Case Study Documentation
Retail Sales Analysis

NAME: HARSHA M R

Objective

Objective of the Case Study:

- To build an end-to-end data pipeline for retail sales data.
- To ingest CSV data from Azure Blob Storage into Databricks.
- To clean, transform, and summarize the dataset using PySpark.
- To store the transformed data in Snowflake as raw and summary tables.
- To visualize insights using Power BI.

Dataset Description

Source: Azure Blob Storage

File: Retail_Sales.csv

FLOW:

```
Azure Blob Storage (Retail_Sales.csv)
        |
        ▼
  Databricks (PySpark)
    ├── Data Cleaning
    ├── Date Transformation
    ├── Type Casting (Sales, Profit, Quantity)
    └── Summary Table (Category-wise Sales)
        |
        ▼
  Snowflake (Data Warehouse)
    ├── RETAIL_SALES_RAW
    └── RETAIL_SALES_SUMMARY
        |
        ▼
     Power BI
    ├── Sales Trend Analysis
    ├── Category-wise Sales
    └── Top Products Analysis
```

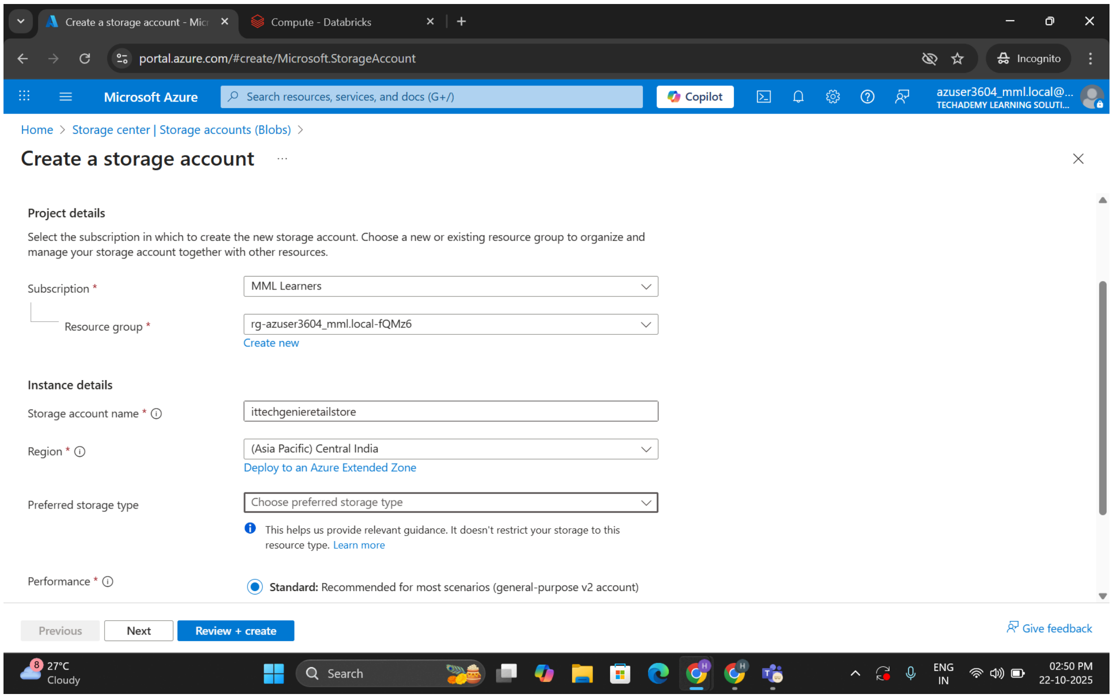
Environment Setup

Tools & Platforms Used:

- Databricks: Apache Spark environment for data processing
- Snowflake: Cloud Data Warehouse for storing raw and summary tables
- Azure Blob Storage: Source for CSV dataset
- Power BI Desktop: Visualization tool
- Python (PySpark): For data transformation

Step 1: Azure Blob Storage Setup

1. **Create Storage Account** in Azure Portal:
    - Name: ittechgenieretailstore
    - Type: Standard, StorageV2
    - Region: Central India



Create a storage account - Micr
Compute - Databricks
portal.azure.com/#create/Microsoft.StorageAccount
Incognito
Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
azuser3604_mml.local@...
TECHADEMY LEARNING SOLUTI...
Home > Storage center | Storage accounts (Blobs) >

Create a storage account

Project details

Select the subscription in which to create the new storage account. Choose a new or existing resource group to organize and manage your storage account together with other resources.

Subscription * : MML Learners

Resource group * : rg-azuser3604_mml.local-fQMz6

Create new

Instance details

Storage account name * : ittechgenieretailstore

Region * : (Asia Pacific) Central India

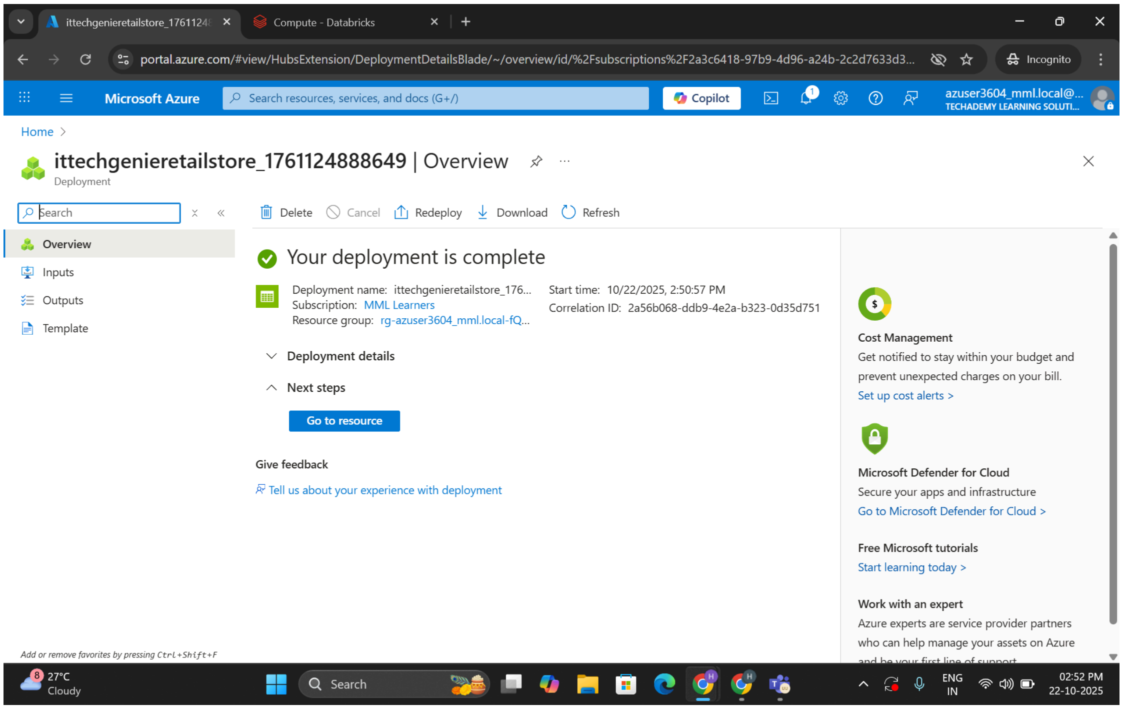
Deploy to an Azure Extended Zone

Preferred storage type: Choose preferred storage type

This helps us provide relevant guidance. It doesn't restrict your storage to this resource type. Learn more

