





Used Autoloader to load incremental data & copied into the Bronze layer which acts as a staging layer inside Azure Databricks Database File System(DBFS)


Real_Time_Retail_Sales_Analytics
Python
Tabs: OFF
File Edit View Run Help
Last edit was 21 hours ago
Run all
Terminated
Schedule
Share
3 days ago (2s)
1
from pyspark.sql.functions import *
from pyspark.sql.types import *
May 13, 2025 (<1s)
2
container = "bronze"
storage_account = "streamdatalakedemo"
input_path = f"abfss://{container}@{storage_account}.dfs.core.windows.net/"
May 13, 2025
3
Python
df = spark.readStream \
.format("cloudFiles") \
.option("cloudFiles.format", "csv") \
.option("cloudFiles.schemaLocation", "/mnt/retail/schema") \
.option("header", "true") \
.load(input_path)
df.display()
(1) Spark Jobs
display_query_3 (id: bf1e1ae3-a6b6-4a63-8e52-f86531c7bf26) Last updated: 13 days ago
Table
Transaction_ID
Date
Customer_ID
Gender
Age
Product_Category
Quantity
1 1 2023-11-24 CUST001 Male 34 Beauty 3 50
2 2 2023-02-27 CUST002 Female 26 Clothing 2 500
3 3 2023-01-13 CUST003 Male 50 Electronics 1 30
4 4 2023-05-21 CUST004 Male 37 Clothing 1 500


May 13, 2025
4
df.writeStream\
.outputMode("append")\
.format("delta")\
.option("checkpointLocation", "/mnt/retail_sales/bronze/retail")\
.toTable("retail_sales.bronze.retail")
(1) Spark Jobs
d218c72c-ece1-4fe7-afcf-5cf455da2702 Last updated: 13 days ago
Out[9]: <pyspark.sql.streaming.query.StreamingQuery at 0x7f50104fe790>

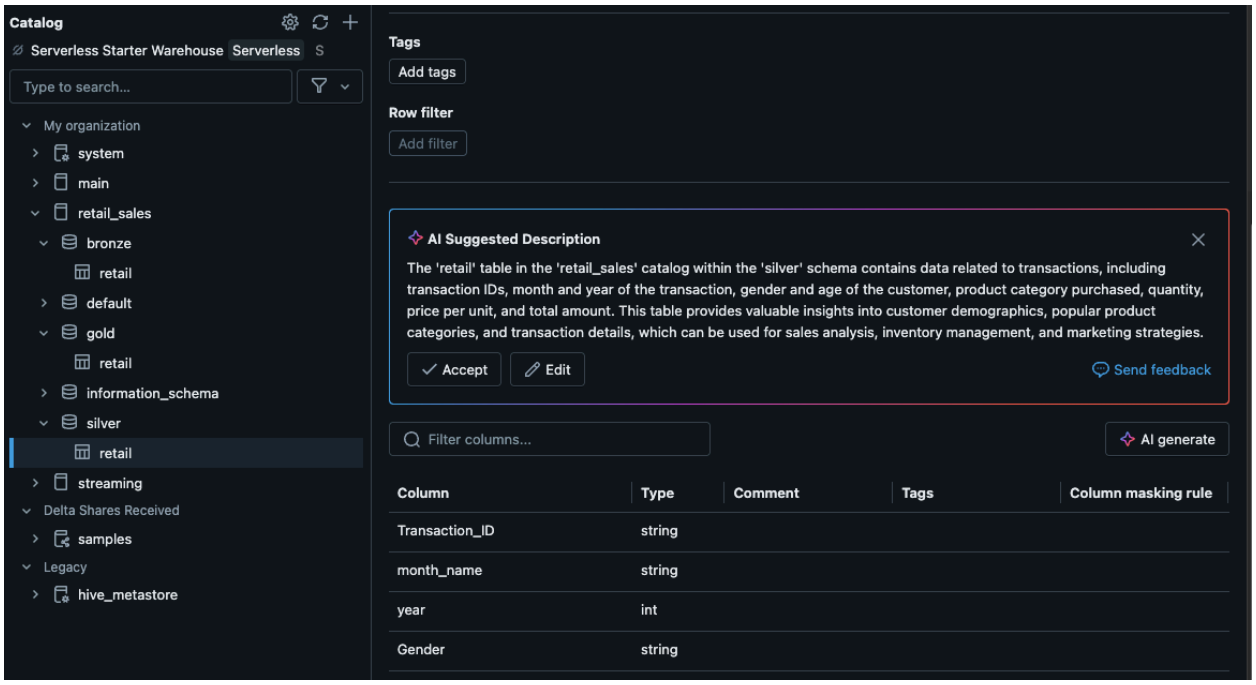
Basic Transformation:

```
May 17, 2025 5 Python

month_name_expr = (
    when(col("month") == 1, "January")
    .when(col("month") == 2, "February")
    .when(col("month") == 3, "March")
    .when(col("month") == 4, "April")
    .when(col("month") == 5, "May")
    .when(col("month") == 6, "June")
    .when(col("month") == 7, "July")
    .when(col("month") == 8, "August")
    .when(col("month") == 9, "September")
    .when(col("month") == 10, "October")
    .when(col("month") == 11, "November")
    .when(col("month") == 12, "December")
)

df = spark.readStream\
    .format('delta')\
    .table("retail_sales.bronze.retail")\
    .withColumn("month", month(col('Date')))\
    .withColumn("year", year(col('Date')))\
    .withColumn("month_name", month_name_expr)\
    .select("Transaction_ID", "month_name", "year", "Gender", "Age", "Product_Category", "Quantity", "Price_per_unit",
    "Total_Amount")

df.display()

(1) Spark Jobs
```

```
May 17, 2025 6 Python

df.writeStream\
    .outputMode('append')\
    .format('delta')\
    .option("checkpointLocation", "/mnt/retail_sales/silver/retail")\
    .toTable("retail_sales.silver.retail")


(2) Spark Jobs

087a5533-9113-4e8b-902a-626e2b72835b  Last updated: 10 days ago

Out[15]: <pyspark.sql.streaming.query.StreamingQuery at 0x73dc7d549580>
```

