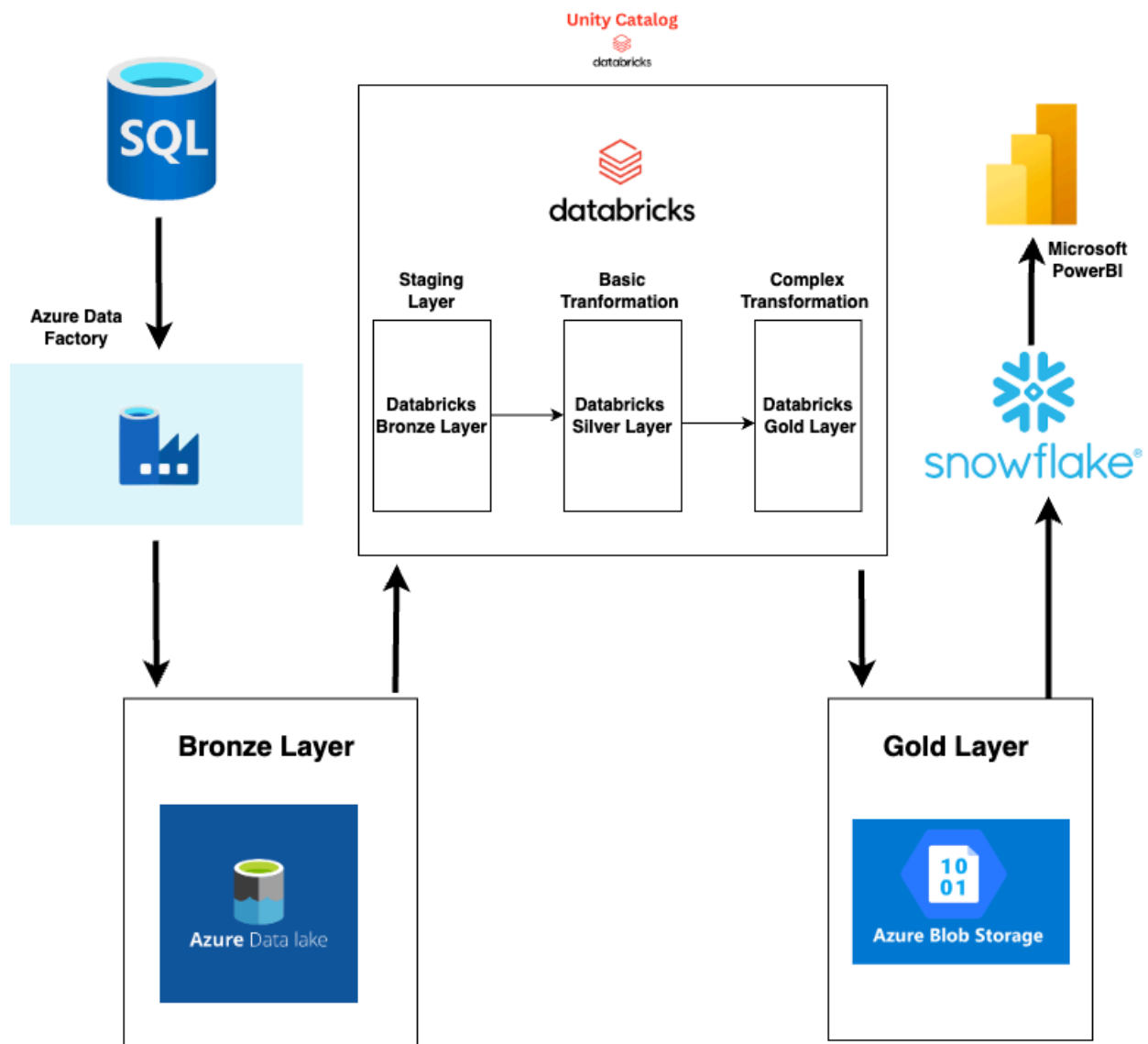


# Real Time Sales Analytics

## Basic Architecture:

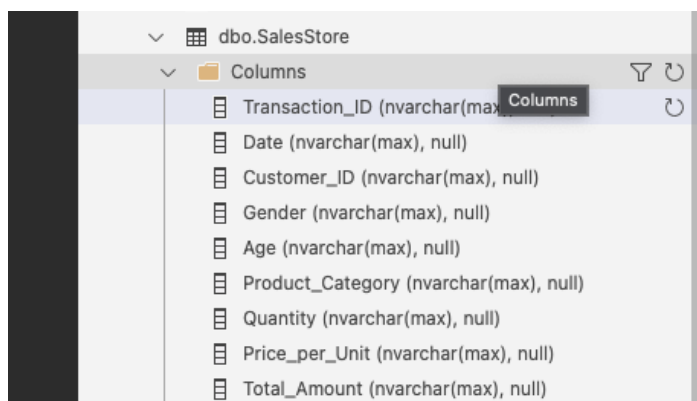


## Data Integration:

### 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

✓ Validate ▶ Debug ⚙️ Add trigger

{} ⓘ ...