Performance * : ● **Standard:** Recommended for most scenarios (general-purpose v2 account)

Previous
Next
Review + create
Give feedback
27°C
Cloudy
Search
ENG
IN
02:50 PM
22-10-2025



ittechgenieretailstore_1761124
Compute - Databricks
portal.azure.com/#view/HubsExtension/DeploymentDetailsBlade/~/overview/id/%2Fsubscriptions%2F2a3c6418-97b9-4d96-a24b-2c2d7633d3...
Incognito
Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
azuser3604_mml.local@...
TECHADEMY LEARNING SOLUTI...
Home >

ittechgenieretailstore_1761124888649 | Overview

Deployment

Search
Overview
Inputs
Outputs
Template
Delete
Cancel
Redeploy
Download
Refresh

Your deployment is complete

Deployment name: ittechgenieretailstore_176...
Subscription: MML Learners
Resource group: rg-azuser3604_mml.local-fQ...

Start time: 10/22/2025, 2:50:57 PM
Correlation ID: 2a56b068-ddb9-4e2a-b323-0d35d751

Deployment details

Next steps

Go to resource

Give feedback

Tell us about your experience with deployment

Cost Management
Get notified to stay within your budget and prevent unexpected charges on your bill.
Set up cost alerts >

Microsoft Defender for Cloud
Secure your apps and infrastructure
Go to Microsoft Defender for Cloud >

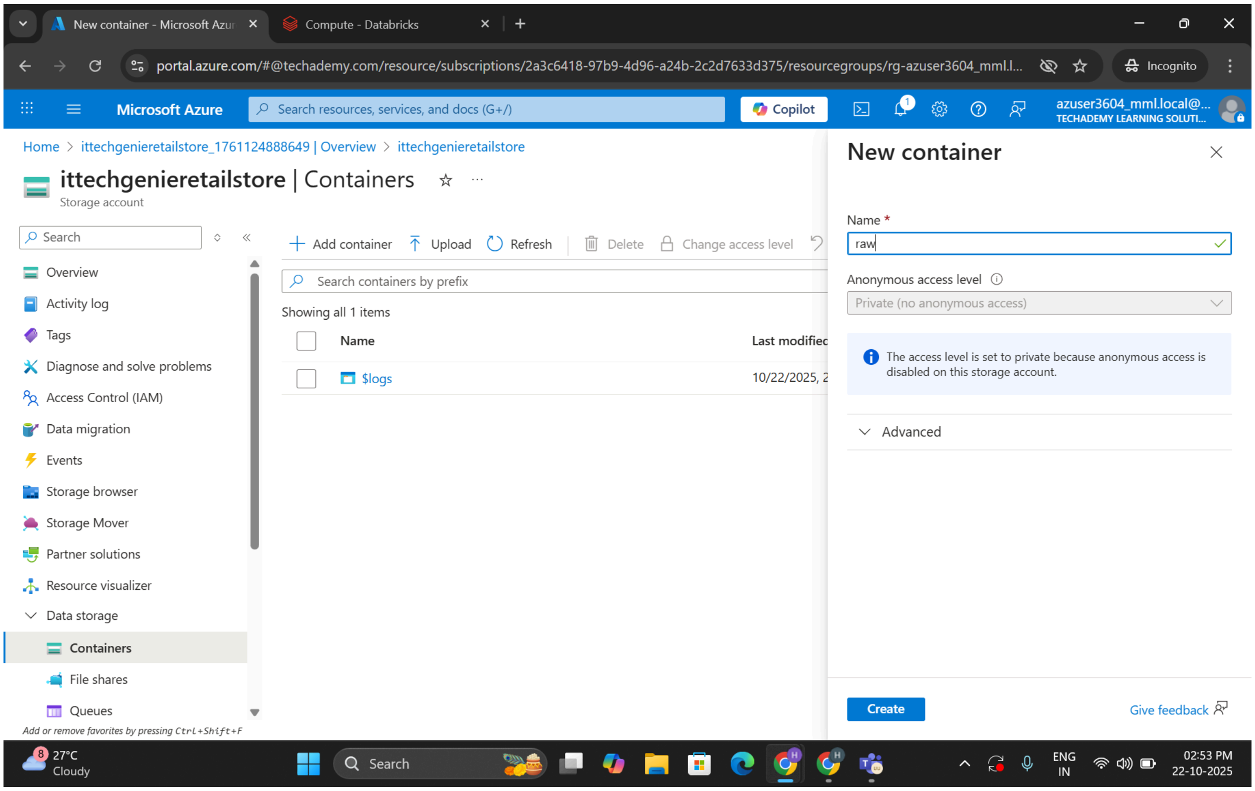
Free Microsoft tutorials
Start learning today >

Work with an expert
Azure experts are service provider partners who can help manage your assets on Azure and be your first line of support

2. Create Container:

- Container Name: `raw`

- Public access level: Private



New container - Microsoft Azure
Compute - Databricks
portal.azure.com/#@techademy.com/resource/subscriptions/2a3c6418-97b9-4d96-a24b-2c2d7633d375/resourcegroups/rg-azuser3604_mml.l...
Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
azuser3604_mml.local@...
TECHADEMY LEARNING SOLUTI...
Home > ittechgenieretailstore_1761124888649 | Overview > ittechgenieretailstore
ittechgenieretailstore | Containers
Storage account
Search
Overview
Activity log
Tags
Diagnose and solve problems
Access Control (IAM)
Data migration
Events
Storage browser
Storage Mover
Partner solutions
Resource visualizer
Data storage
Containers
File shares
Queues
Add or remove favorites by pressing Ctrl+Shift+F
Add container
Upload
Refresh
Delete
Change access level
Search containers by prefix
Showing all 1 items
Name
Last modified
$logs
10/22/2025, 2
Create
Give feedback
27°C
Cloudy
Search
ENG
IN
02:53 PM
22-10-2025