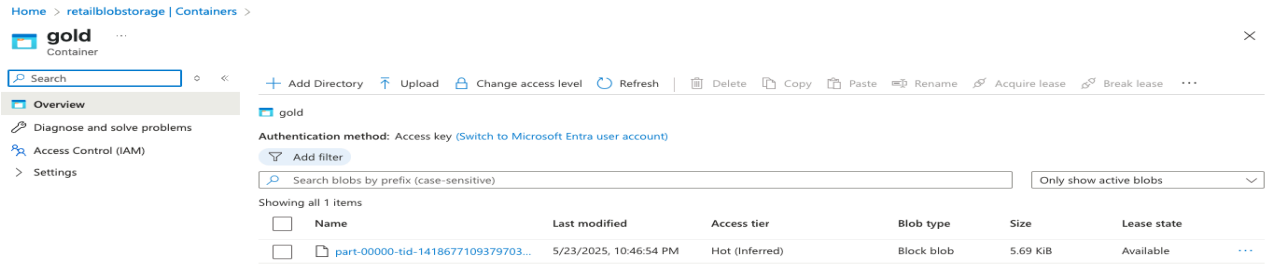
Complex Transformations:

```python
agg_df = df.groupBy(
    "month_name",
    "year",
    "Gender",
    "Age",
    "Product_Category"
).agg(
    sum("Quantity").alias("total_quantity"),
    sum("Total_Amount").alias("total_sales_amount")
)
```




```python
df = spark.read.format("delta").load("abfss://gold@streamdatalakedemo.dfs.core.windows.net/gold/retail_aggregated")

# Coalesce to 1 file for Power BI consumption
df.coalesce(1).write \
    .mode("overwrite") \
    .parquet("abfss://gold@retailblobstoragedemo.blob.core.windows.net/powerbi_export")
```

