

# **END TO END DATA ENGINEERING PROJECT**

**BY:**

I. Ashik Roshan

Associate Data Engineer

## **PROJECT DESCRIPTION:**

My project involves gathering data from on-premises sources using Python. After that, I import this data into SQL Server Management Studio (SSMS) using Python. Following that, I move the data to Azure Blob Storage from SSMS using Azure Data Factory. Then, I bring the data from Azure Blob Storage into Snowflake, using shared access security keys, and perform necessary transformations within Snowflake. Lastly, I load the transformed data into Power BI from Snowflake and create a dashboard.

## **THE ESSENTIAL TOOLS REQUIRED FOR THE PROJECT:**

### **PYTHON:**

Python is a versatile and widely-used programming language. It's commonly used for data manipulation, analysis, and scripting tasks. In your project, Python can be used for data extraction, transformation, and scripting.

### **SQL SERVER MANAGEMENT STUDIO (SSMS):**

SSMS is a Microsoft application used for managing and interacting with Microsoft SQL Server databases. It provides a graphical interface for database administration, querying, and development.

### **AZURE FREE TRIAL SUBSCRIPTION:**

Azure is Microsoft's cloud computing platform. A free trial subscription provides you with access to Azure services, allowing you

to host and manage your applications, data, and infrastructure in the cloud.

### **AZURE BLOB STORAGE:**

Azure Blob Storage is a cloud-based object storage service in Microsoft Azure. It's used for storing and managing unstructured data, such as documents, images, and backup files. In your project, it can be a repository for your data files.

### **Azure Data Factory:**

Azure Data Factory is a cloud-based data integration service that allows you to create, schedule, and manage data-driven workflows. It's used for data movement and data transformation tasks in the cloud.

### **SNOWFLAKE:**

Snowflake is a cloud-based data warehousing platform that offers scalable and fully-managed data storage and analytics. It's designed for data warehousing and data analytics, including data transformation and analysis.

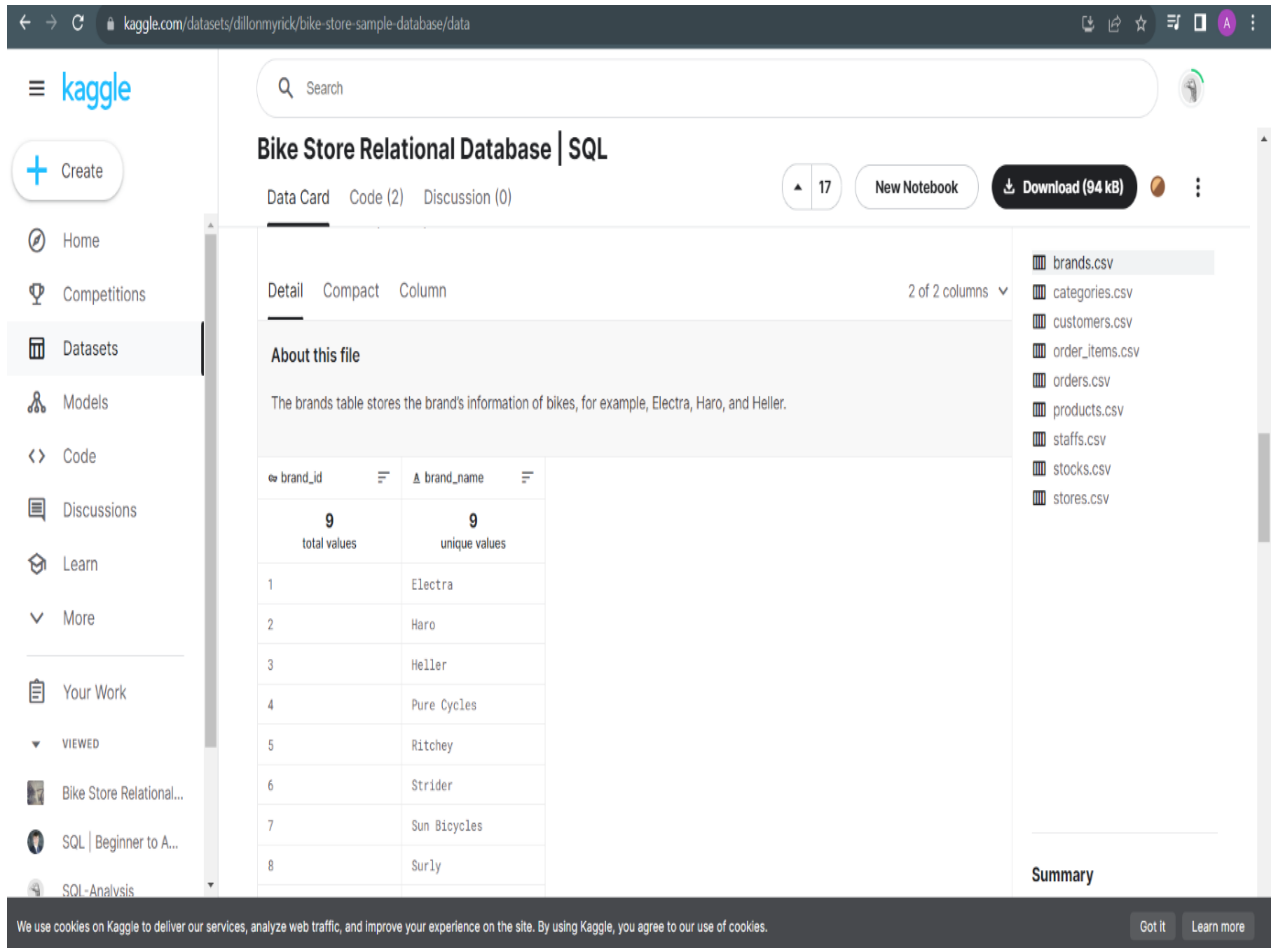
### **POWER BI:**

Power BI is a business intelligence and data visualization tool by Microsoft. It allows you to connect to various data sources, create interactive reports and dashboards, and share insights with others.

## STAGE BY STAGE PROCEDURE:

### ➤ TO OBTAIN THE DATA SOURCE:

1. Visit Kaggle or use Google Dataset Search website.
2. Use the search Engine to find a dataset with relational data.
3. Select a relevant dataset and download it.
4. Securely save the downloaded dataset to your local storage.



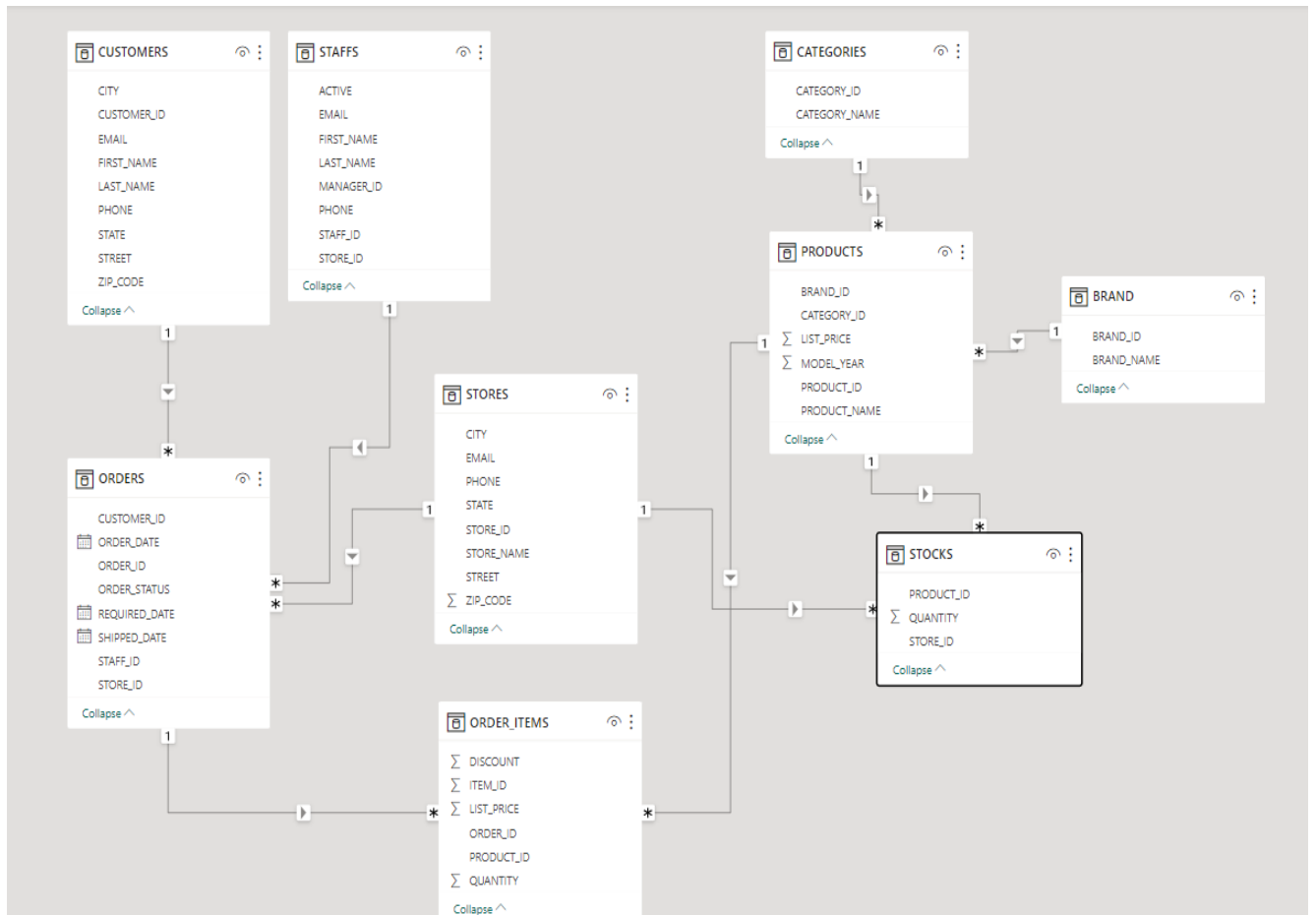
The screenshot shows the Kaggle website interface for the 'Bike Store Relational Database | SQL' dataset. The page is titled 'Bike Store Relational Database | SQL' and includes a search bar, a 'Create' button, and a sidebar with navigation options like Home, Competitions, Datasets, Models, Code, Discussions, Learn, More, Your Work, and VIEWED. The main content area displays the dataset details, including a table with 2 columns: 'brand\_id' and 'brand\_name'. The table has 9 rows of data, including brands like Electra, Haro, Heller, Pure Cycles, Ritchey, Strider, Sun Bicycles, and Surly. The page also shows a sidebar with navigation options and a list of CSV files available for download.