Parameters Variables Settings **Output**

**Pipeline run ID** 4ff009e1-c1dd-4186-b979-c3e0c3edffc6 **Pipeline status** ✓ Succeeded [View debug run consumption](#)

All status ▾ [Monitor in Azure Metrics](#) [Export to CSV](#) ▾

Showing 1 - 2 of 2 items

Activity name	Activity st...	Activit...	Run start	Duration	Integration runtime	User prop...	Activity run ID
Retail_Copy_data	✓ Succeeded	Copy data	5/26/2025, 8:51:10 PM	14s	AutoResolveIntegrationRuntime (Central US)		38dd1ca5-a41f-4381-8509-00995fbc
Retail_Lookup	✓ Succeeded	Lookup	5/26/2025, 8:51:05 PM	5s	AutoResolveIntegrationRuntime (Central US)		89db65ca-b010-4b3e-abb1-04e0a8a

## Data lake Gen 2 output:

[Home](#) > [streamdatalakedemo](#) | [Containers](#) >

**bronze** Container

Search

+ Add Directory ↑ Upload ↻ Refresh 🗑️ Delete 📄 Copy 📄 Paste 🔄 Rename 🔑 Acquire lease 🔑 Break lease 🛠️ Edit columns

**Overview**

🔧 Diagnose and solve problems

🔑 Access Control (IAM)

> Settings

Authentication method: Access key ([Switch to Microsoft Entra user account](#))

Search blobs by prefix (case-sensitive)

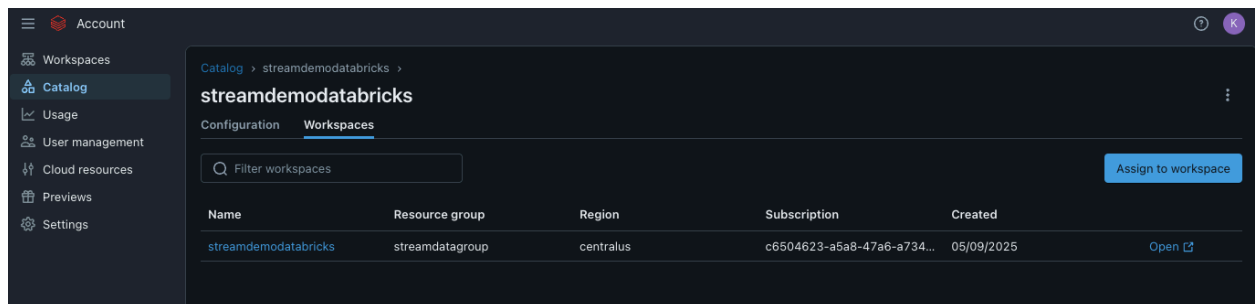
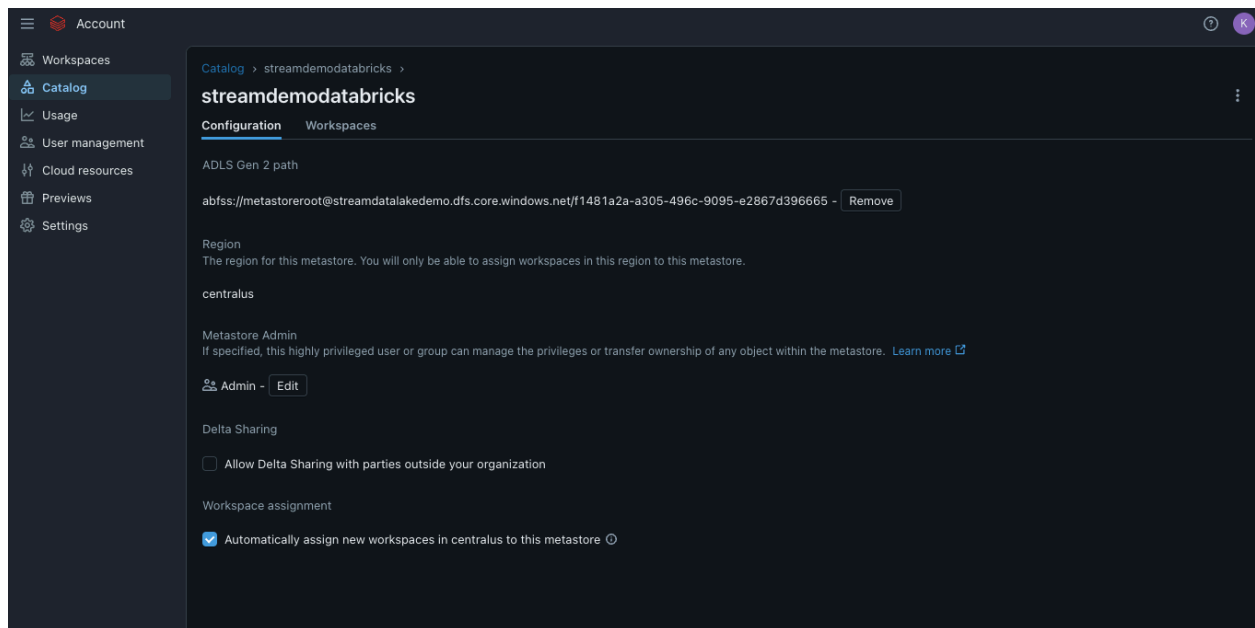
Only show active objects ▾