New container

Name *: raw

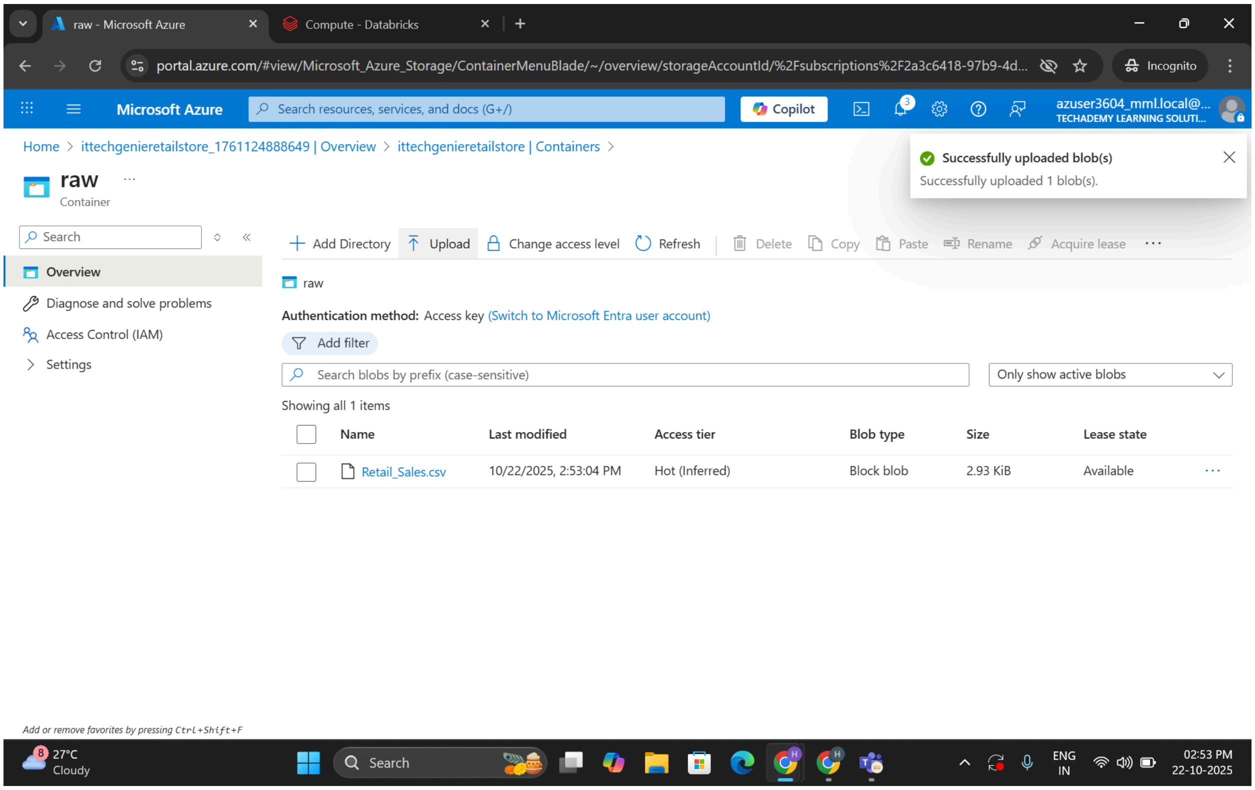
Anonymous access level: Private (no anonymous access)

The access level is set to private because anonymous access is disabled on this storage account.

Advanced

3. Upload CSV:

- File: Retail_Sales.csv
- Path: raw/Retail_Sales.csv


raw - Microsoft Azure
Compute - Databricks
portal.azure.com/#view/Microsoft_Azure_Storage/ContainerMenuBlade/~/overview/storageAccountId/%2Fsubscriptions%2F2a3c6418-97b9-4d...
Incognito
Microsoft Azure
Search resources, services, and docs (G+/)
Copilot
azuser3604_mml.local@...
TECHADEMY LEARNING SOLUTI...
Home > ittechgenieretailstore_1761124888649 | Overview > ittechgenieretailstore | Containers >
raw
Container
Search
Overview
Diagnose and solve problems
Access Control (IAM)
Settings
Add Directory
Upload
Change access level
Refresh
Delete
Copy
Paste
Rename
Acquire lease
Add filter
Search blobs by prefix (case-sensitive)
Only show active blobs
Add or remove favorites by pressing Ctrl+Shift+F
27°C
Cloudy
Search
ENG
IN
02:53 PM
22-10-2025

Successfully uploaded blob(s)

Successfully uploaded 1 blob(s).

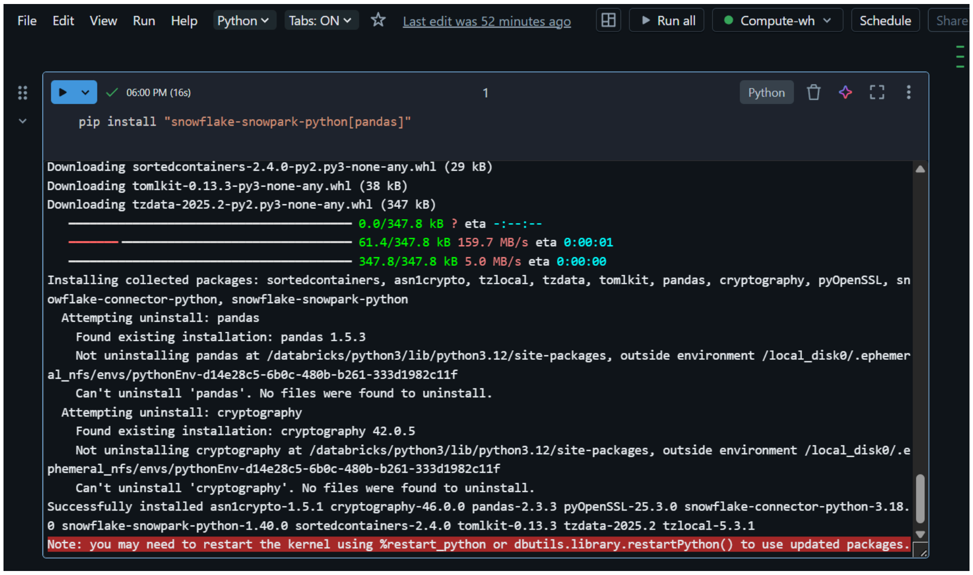
raw

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

Showing all 1 items

	Name	Last modified	Access tier	Blob type	Size	Lease state
□	Retail_Sales.csv	10/22/2025, 2:53:04 PM	Hot (Inferred)	Block blob	2.93 KiB	Available