brand_id	brand_name
1	Electra
2	Haro
3	Heller
4	Pure Cycles
5	Ritchey
6	Strider
7	Sun Bicycles
8	Surly

Summary

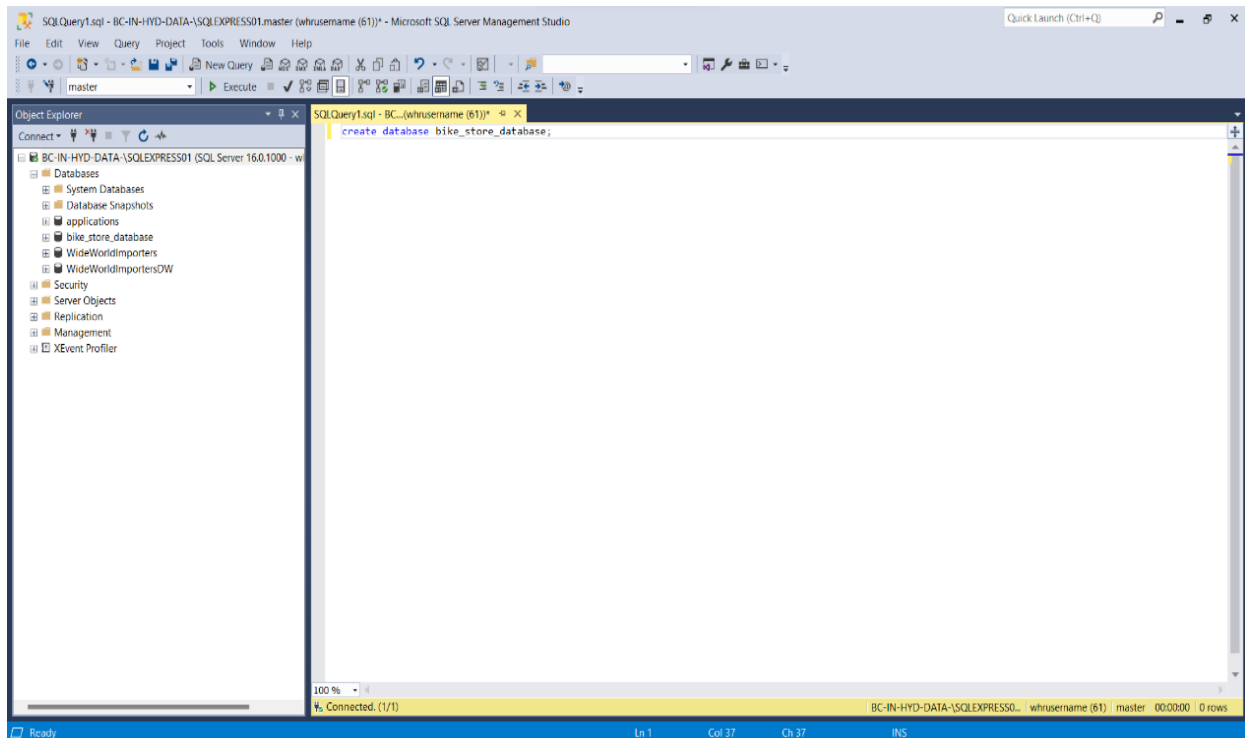
## ➤ DATA MODELING WITH POWER BI:

1. Import Data: Start by bringing your data into Power BI from Data sources.
2. Tables and Relationships: Organize data into tables and connect them together.
3. Primary and Foreign Keys: Use unique IDs to link tables together.