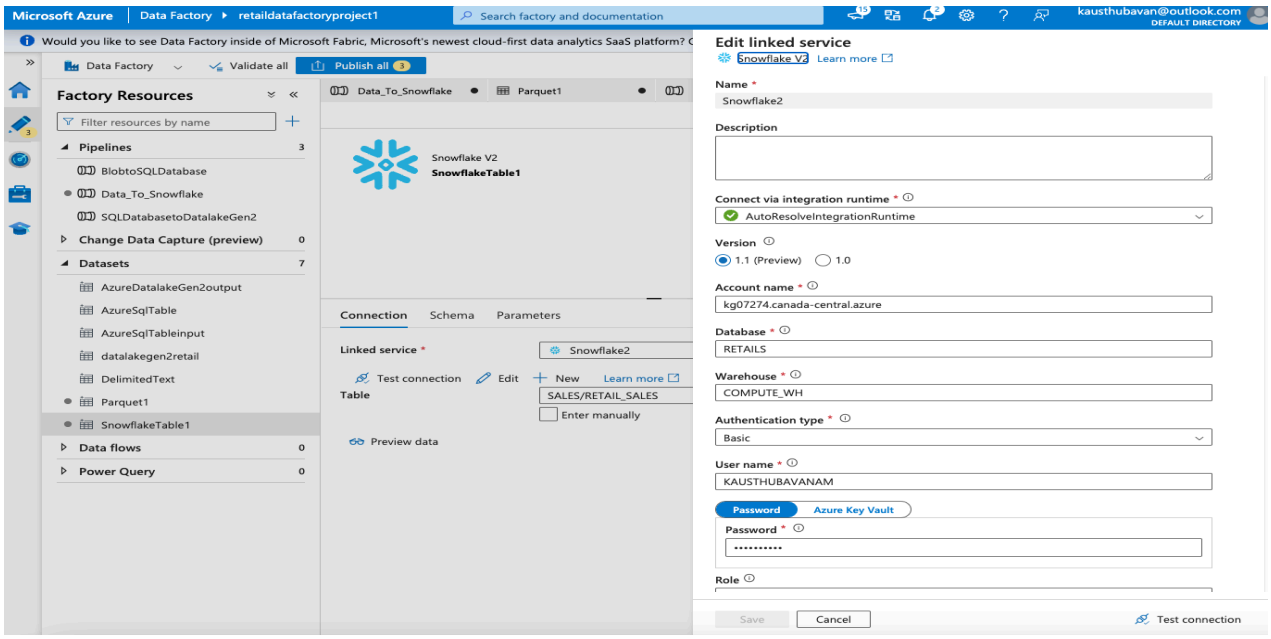
Data Inside Blob Storage (Gold Layer):