Step 2: Load CSV into Databricks



```
pip install "snowflake-snowpark-python[pandas]"
```

```
Downloading sortedcontainers-2.4.0-py2.py3-none-any.whl (29 kB)
Downloading tomlkit-0.13.3-py3-none-any.whl (38 kB)
Downloading tzdata-2025.2-py2.py3-none-any.whl (347 kB)
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 0.0/347.8 kB ? eta -:--:--
   ━━━━━━ ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 61.4/347.8 kB 159.7 MB/s eta 0:00:01
   ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 347.8/347.8 kB 5.0 MB/s eta 0:00:00
Installing collected packages: sortedcontainers, asn1crypto, tzlocal, tzdata, tomlkit, pandas, cryptography, pyOpenSSL, sn
owflake-connector-python, snowflake-snowpark-python
  Attempting uninstall: pandas
    Found existing installation: pandas 1.5.3
    Not uninstalling pandas at /databricks/python3/lib/python3.12/site-packages, outside environment /local_disk0/.ephemer
al_nfs/envs/pythonEnv-d14e28c5-6b0c-480b-b261-333d1982c11f
    Can't uninstall 'pandas'. No files were found to uninstall.
  Attempting uninstall: cryptography
    Found existing installation: cryptography 42.0.5
    Not uninstalling cryptography at /databricks/python3/lib/python3.12/site-packages, outside environment /local_disk0/.e
phemeral_nfs/envs/pythonEnv-d14e28c5-6b0c-480b-b261-333d1982c11f
    Can't uninstall 'cryptography'. No files were found to uninstall.
Successfully installed asn1crypto-1.5.1 cryptography-46.0.0 pandas-2.3.3 pyOpenSSL-25.3.0 snowflake-connector-python-3.18.
0 snowflake-snowpark-python-1.40.0 sortedcontainers-2.4.0 tomlkit-0.13.3 tzdata-2025.2 tzlocal-5.3.1
Note: you may need to restart the kernel using %restart_python or dbutils.library.restartPython() to use updated packages.
```

```python
from snowflake.snowpark import Session
from snowflake.snowpark.functions import col, trim, upper, to_date, sum as _sum


connection_parameters = {
    "account": "EBDQDYC-NB21316",
    "user": "HARSHA",
    "password": "HarshaRadhakrishnan02",
    "role": "ACCOUNTADMIN",
    "warehouse": "CASE_WH",
    "database": "ITTECHGENIE_DB",
    "schema": "PUBLIC"
}
session = Session.builder.configs(connection_parameters).create()


columns = [
    "OrderID", "OrderDate", "MonthOfSale", "CustomerID", "CustomerName",
    "Country", "Region", "City", "Category", "Subcategory",
    "Quantity", "Discount", "Sales", "Profit"
]
```

```
df = session.read.option("skip_header",
1).csv("@retail_stage/Retail_Sales.csv")



df = df.to_df(*columns)


df_clean = (

    df.with_column("OrderDate", to_date(trim(col("OrderDate")), "YYYY-MM-DD"))

      .with_column("Category", upper(col("Category")))

      .with_column("Quantity", col("Quantity").cast("INTEGER"))

      .with_column("Discount", col("Discount").cast("FLOAT"))

      .with_column("Sales", col("Sales").cast("FLOAT"))

      .with_column("Profit", col("Profit").cast("FLOAT"))

)



df_summary =
df_clean.group_by("Category").agg(_sum(col("Sales")).alias("Category_Sales"))



df_clean.write.save_as_table("RETAIL_SALES_RAW", mode="overwrite")

df_summary.write.save_as_table("RETAIL_SALES_SUMMARY", mode="overwrite")
```

Explanation:

- Removed spaces from OrderDate.

- Converted OrderDate to DATE type (yyyy-MM-dd).

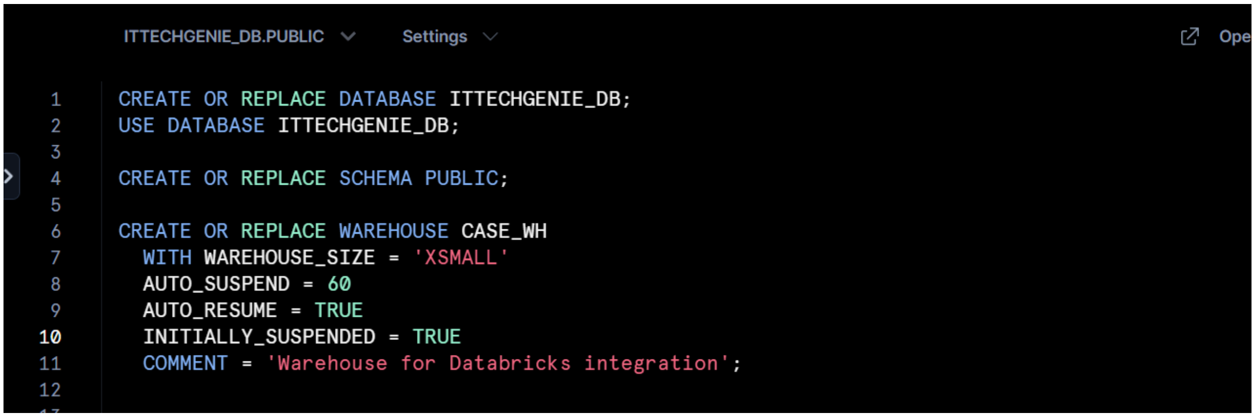
- Uppercased Category.

- Casted numeric columns to correct types.

- Created summary table by category.

Step 4: Snowflake Setup

- Create Warehouse: CASE_WH
- Create Database: ITTECHGENIE_DB
- Create Schema: PUBLIC
- Create Stage

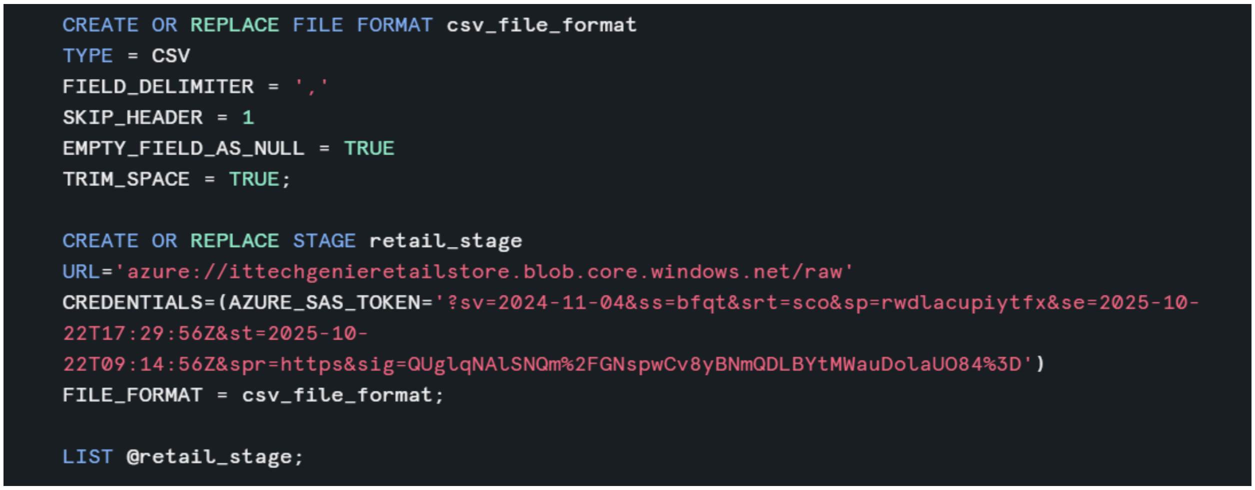


```
CREATE OR REPLACE DATABASE ITTECHGENIE_DB;
USE DATABASE ITTECHGENIE_DB;

CREATE OR REPLACE SCHEMA PUBLIC;

CREATE OR REPLACE WAREHOUSE CASE_WH
  WITH WAREHOUSE_SIZE = 'XSMALL'
  AUTO_SUSPEND = 60
  AUTO_RESUME = TRUE
  INITIALLY_SUSPENDED = TRUE
  COMMENT = 'Warehouse for Databricks integration';