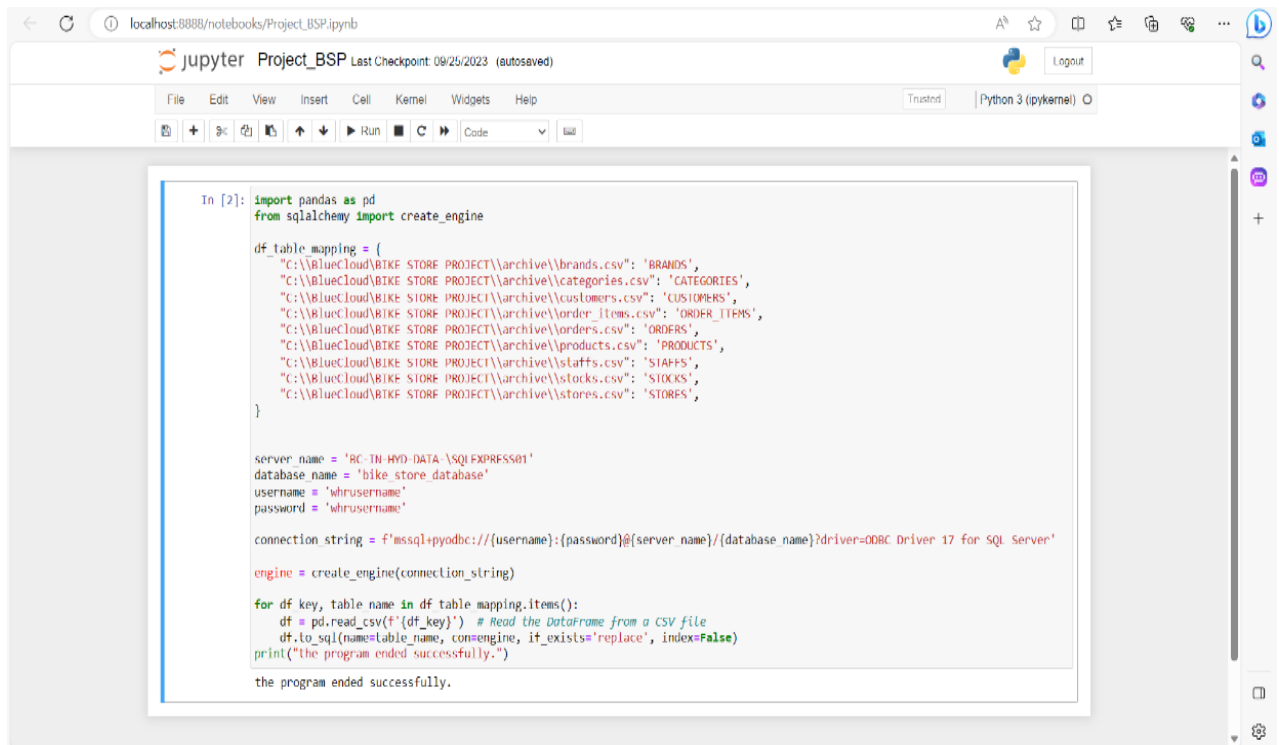


## ➤ LOADING ON-PREMISES DATA INTO SSMS WITH PYTHON:

1. Open SSMS and Create a Database.



2. Begin by opening your preferred Integrated Development Environment (IDE) for running Python scripts.
3. In this project, the chosen tool for running Python scripts is **Jupyter Notebook**.
4. Open Jupyter Notebook to load data from an on-premises source into SQL Server Management Studio (SSMS). Following this, I'll provide a detailed overview of the Python script used for this purpose in the subsequent points.



```
In [2]: import pandas as pd
from sqlalchemy import create_engine

df_table_mapping = {
    "C:\\BlueCloud\\BIKE STORE PROJECT\\archive\\brands.csv": 'BRANDS',
    "C:\\BlueCloud\\BIKE STORE PROJECT\\archive\\categories.csv": 'CATEGORIES',
    "C:\\BlueCloud\\BIKE STORE PROJECT\\archive\\customers.csv": 'CUSTOMERS',
    "C:\\BlueCloud\\BIKE STORE PROJECT\\archive\\order_items.csv": 'ORDER_ITEMS',
    "C:\\BlueCloud\\BIKE STORE PROJECT\\archive\\orders.csv": 'ORDERS',
    "C:\\BlueCloud\\BIKE STORE PROJECT\\archive\\products.csv": 'PRODUCTS',
    "C:\\BlueCloud\\BIKE STORE PROJECT\\archive\\staffs.csv": 'STAFFS',
    "C:\\BlueCloud\\BIKE STORE PROJECT\\archive\\stocks.csv": 'STOCKS',
    "C:\\BlueCloud\\BIKE STORE PROJECT\\archive\\stores.csv": 'STORES',
}

server_name = 'BC-IN-HYD-DATA\\SQL EXPRESS01'
database_name = 'bike_store_database'
username = 'whrusername'
password = 'whrusername'

connection_string = f'mssql+pyodbc://{username}:{password}@{server_name}/{database_name}?driver=ODBC Driver 17 for SQL Server'

engine = create_engine(connection_string)

for df_key, table_name in df_table_mapping.items():
    df = pd.read_csv(df_key) # Read the Dataframe from a CSV file
    df.to_sql(name=table_name, con=engine, if_exists='replace', index=False)
print("the program ended successfully.")

the program ended successfully.
```

## 5. Python Script Idea:

### Import necessary libraries:

Import Pandas and SQLAlchemy to work with data and databases.

### Create a dictionary:

Define a dictionary (df\_table\_mapping) that connects CSV file paths to database table names.

### Set up database connection details:

Define server name, database name, username, and password for connecting to the SQL Server database.

### Build a connection string:

Create a connection string using the database configuration details.

### Create an SQLAlchemy engine:

Set up a connection engine using the connection string.

## Loop through CSV files and tables:

For each CSV file and its corresponding table name in the dictionary:

- Read the data from the CSV file into a Pandas DataFrame.
- Write the DataFrame's data into the specified SQL Server table.
- If the table already exists, replace its contents with the new data.
- Print a success message: Indicate that the program has finished successfully.

6.Open SSMS and check Data is Loaded or not.

The screenshot shows the Microsoft SQL Server Enterprise Manager (SSMS) interface. The left pane displays the Object Explorer with the 'bike\_store\_database' selected. The right pane shows a query window with the following SQL query:

```
--create database bike_store_database;

select * from [dbo].[BRANDS];
select * from [dbo].[CATEGORIES];
select * from [dbo].[CUSTOMERS];
select * from [dbo].[ORDER_ITEMS];
select * from [dbo].[ORDERS];
select * from [dbo].[PRODUCTS];
select * from [dbo].[STAFFS];
select * from [dbo].[STOCKS];
select * from [dbo].[STORES];
```

The query results are displayed in the bottom pane, showing two tables. The first table is a list of staff members, and the second table is a list of products.

staff_id	first_name	last_name	email	phone	active	store_id	manager_id
3	Gemma	Serrano	gemma.serrano@bikes shop	(831) 555-5556	1	1	2
4	Virgie	Wiggins	virgie.wiggins@bikes shop	(831) 555-5557	1	1	2
5	Jannette	David	jannette.david@bikes shop	(516) 379-4444	1	2	1
6	Marcelene	Boyer	marcelene.boyer@bikes shop	(516) 379-4445	1	2	5
7	Venita	Daniel	venita.daniel@bikes shop	(516) 379-4446	1	2	5
8	Kali	Vargas	kali.vargas@bikes shop	(972) 530-5555	1	3	1
9	Layla	Terrell	layla.terrell@bikes shop	(972) 530-5556	1	3	7
10	Bernard	Houston	bernardine.houston@bikes...	(972) 530-5557	1	3	7

store_id	product_id	quantity
40	1	24
41	1	10
42	1	0
43	1	2
44	1	1
45	1	15
46	1	19
47	1	21

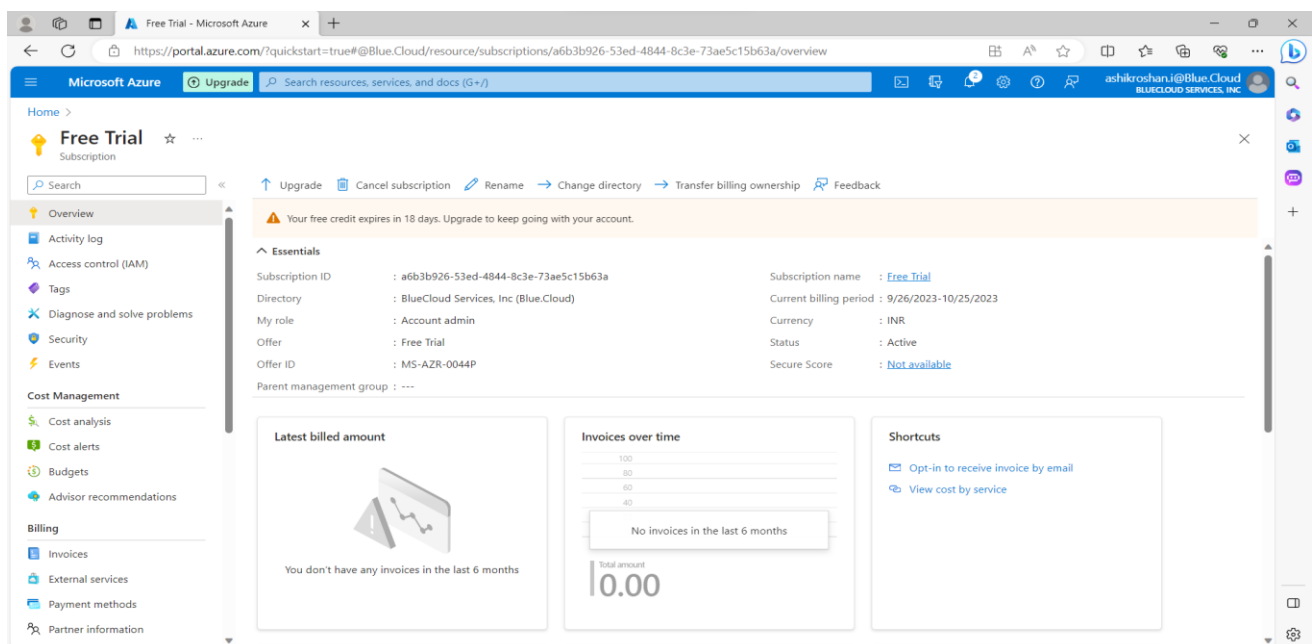
store_id	store_name	phone	email	street	city	state	zip_code
1	Santa Cruz Bikes	(831) 476-4321	santacruz@bikes shop	3700 Portola Drive	Santa Cruz	CA	95060
2	Baldwin Bikes	(516) 379-8888	baldwin@bikes shop	4200 Chestnut Lane	Baldwin	NY	11432
3	Rowlett Bikes	(972) 530-5555	rowlett@bikes shop	8000 Fairway Aven...	Rowlett	TX	75088

The status bar at the bottom indicates that the query was executed successfully, returning 3 rows.



## ➤ CREATE AN AZURE FREE TRIAL SUBSCRIPTION:

1. Go to Azure Free Trial Page: Visit [Azure Free Trial](#).
2. Sign In/Create Account: Sign in with your Microsoft account or create one.
3. Provide Info: Fill in your details and a payment method.
4. Verify Identity: Confirm your identity with a code.
5. Accept Terms: Agree to the terms.
6. Add Payment Info: Enter card details (no charges for now).
7. Activate Free Trial: Get free Azure credits.
8. Access Azure Portal: Go to [Azure Portal](#).
9. Explore Azure: Try out Azure services with your free credits.
10. Check Usage: Keep an eye on your credit balance.
11. Upgrade or Cancel: Upgrade if you need more or cancel when done.