Showing all 1 items

<input type="checkbox"/>	Name	Last modified	Access tier	Blob type	Size	Lease state
<input type="checkbox"/>	<a href="#">dbo.SalesStore.csv</a>	5/26/2025, 8:51:23 PM	Hot (Inferred)	Block blob	67.94 KiB	Available

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\_AnalyticsPythonTabs: OFF

FileEditViewRunHelpLast edit was 21 hours ago

Run allTerminatedScheduleShare

3 days ago (2s)1

1 from pyspark.sql.functions import \*  
2 from pyspark.sql.types import \*

May 13, 2025 (<1s)2

1 container = "bronze"  
2 storage\_account = "streamdatalakedemo"  
3 input\_path = f"abfss://{container}@{storage\_account}.dfs.core.windows.net/"

May 13, 20253Python

1 df = spark.readStream \  
2 .format("cloudFiles") \  
3 .option("cloudFiles.format", "csv") \  
4 .option("cloudFiles.schemaLocation", "/mnt/retail/schema") \  
5 .option("header", "true") \  
6 .load(input\_path)  
7  
8 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	Pr
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, 20254

1 df.writeStream\  
2 .outputMode("append")\  
3 .format("delta")\  
4 .option("checkpointLocation", "/mnt/retail\_sales/bronze/retail")\  
5 .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
```

```
1 month_name_expr = (  
2     when(col("month") == 1, "January")  
3     .when(col("month") == 2, "February")  
4     .when(col("month") == 3, "March")  
5     .when(col("month") == 4, "April")  
6     .when(col("month") == 5, "May")  
7     .when(col("month") == 6, "June")  
8     .when(col("month") == 7, "July")  
9     .when(col("month") == 8, "August")  
10    .when(col("month") == 9, "September")  
11    .when(col("month") == 10, "October")  
12    .when(col("month") == 11, "November")  
13    .when(col("month") == 12, "December")  
14 )  
15  
16 df = spark.readStream\  
17     .format('delta')\  
18     .table("retail_sales.bronze.retail")\  
19     .withColumn("month", month(col('Date')))\  
20     .withColumn("year", year(col('Date')))\  
21     .withColumn("month_name", month_name_expr)\  
22     .select("Transaction_ID", "month_name", "year", "Gender", "Age", "Product_Category", "Quantity", "Price_per_unit",  
23            "Total_Amount")  
24 df.display()
```

▶ (1) Spark Jobs

```
May 17, 2025 6 Python
```

```
1 df.writeStream\  
2     .outputMode('append')\  
3     .format('delta')\  
4     .option("checkpointLocation", "/mnt/retail_sales/silver/retail")\  
5     .toTable("retail_sales.silver.retail")  
6  
7
```

▶ (2) Spark Jobs

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

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

## Complex Transformations:

```
1
2 agg_df = df.groupBy(
3     "month_name",
4     "year",
5     "Gender",
6     "Age",
7     "Product_Category"
8 ).agg(
9     sum("Quantity").alias("total_quantity"),
10    sum("Total_Amount").alias("total_sales_amount")
11 )
12
```

Catalog

Serverless Starter Warehouse Serverless S

Type to search...

My organization

system

main

retail\_sales

bronze

retail

default

gold

retail

information\_schema

silver

retail

streaming

Delta Shares Received

samples

Legacy

hive\_metastore

Tags

Add tags

Row filter

Add filter

AI Suggested Description

The 'retail' table in the 'retail\_sales' catalog within the 'silver' schema contains data related to transactions, including transaction IDs, month and year of the transaction, gender and age of the customer, product category purchased, quantity, price per unit, and total amount. This table provides valuable insights into customer demographics, popular product categories, and transaction details, which can be used for sales analysis, inventory management, and marketing strategies.

Accept Edit Send feedback

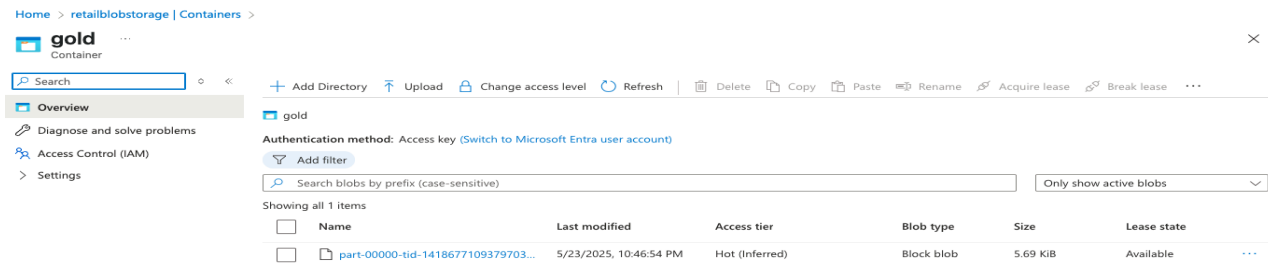
Filter columns...

AI generate

Column	Type	Comment	Tags	Column masking rule
Transaction_ID	string			
month_name	string			
year	int			
Gender	string			