```

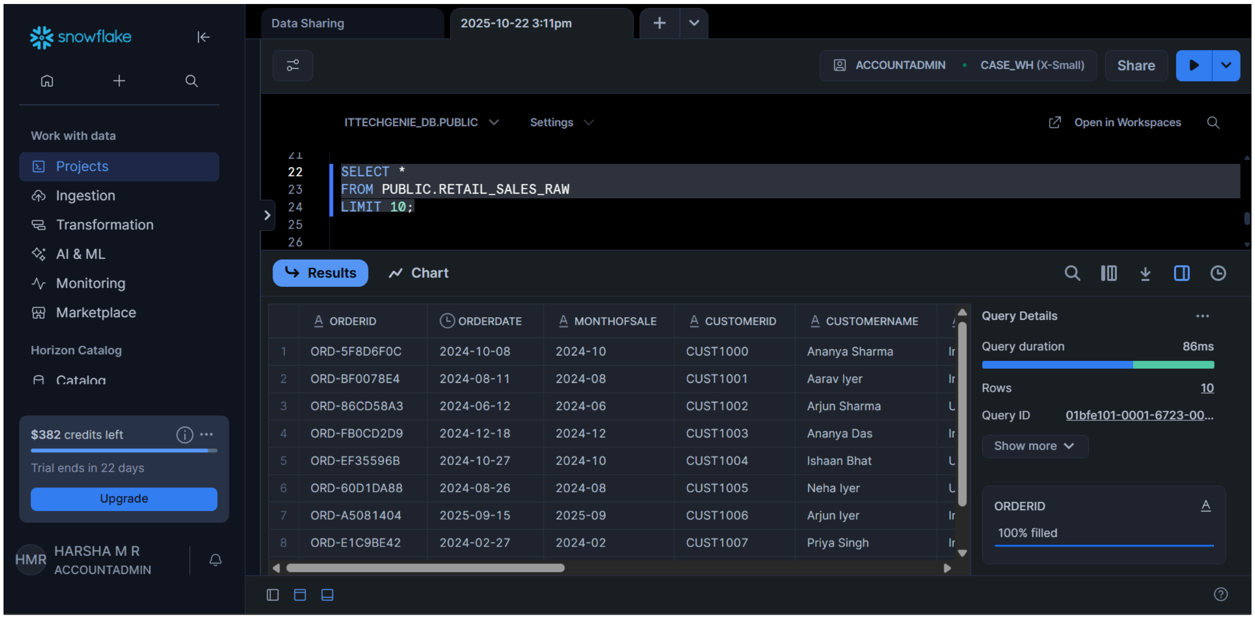


```
CREATE OR REPLACE FILE FORMAT csv_file_format
TYPE = CSV
FIELD_DELIMITER = ','
SKIP_HEADER = 1
EMPTY_FIELD_AS_NULL = TRUE
TRIM_SPACE = TRUE;

CREATE OR REPLACE STAGE retail_stage
URL='azure://ittechgenieretailstore.blob.core.windows.net/raw'
CREDENTIALS=(AZURE_SAS_TOKEN='?sv=2024-11-04&ss=bfqt&srt=sco&sp=rwdlacupiytfx&se=2025-10-
22T17:29:56Z&st=2025-10-
22T09:14:56Z&spr=https&sig=QUglqNALSNQm%2FGNspwCv8yBNmQDLBYtMWauDolaUO84%3D')
FILE_FORMAT = csv_file_format;

LIST @retail_stage;
```

Step 6: Verify Data in Snowflake



```
SELECT *
FROM PUBLIC.RETAIL_SALES_RAW
LIMIT 10;
```

	ORDERID	ORDERDATE	MONTHOFSALE	CUSTOMERID	CUSTOMERNAME	
1	ORD-5F8D6F0C	2024-10-08	2024-10	CUST1000	Ananya Sharma	Ir
2	ORD-BF0078E4	2024-08-11	2024-08	CUST1001	Aarav Iyer	Ir
3	ORD-86CD58A3	2024-06-12	2024-06	CUST1002	Arjun Sharma	U
4	ORD-FB0CD2D9	2024-12-18	2024-12	CUST1003	Ananya Das	Ir
5	ORD-EF35596B	2024-10-27	2024-10	CUST1004	Ishaan Bhat	U
6	ORD-60D1DA88	2024-08-26	2024-08	CUST1005	Neha Iyer	U
7	ORD-A5081404	2025-09-15	2025-09	CUST1006	Arjun Iyer	Ir
8	ORD-E1C9BE42	2024-02-27	2024-02	CUST1007	Priya Singh	Ir

Query Details

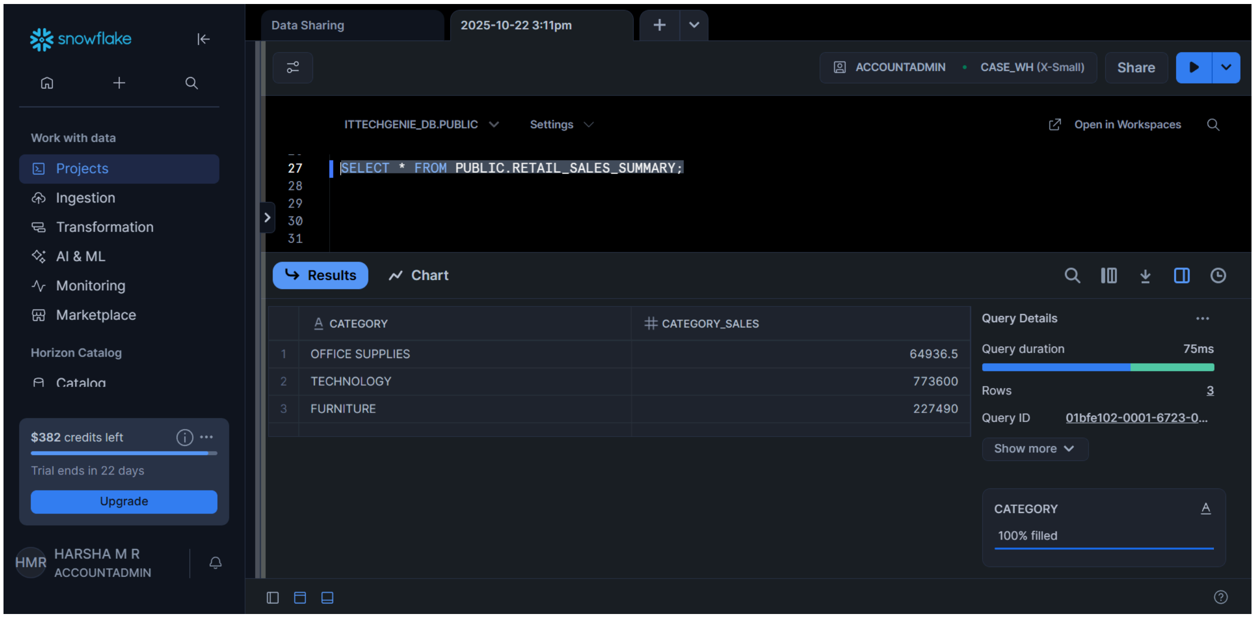
Query duration 86ms

Rows 10

Query ID 01bfe101-0001-6723-00...

ORDERID

100% filled



```
SELECT * FROM PUBLIC.RETAIL_SALES_SUMMARY;
```

	CATEGORY	CATEGORY_SALES
1	OFFICE SUPPLIES	64936.5
2	TECHNOLOGY	773600
3	FURNITURE	227490

Query Details

Query duration 75ms

Rows 3

Query ID 01bfe102-0001-6723-0...