## ➤ CREATE AN AZURE BLOB STORAGE:

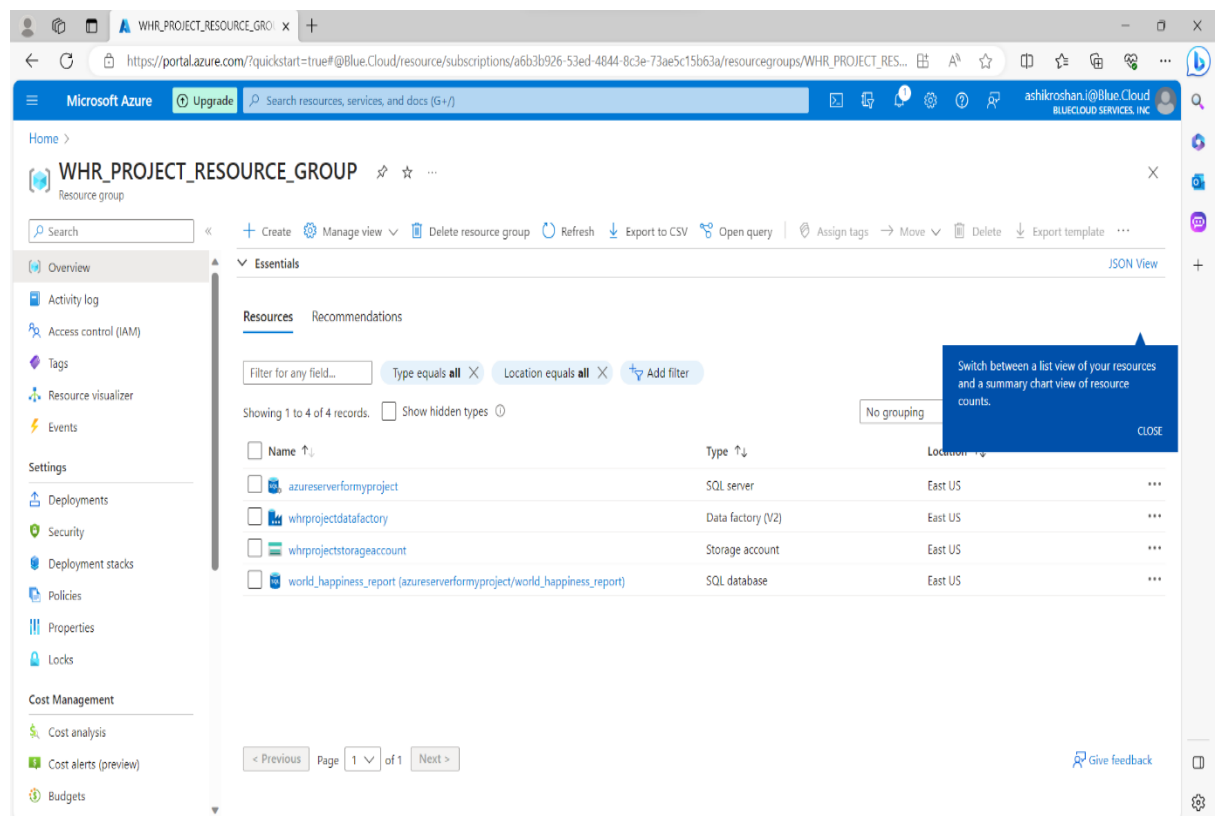
### SIGN IN OR CREATE AN AZURE ACCOUNT:

Log in to the Azure Portal or create an account if you don't have one.

### CREATE A RESOURCE GROUP:

Before creating the storage account, consider creating a resource group to help organize your resources.

Go to "Create a resource group" and provide a name and region for the resource group.



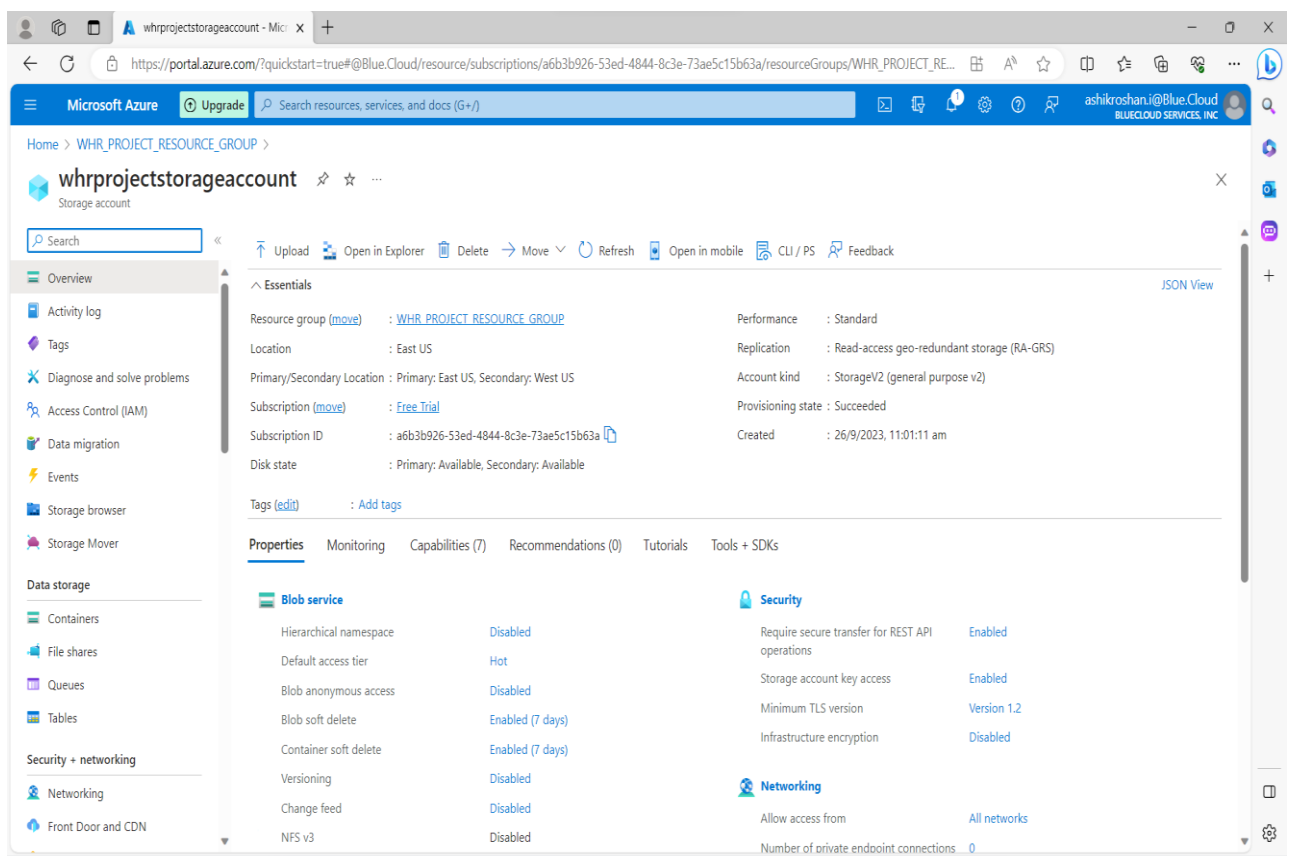
### CREATE A STORAGE ACCOUNT:

1. Go to "Create a resource."

2. Search for "Storage account" and select it.

3. Click "Create" and fill in the details:

- Subscription
- Resource Group (choose the one you just created or an existing one)
- Unique Storage account name
- Region
- Performance (Standard or Premium)
- Replication (data redundancy)
- Review and click "Create."