```
1 df = spark.read.format("delta").load("abfss://gold@streamdatalakedemo.dfs.core.windows.net/gold/retail_aggregated")
2
3 # Coalesce to 1 file for Power BI consumption
4 df.coalesce(1).write \
5     .mode("overwrite") \
6     .parquet("abfss://gold@retailblobstoragedemo.blob.core.windows.net/powerbi_export")
```

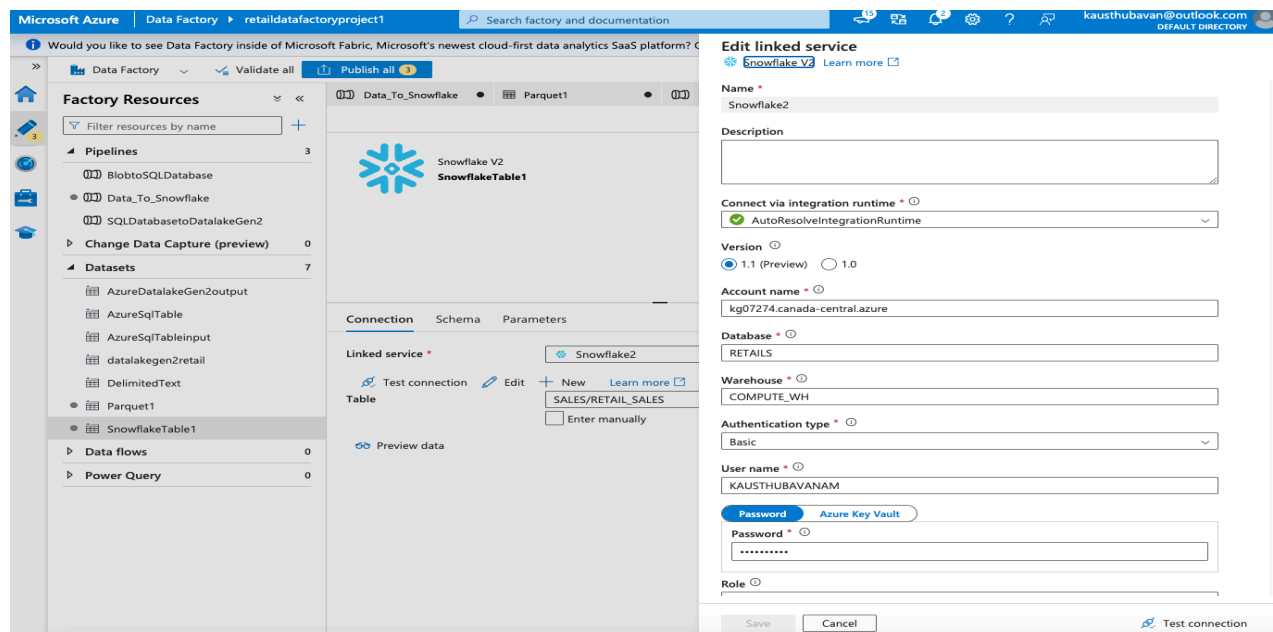
## Data Inside Blob Storage (Gold Layer):



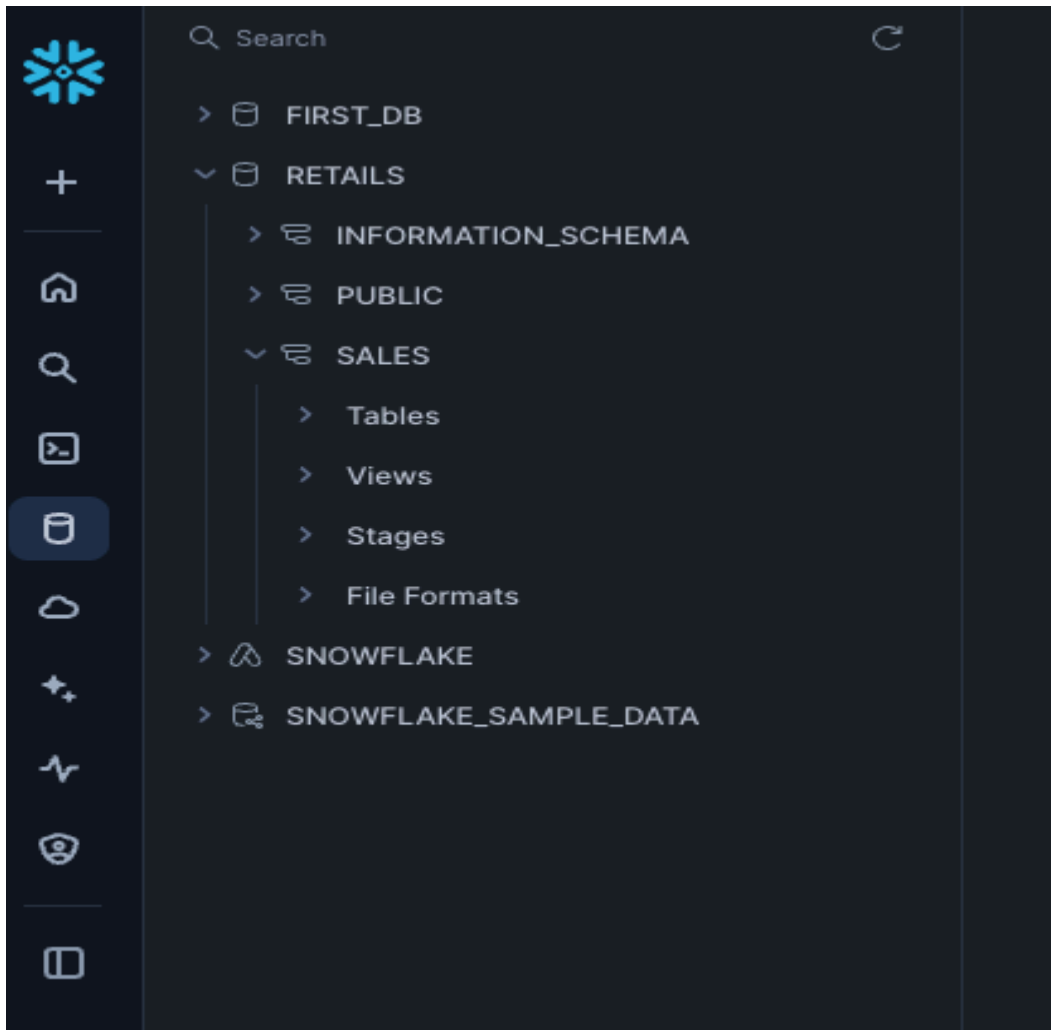
## Ingesting Data into Snowflake using Azure Data Factory:

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







Warehouses								
<div><div>Search</div><div>Type AllSize AllStatus All2 Warehouses</div><div>Columns</div></div>								
NAME	TYPE	SIZE	STATUS <span>↑</span>	CLUSTERS	RUNNING <span>⌚</span>	QUEUED <span>⌚</span>	OWNER	RESUMED
COMPUTE_WH	Standard	X-Small	Started	1	0	0	ACCOUNTADM...	4 minutes ago

Users Roles						
<div><div>UsersInvited</div><div>Create user</div><div>Invite</div></div>						
<div>2 Users</div> <div><div>Search</div><div>Owner AllStatus All</div></div>						
NAME <span>↑</span>	DISPLAY NAME	STATUS	LAST LOGIN	MFA	OWNER	
KAUSTHUBAVAN...	KAUSTHUBAVANAM	<span>Enabled</span>	just now	No	ACCOUNTADMIN	...

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