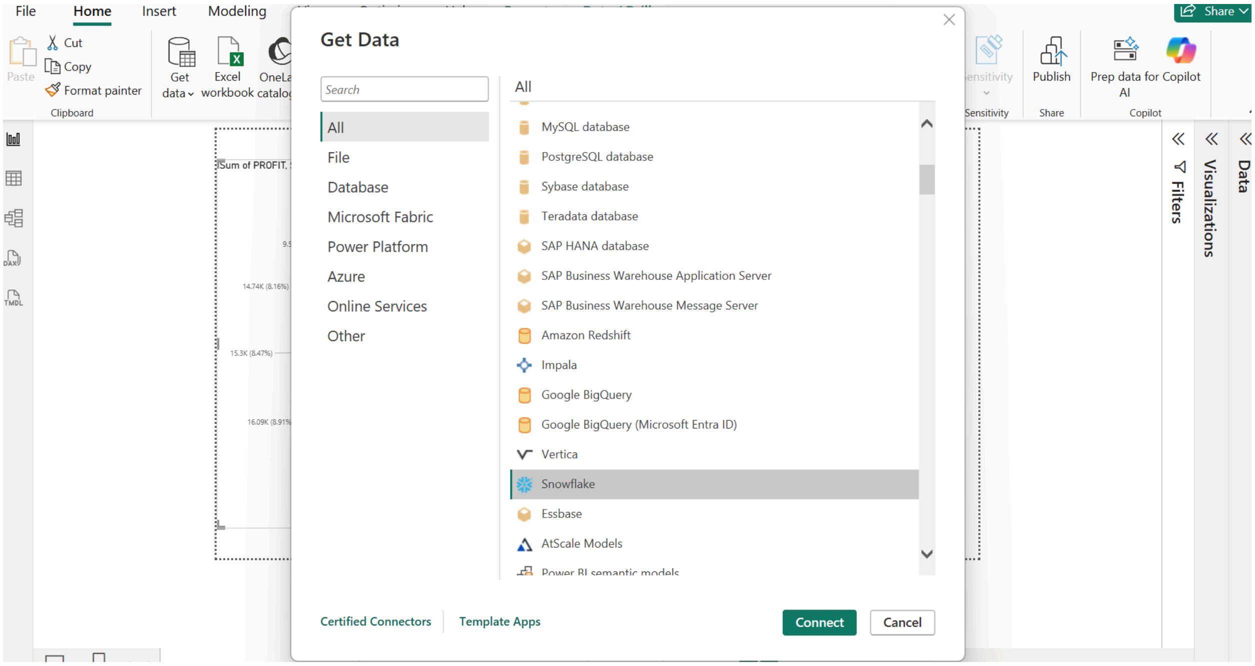
Show more

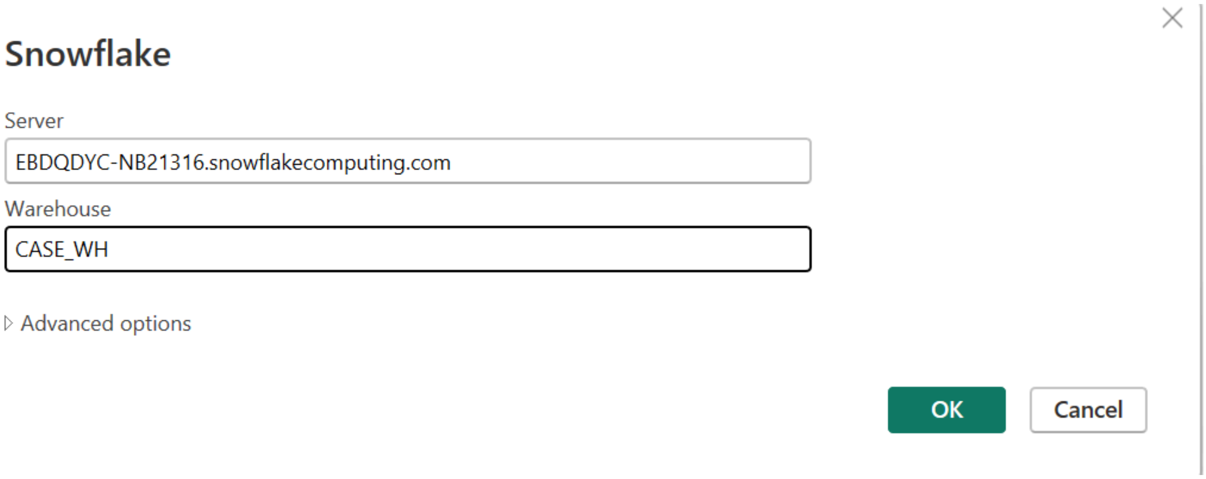
CATEGORY

100% filled

Step 7: Power BI Connection

- Open Power BI Desktop → Get Data → Snowflake.
- Enter Server, Warehouse, Database, Schema.
- Load RETAIL_SALES_RAW and RETAIL_SALES_SUMMARY.