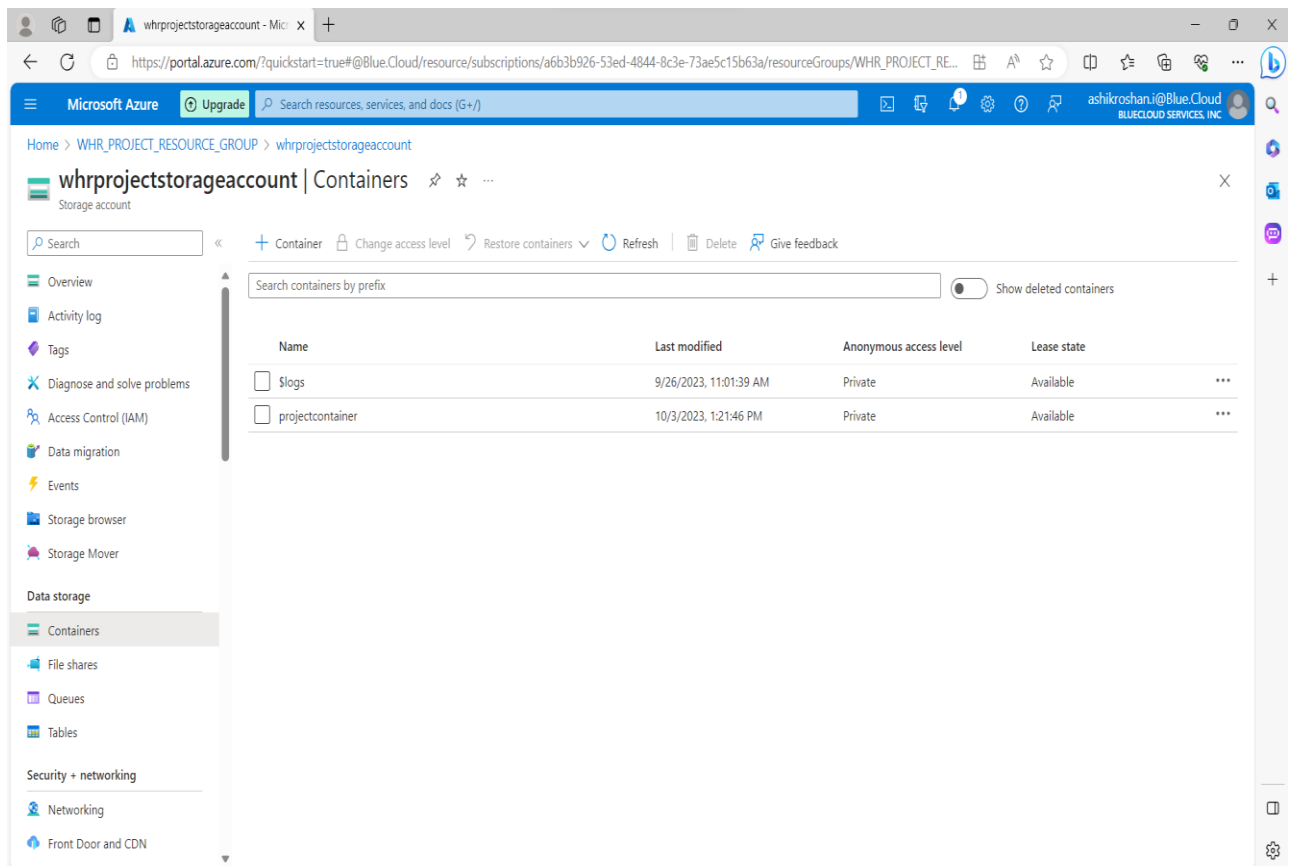


## ACCESS BLOB STORAGE:

Find and select your newly created storage account.

## CREATE BLOB CONTAINERS:

- In your storage account, go to "Containers."
- Click "+ Container" to make a new one.
- Give it a unique name and choose access level (e.g., private).

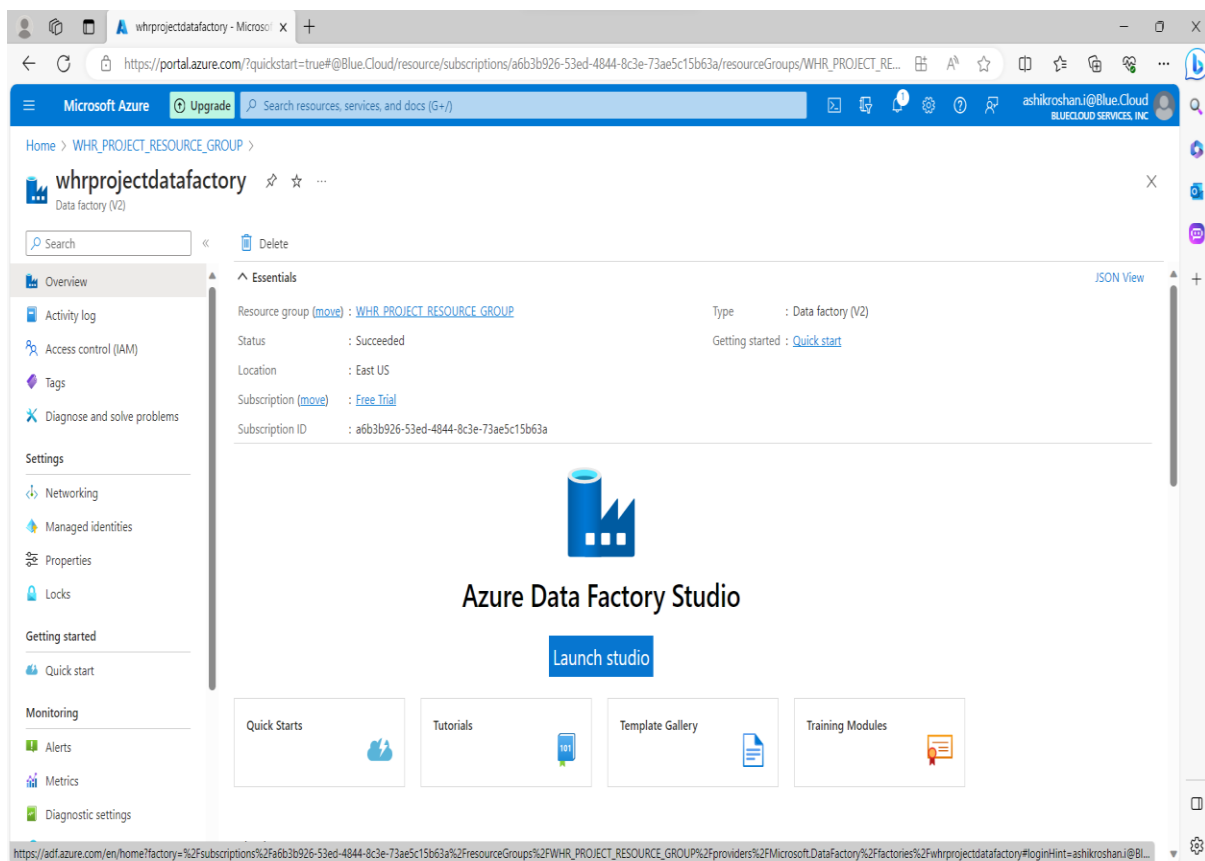


## MANAGE DATA:

- Set permissions and access controls as needed.
- Manage your data using Azure tools or SDKs.

## ➤ CREATE AN AZURE DATA FACTORY:

1. Click on "Your Resource Group" in the Azure Portal.
2. Search for "Data Factory" and select.
3. Click the "Create" button.
4. Configure Basic Settings:
  - Fill the required details for your Data Factory:
    - Subscription: Choose your Azure subscription.
    - Resource Group: Select an existing one or create a new one.
    - Region: Choose the region where you want to deploy your Data Factory.
    - Version: Select the version of Azure Data Factory (V1 or V2).



➤ **LOADING DATA FROM SSMS TO AZURE BLOB STORAGE WITH AZURE DATA FACTORY:**

- Open azure data factory.
- Create linked services:

In your Azure Data Factory, create linked services to connect to both your SSMS database and your Azure Blob Storage account.

- Create a Data Pipeline:

Inside your Data Factory, create a new data pipeline.

- Add activities:

In the data pipeline, add activities to define the data movement. For loading data from SSMS to Azure Blob Storage, Use activities such as “lookup”, “foreach”, “copy activity”.

- Configure Source Dataset:

In the "Copy Data" activity, configure the source dataset. This should be your SSMS database.

- Configure Sink Dataset:

Configure the sink dataset as your Azure Blob Storage account and specify the destination container and file settings.

- Schedule or Trigger the Pipeline:

Set up a schedule or trigger to specify when and how often the data transfer should occur.

- Debug and Test:

Use the debug functionality within Azure Data Factory to test your pipeline and ensure it's working as expected.