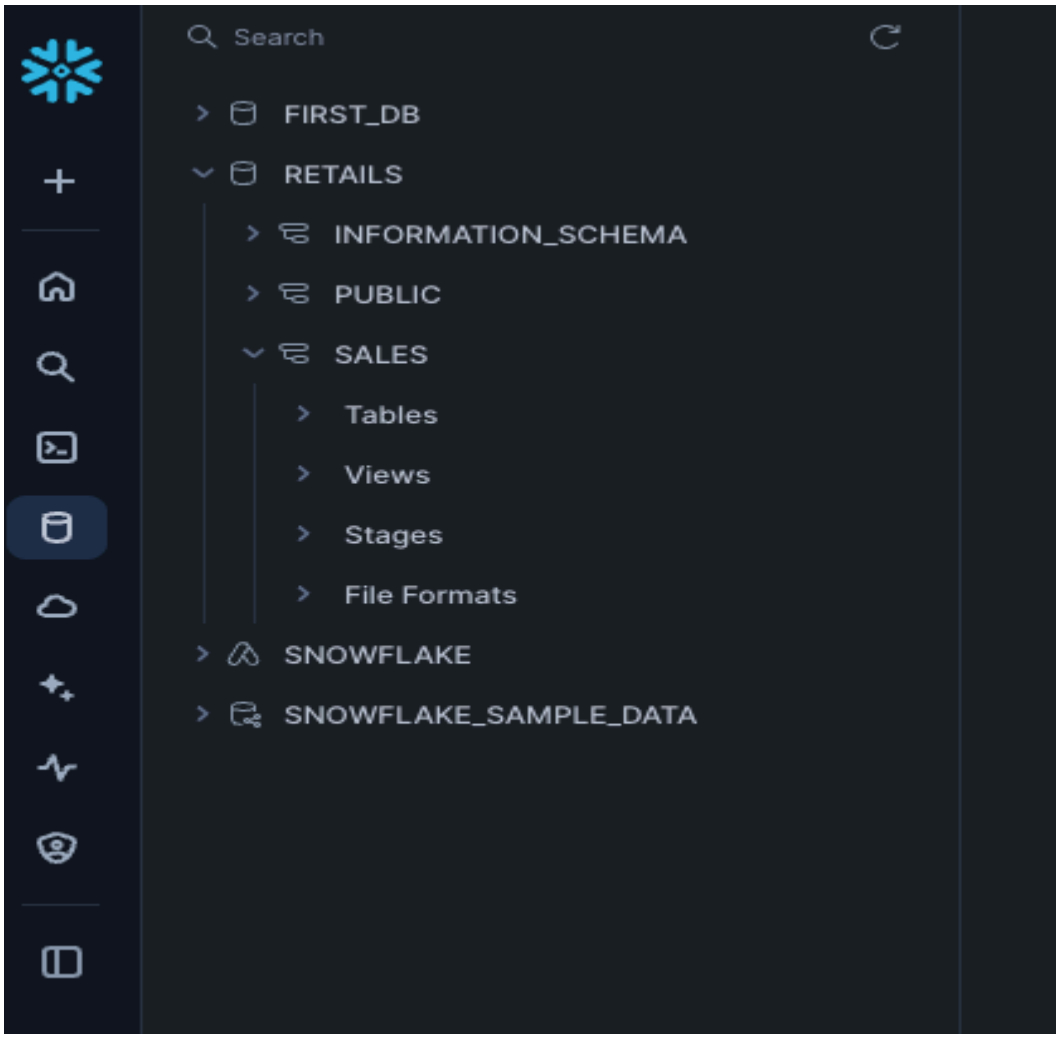


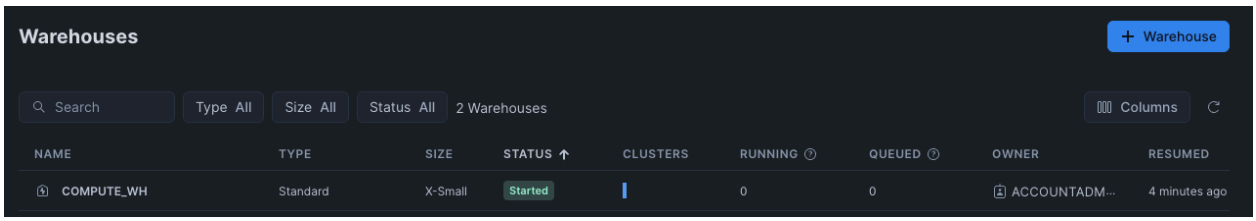
Ingesting Data into Snowflake using Azure Data Factory:

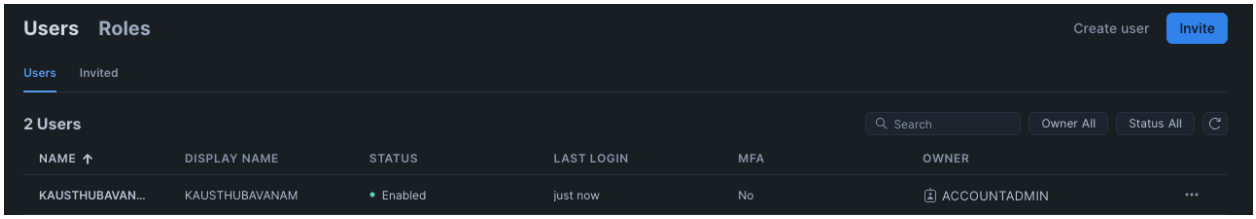
Used Snowflake as the Data Warehouse in this project because it's an efficient Centralized Data Repository which provides multi cluster Warehouse & helps in storage and exposes the data to a wide range of Cloud Providers & allows multiple users to work on ML applications, PowerBI Visualizations.

Connection for ADF to Snowflake




Search
FIRST_DB
RETAILS
INFORMATION_SCHEMA
PUBLIC
SALES
Tables
Views
Stages
File Formats
SNOWFLAKE
SNOWFLAKE_SAMPLE_DATA


Warehouses
+ Warehouse
Search
Type All
Size All
Status All
2 Warehouses
Columns
NAME
TYPE
SIZE
STATUS
CLUSTERS
RUNNING
QUEUED
OWNER
RESUMED
COMPUTE_WH
Standard
X-Small
Started
0
0
ACCOUNTADM...
4 minutes ago


Users
Roles
Create user
Invite
Users
Invited
2 Users
Search
Owner All
Status All
NAME
DISPLAY NAME
STATUS
LAST LOGIN
MFA
OWNER
KAUSTHUBAVAN...
KAUSTHUBAVANAM
Enabled
just now
No
ACCOUNTADMIN

Extracted Data From Snowflake to PowerBI to perform Visualizations on the Retail sales data.