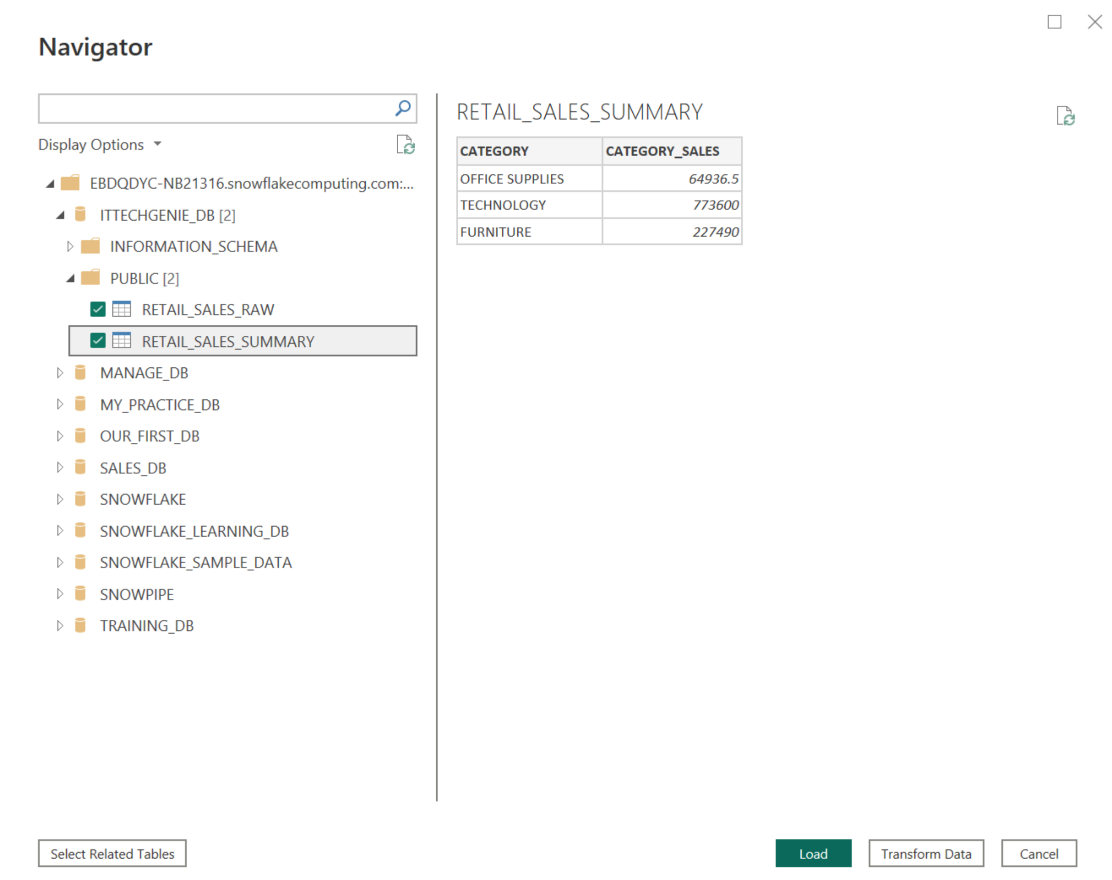
Snowflake

Server

EBDQDYC-NB21316.snowflakecomputing.com

Warehouse

CASE_WH



Navigator

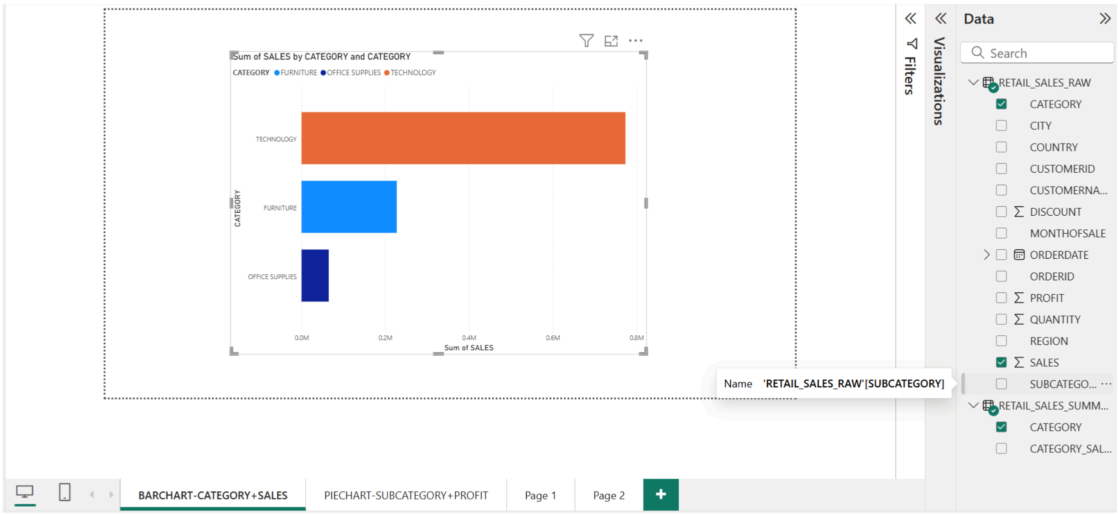
Display Options

- EBDQDYC-NB21316.snowflakecomputing.com:...
  - ITTECHGENIE_DB [2]
    - INFORMATION_SCHEMA
    - PUBLIC [2]
      - ☑ RETAIL_SALES_RAW
      - ☑ RETAIL_SALES_SUMMARY
  - MANAGE_DB
  - MY_PRACTICE_DB
  - OUR_FIRST_DB
  - SALES_DB
  - SNOWFLAKE
  - SNOWFLAKE_LEARNING_DB
  - SNOWFLAKE_SAMPLE_DATA
  - SNOWPIPE
  - TRAINING_DB

RETAIL_SALES_SUMMARY

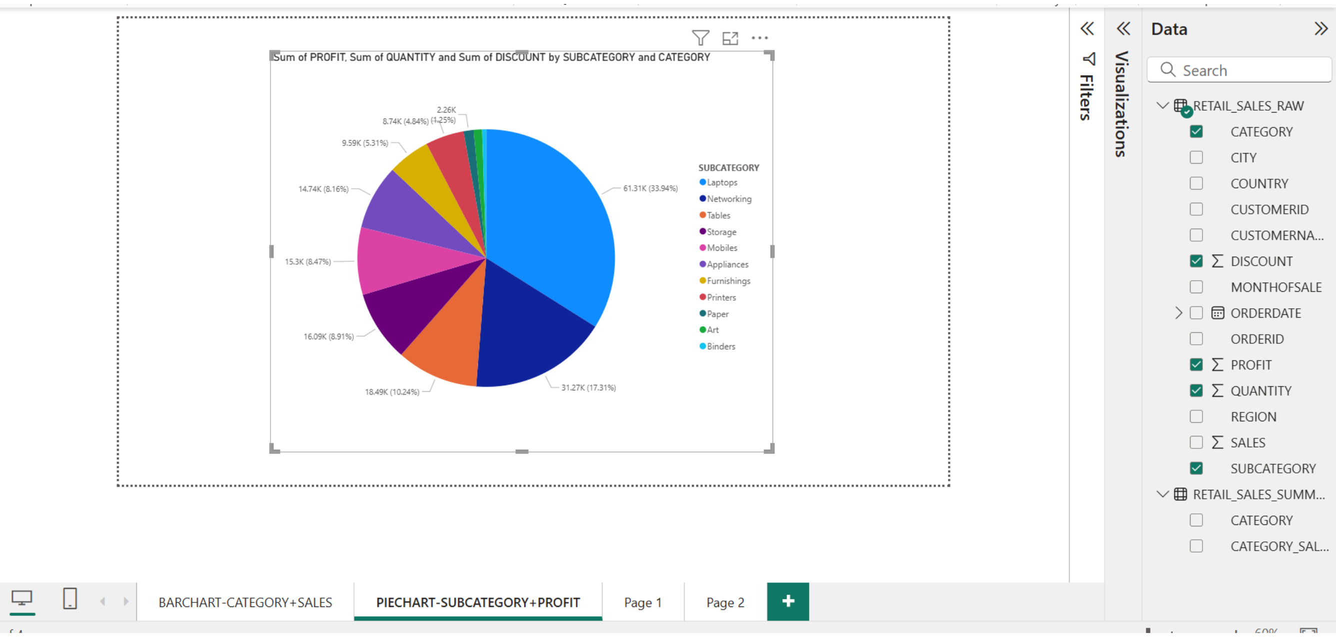
CATEGORY	CATEGORY_SALES
OFFICE SUPPLIES	64936.5
TECHNOLOGY	773600
FURNITURE	227490

Step 8: Create Power BI Visuals



Sum of SALES by CATEGORY and CATEGORY
CATEGORY ● FURNITURE ● OFFICE SUPPLIES ● TECHNOLOGY
CATEGORY
TECHNOLOGY
FURNITURE
OFFICE SUPPLIES
0.0M
0.2M
0.4M
0.6M
0.8M
Sum of SALES
Filters
Visualizations
Data
Search
Name 'RETAIL_SALES_RAW'[SUBCATEGORY]
BARCHART-CATEGORY+SALES
PIECHART-SUBCATEGORY+PROFIT
Page 1
Page 2

- RETAIL_SALES_RAW
  - ☑ CATEGORY
  - ☐ CITY
  - ☐ COUNTRY
  - ☐ CUSTOMERID
  - ☐ CUSTOMERNA...
  - ☐ Σ DISCOUNT
  - ☐ MONTHOFSALE
  - ☐ ORDERDATE
  - ☐ ORDERID
  - ☐ Σ PROFIT
  - ☐ Σ QUANTITY
  - ☐ REGION
  - ☑ Σ SALES
  - ☐ SUBCATEGO...
- RETAIL_SALES_SUMM...
  - ☑ CATEGORY
  - ☐ CATEGORY_SAL...