- Monitor and Troubleshoot:

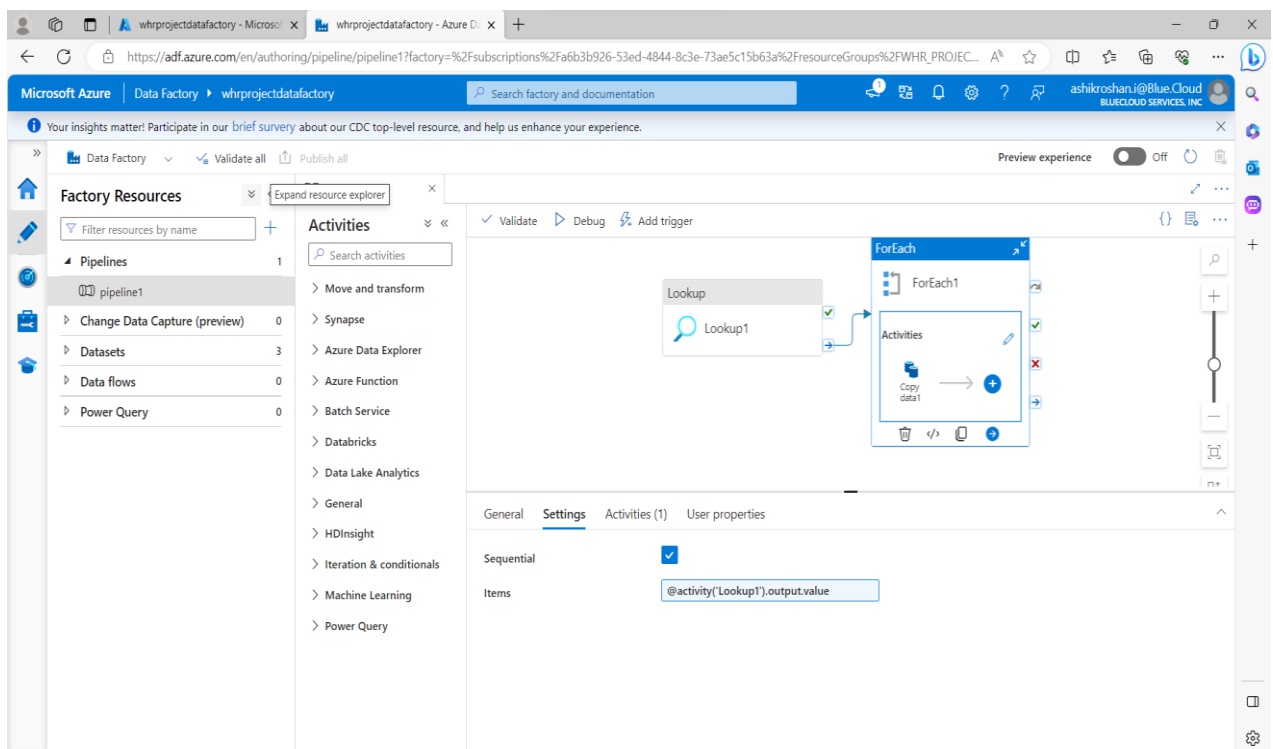
Monitor the pipeline's execution to track progress and troubleshoot any errors or issues.

- Run the Pipeline:

Execute the pipeline to initiate the data transfer from SSMS to Azure Blob Storage.

- Verify Data in Blob Storage:

After the pipeline has run successfully, check your Azure Blob Storage container to ensure the data has been loaded correctly.



➤ **CREATE AN SNOWFLAKE ACCOUNT:**

1. Go to the Snowflake website at <https://www.snowflake.com/>.
2. Look for a "Free Trial" or "Try Snowflake for Free" option on the website's homepage or in the navigation menu.
3. Fill out the registration form with your personal and contact information. You may need to provide details such as your name, email address, phone number, and company name.
4. You may be required to verify your email address by clicking on a confirmation link sent to your email.
5. Select your preferred cloud platform for Snowflake (e.g., AWS, Azure, GCP). Snowflake is available on multiple cloud providers.
6. Follow the on-screen instructions to configure your Snowflake account settings. This typically includes setting a username and password.
7. After completing the registration process, you should be able to access your Snowflake trial account through a web-based interface.
8. Once you've logged in, you can start exploring Snowflake's features, including data loading, SQL querying, and data warehousing.



ARI

ASHIK Roshan I

ACCOUNTADMIN

Worksheets

Dashboards

Streamlit

Apps

Data

Marketplace

Activity

Admin

Help & Support

L

2 days left in trial

Upgrade

VW74842

app.snowflake.com/uugogwd/vw74842/worksheets

☆

A

Worksheets

Q Search

...

+

Recent

Shared with me

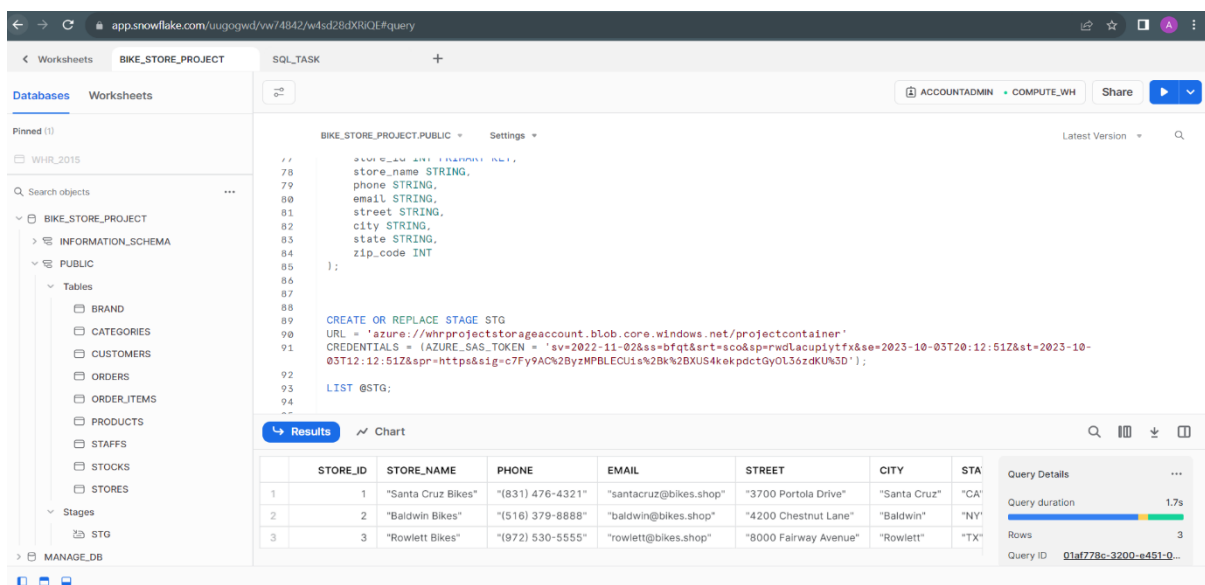
My Worksheets

Folders

TITLE	TYPE	VIEWED ↓	UPDATED	ROLE
<div></div> BIKE_STORE_PROJE...	SQL	just now	1 minute ago	ACCOUNTADMIN
<div></div> SQL_TASK	SQL	1 day ago	1 minute ago	ACCOUNTADMIN
<div></div> 2023-09-28 5:30pm	SQL	1 week ago	1 week ago	ACCOUNTADMIN
<div></div> 2023-09-28 5:07pm	SQL	1 week ago	1 week ago	ACCOUNTADMIN
<div></div> Tutorial 1: Sa... Bench...	SQL	—	4 weeks ago	ACCOUNTADMIN
<div></div> Tutorial 2: Sa... Bench...	SQL	—	4 weeks ago	ACCOUNTADMIN
<div></div> Tutorial 3: TP... Bench...	SQL	—	4 weeks ago	ACCOUNTADMIN
<div></div> Tutorial 4: TP... Bench...	SQL	—	4 weeks ago	ACCOUNTADMIN
<div></div> Benchmarking Tutori...	Folder	—	4 weeks ago	—

## ➤ LOADING DATA FROM AZURE BLOB STORAGE TO SNOWFLAKE:

1. Open your snowflake account and Create a Snowflake SQL worksheet.
2. create Snowflake stage. This stage acts as a connection point between Snowflake and Azure Blob Storage.
3. In the stage creation code, specify the Azure Blob Storage URL that points to your container and provide the SAS (Shared Access Signature) token. The SAS token contains authentication and access permissions.
4. Create a File Format and Specify the File Format of the data.
5. Create a table and ensure that the columns have the appropriate data types as in the Azure blob storage file.
6. Use Snowflake's SQL “COPY INTO” command to initiate the data transfer from Stage to Targeted Table.



The screenshot displays the Snowflake SQL Worksheet interface. The left sidebar shows the database structure with the 'BIKE\_STORE\_PROJECT' database selected. The main area shows the SQL query and its results.

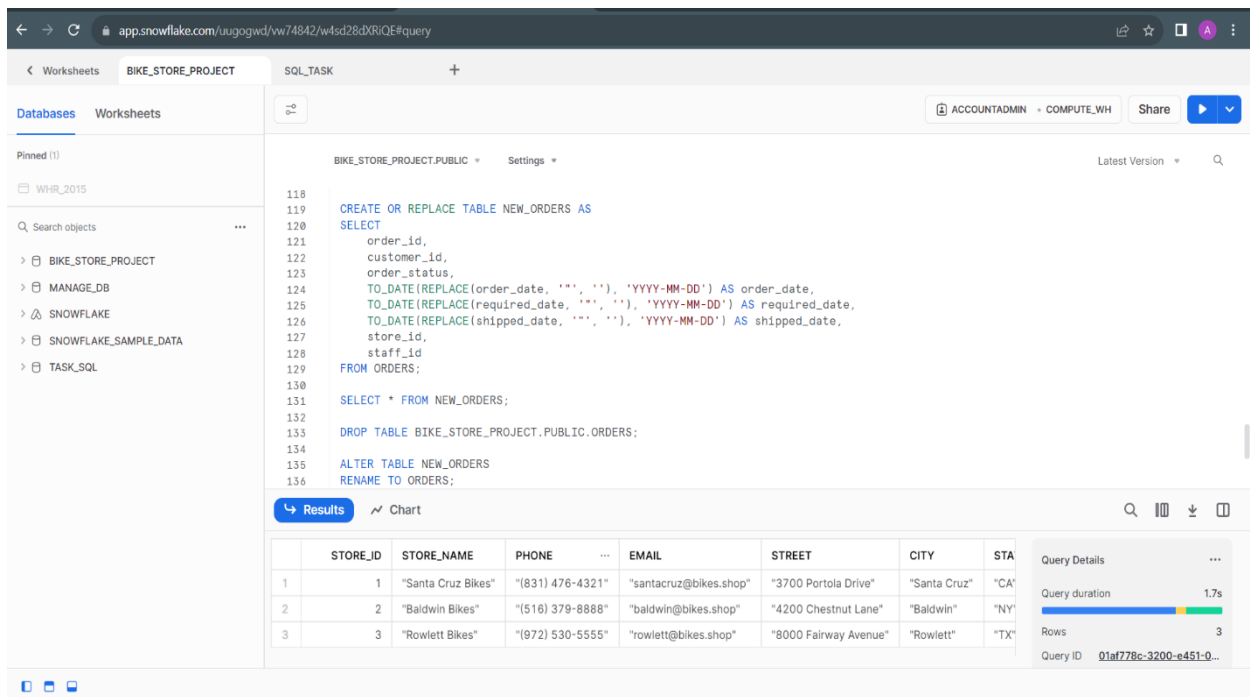
```
BIKE_STORE_PROJECT.PUBLIC > Settings >
78 store_name STRING,
79 phone STRING,
80 email STRING,
81 street STRING,
82 city STRING,
83 state STRING,
84 zip_code INT
85 };
86
87
88
89
90 CREATE OR REPLACE STAGE STG
91 URL = 'azure://whrprojectstorageaccount.blob.core.windows.net/projectcontainer'
92 CREDENTIALS = (AZURE_SAS_TOKEN = 'sv=2022-11-02&ss=bfq&sr=sco&sp=rwdlacuplyfx&se=2023-10-03T20:12:51Z&st=2023-10-03T12:12:51Z&spr=https&sig=c7Fy9AC%2ByzMPBLECUis%2Bk%2BXUS4kekpdctgyOL56zdKU%3D');
93
94 LIST @STG;
```

The results table shows the following data:

	STORE_ID	STORE_NAME	PHONE	EMAIL	STREET	CITY	STA
1	1	"Santa Cruz Bikes"	"(831) 476-4321"	"santacruz@bikes.shop"	"3700 Portola Drive"	"Santa Cruz"	"CA"
2	2	"Baldwin Bikes"	"(516) 379-8888"	"baldwin@bikes.shop"	"4200 Chestnut Lane"	"Baldwin"	"NY"
3	3	"Rowlett Bikes"	"(972) 530-5555"	"rowlett@bikes.shop"	"8000 Fairway Avenue"	"Rowlett"	"TX"

Query Details: Query duration 1.7s, Rows 3, Query ID 01a778c-3209-e451-0...

## ➤ PERFORM TRANSFORMATION ON THE TABLES WITH SNOWFLAKE SQL ACCORDING TO REQUIREMENT:



The screenshot shows the Snowflake web interface. The left sidebar displays the database structure with 'BIKE\_STORE\_PROJECT' selected. The main area shows a SQL query being executed. The query creates a new table 'NEW\_ORDERS' by selecting data from the 'ORDERS' table, transforming the date fields. Below the query, the 'Results' tab is active, showing a table with 3 rows and 8 columns. The 'Query Details' panel on the right indicates the query duration is 1.7s and the number of rows is 3.

```
118 CREATE OR REPLACE TABLE NEW_ORDERS AS
119 SELECT
120     order_id,
121     customer_id,
122     order_status,
123     TO_DATE(REPLACE(order_date, '-', ''), 'YYYY-MM-DD') AS order_date,
124     TO_DATE(REPLACE(required_date, '-', ''), 'YYYY-MM-DD') AS required_date,
125     TO_DATE(REPLACE(shipped_date, '-', ''), 'YYYY-MM-DD') AS shipped_date,
126     store_id,
127     staff_id
128 FROM ORDERS;
129
130 SELECT * FROM NEW_ORDERS;
131
132 DROP TABLE BIKE_STORE_PROJECT.PUBLIC.ORDERS;
133
134 ALTER TABLE NEW_ORDERS
135 RENAME TO ORDERS;
```

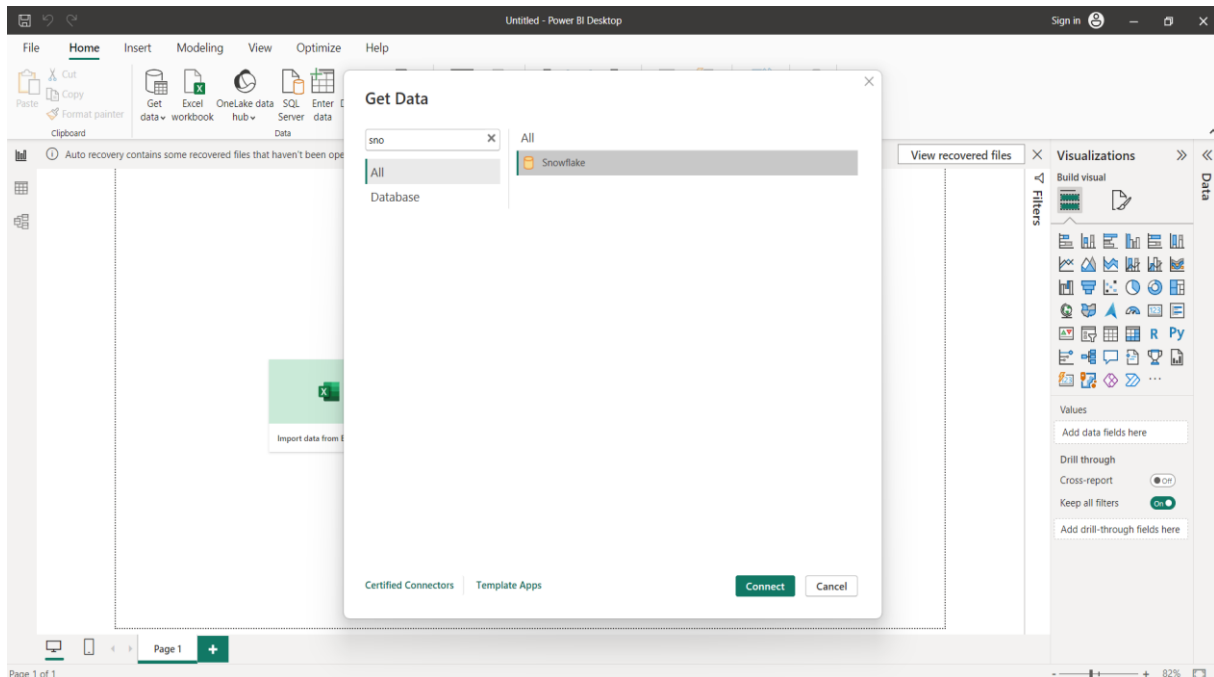
	STORE_ID	STORE_NAME	PHONE	EMAIL	STREET	CITY	STA
1	1	"Santa Cruz Bikes"	"(831) 476-4321"	"santacruz@bikes.shop"	"3700 Portola Drive"	"Santa Cruz"	"CA"
2	2	"Baldwin Bikes"	"(516) 379-8888"	"baldwin@bikes.shop"	"4200 Chestnut Lane"	"Baldwin"	"NY"
3	3	"Rowlett Bikes"	"(972) 530-5555"	"rowlett@bikes.shop"	"8000 Fairway Avenue"	"Rowlett"	"TX"

Query Details  
Query duration: 1.7s  
Rows: 3  
Query ID: 01af778c-3200-e451-0...