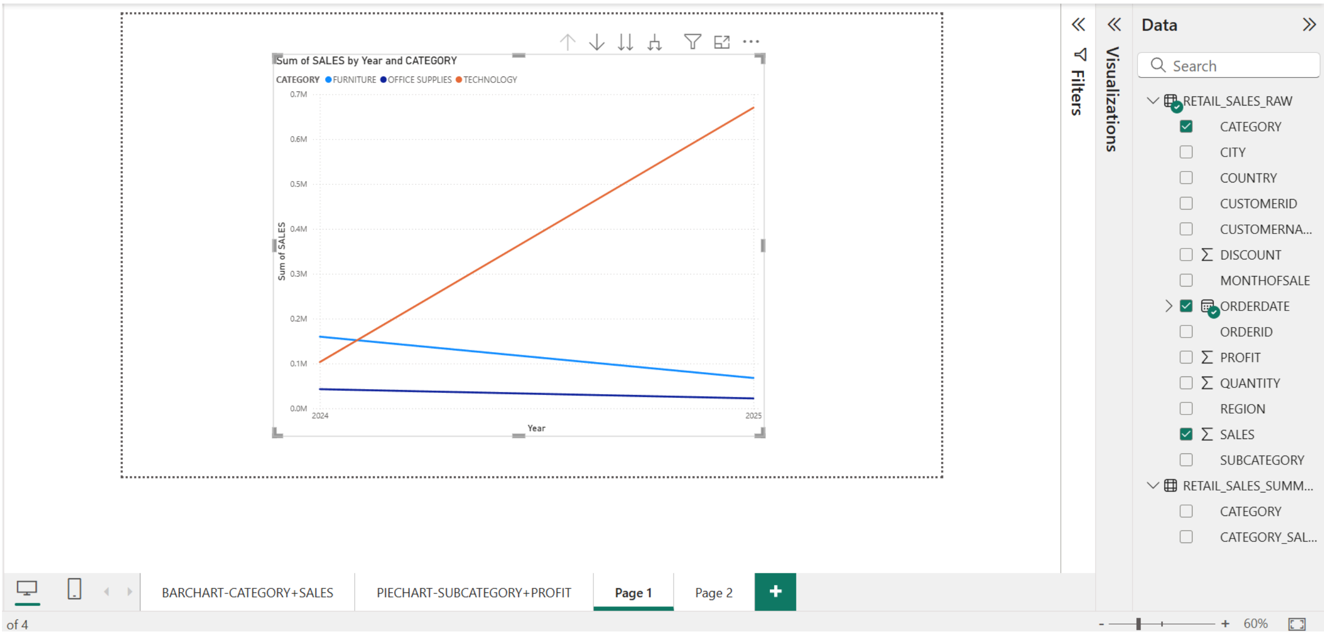


Data

Search

RETAIL_SALES_RAW
- ☑ CATEGORY
- □ CITY
- □ COUNTRY
- □ CUSTOMERID
- □ CUSTOMERNA...
- ☑ Σ DISCOUNT
- □ MONTHOFSALE
- □ ORDERDATE
- □ ORDERID
- ☑ Σ PROFIT
- ☑ Σ QUANTITY
- □ REGION
- □ Σ SALES
- ☑ SUBCATEGORY

RETAIL_SALES_SUMM...
- □ CATEGORY
- □ CATEGORY_SAL...

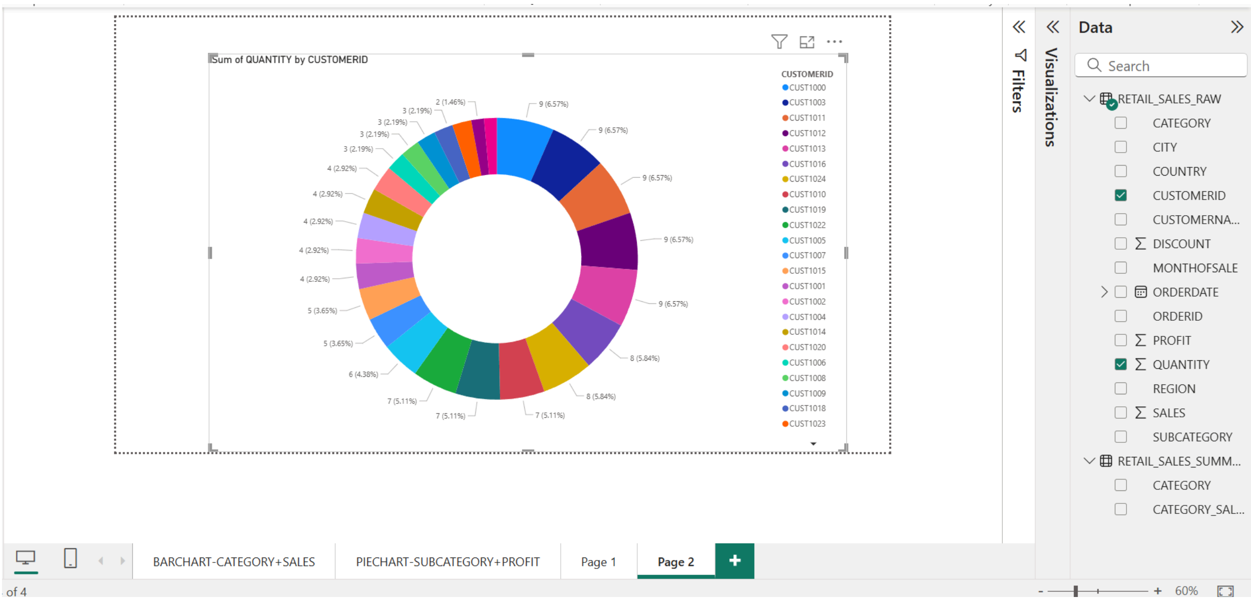


Data

Search

RETAIL_SALES_RAW
- ☑ CATEGORY
- □ CITY
- □ COUNTRY
- □ CUSTOMERID
- □ CUSTOMERNA...
- □ Σ DISCOUNT
- □ MONTHOFSALE
- ☑ ORDERDATE
- □ ORDERID
- □ Σ PROFIT
- □ Σ QUANTITY
- □ REGION
- ☑ Σ SALES
- □ SUBCATEGORY

RETAIL_SALES_SUMM...
- □ CATEGORY
- □ CATEGORY_SAL...



Data

Search

RETAIL_SALES_RAW
- □ CATEGORY
- □ CITY
- □ COUNTRY
- ☑ CUSTOMERID
- □ CUSTOMERNA...
- □ Σ DISCOUNT
- □ MONTHOFSALE
- □ ORDERDATE
- □ ORDERID
- □ Σ PROFIT
- ☑ Σ QUANTITY
- □ REGION
- □ Σ SALES
- □ SUBCATEGORY

RETAIL_SALES_SUMM...
- □ CATEGORY
- □ CATEGORY_SAL...