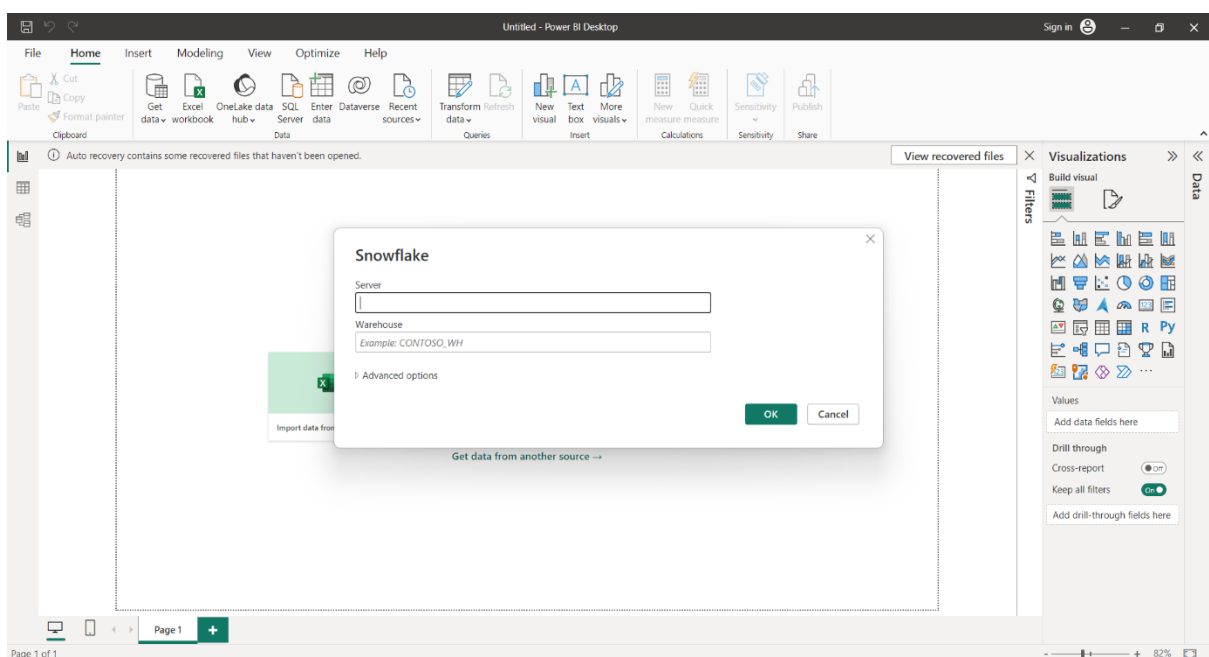
## ➤ LOADING DATA FROM SNOWFLAKE TO POWER BI:

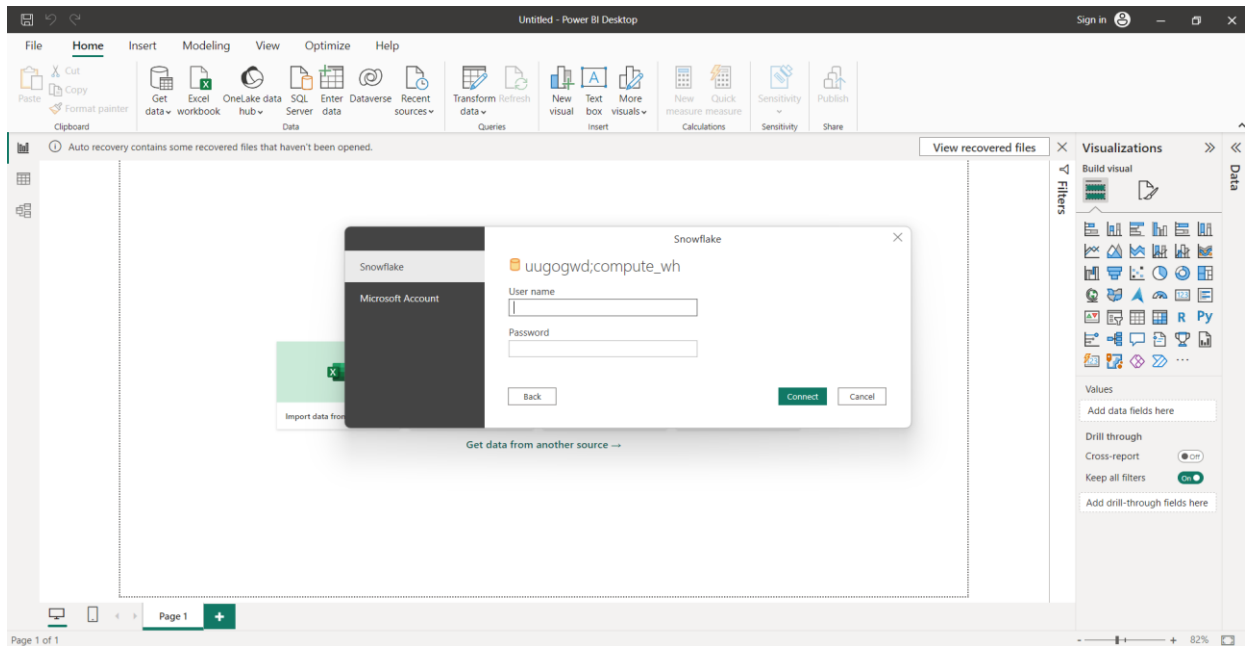
1. Ensure you have access to Snowflake and Power BI.
2. Install the Power BI desktop application on your computer if you haven't already.
3. Make sure your data is stored in Snowflake, and you have the necessary credentials and permissions to access it.
4. Launch Power BI Desktop.
5. Click on the "Home" tab in Power BI Desktop.
6. Click on "Get Data" to open the "Get Data" dialog box.

7. In the "Get Data" dialog box, search for "Snowflake" and select it as your data source.

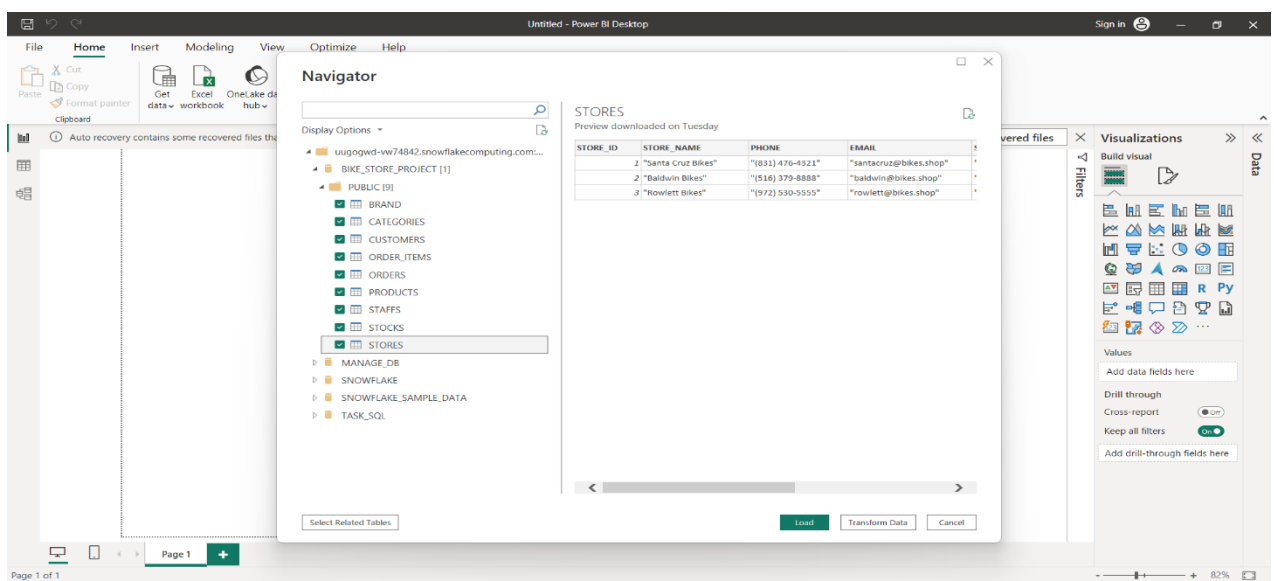


8. In the "Snowflake Database" dialog, enter your Snowflake server, database, warehouse, and authentication credentials (username and password). Click "OK" to establish the connection.





9. Power BI will display a Navigator window with a list of tables and views available in your Snowflake database. Select the tables/views you want to load into Power BI, then click the "Load" button.



➤ **CREATE A DASHBOARD AND REPORT ACCORDING TO REQUIREMENT WITH POWER BI:**

