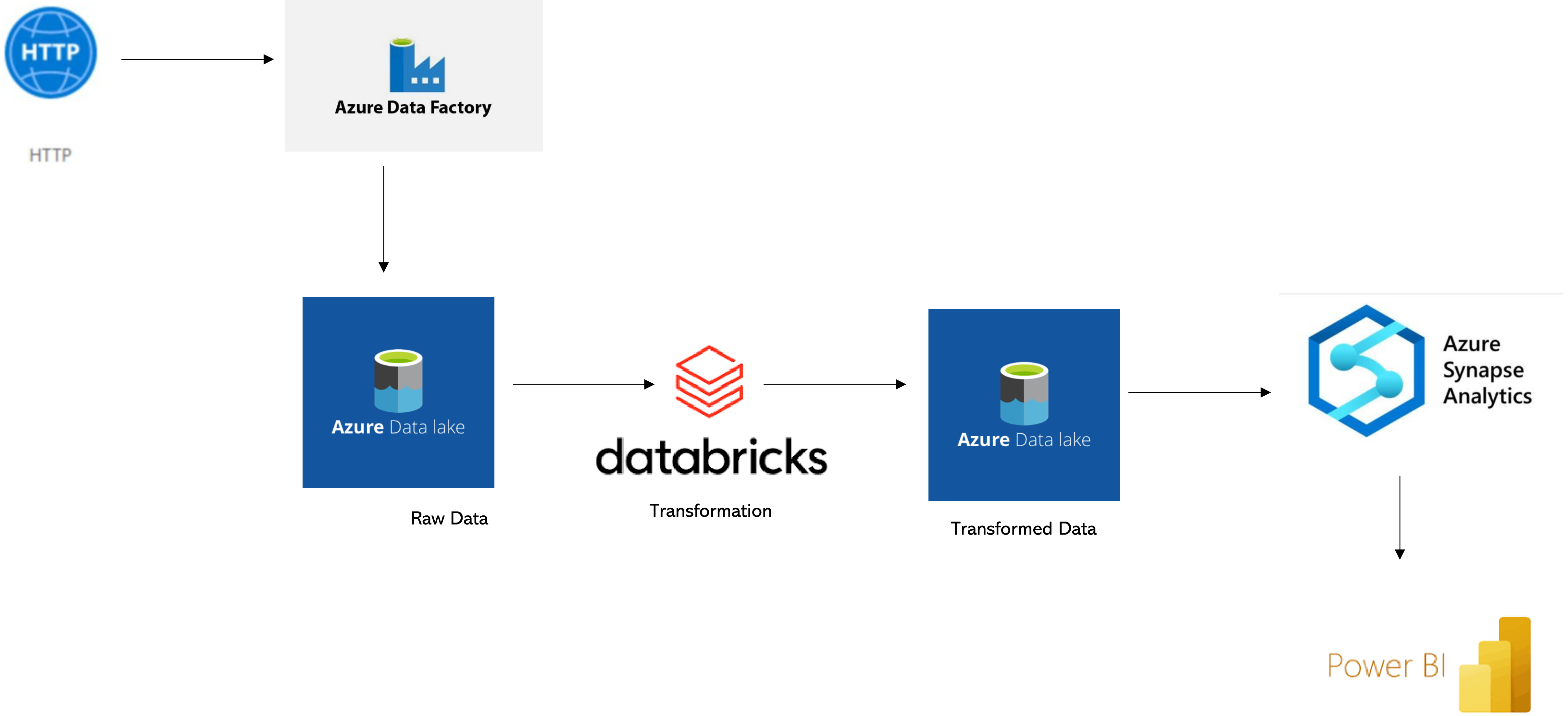
Azure Data Engineering Project

~ Siddhant Prakash More

ARCHITECTURE DIAGRAM

Data Source:
GitHub API



HTTP

Azure Data Factory

Azure Data lake

Raw Data

databricks

Transformation

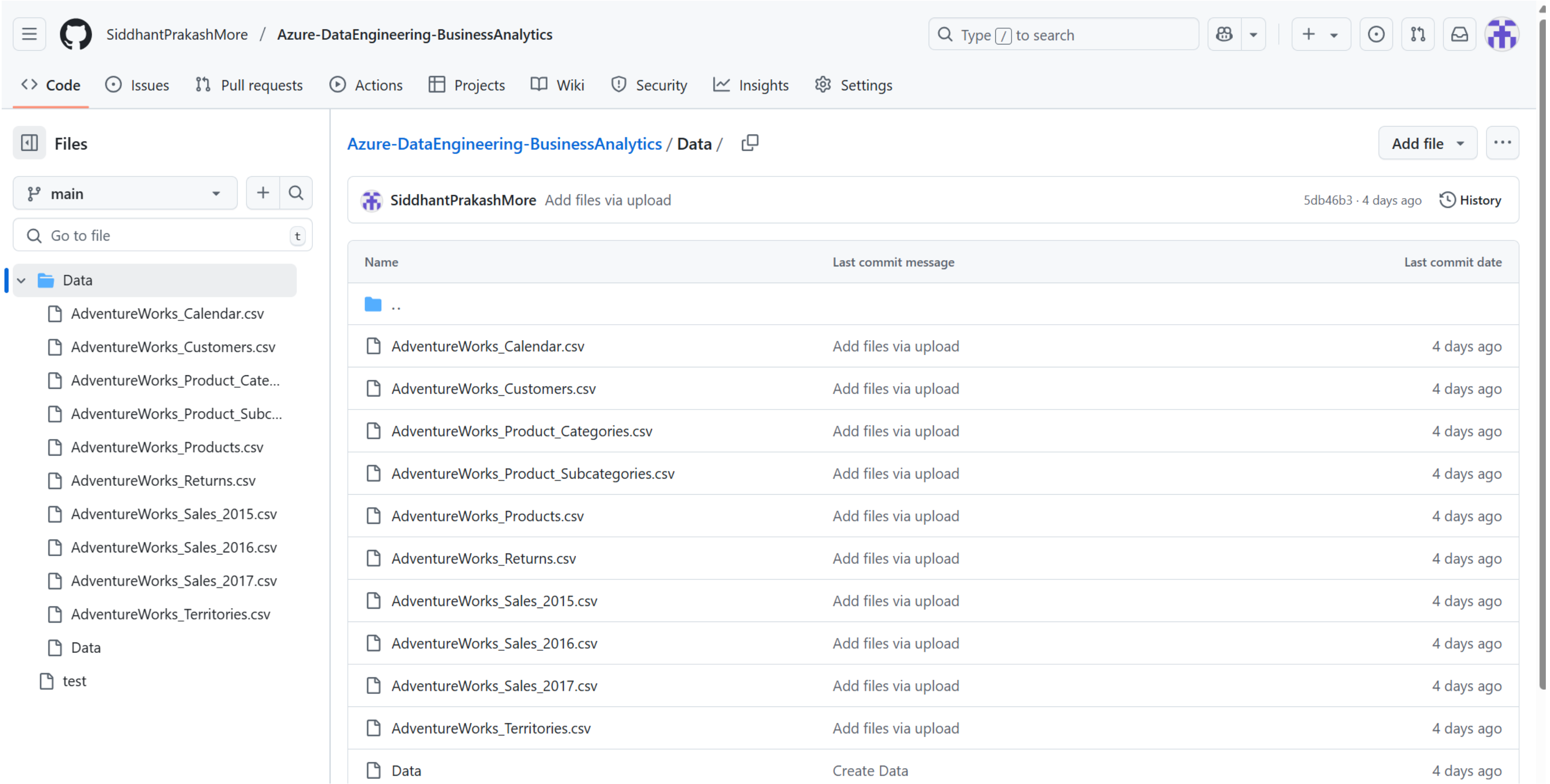
Azure Data lake

Transformed Data

Azure Synapse Analytics

Power BI

DATA SOURCE


SiddhantPrakashMore / Azure-DataEngineering-BusinessAnalytics
Type / to search
Code
Issues
Pull requests
Actions
Projects
Wiki
Security
Insights
Settings
Files
main
Go to file
t
Data
AdventureWorks_Calendar.csv
AdventureWorks_Customers.csv
AdventureWorks_Product_Cate...
AdventureWorks_Product_Subc...
AdventureWorks_Products.csv
AdventureWorks_Returns.csv
AdventureWorks_Sales_2015.csv
AdventureWorks_Sales_2016.csv
AdventureWorks_Sales_2017.csv
AdventureWorks_Territories.csv
Data
test
Add file
History

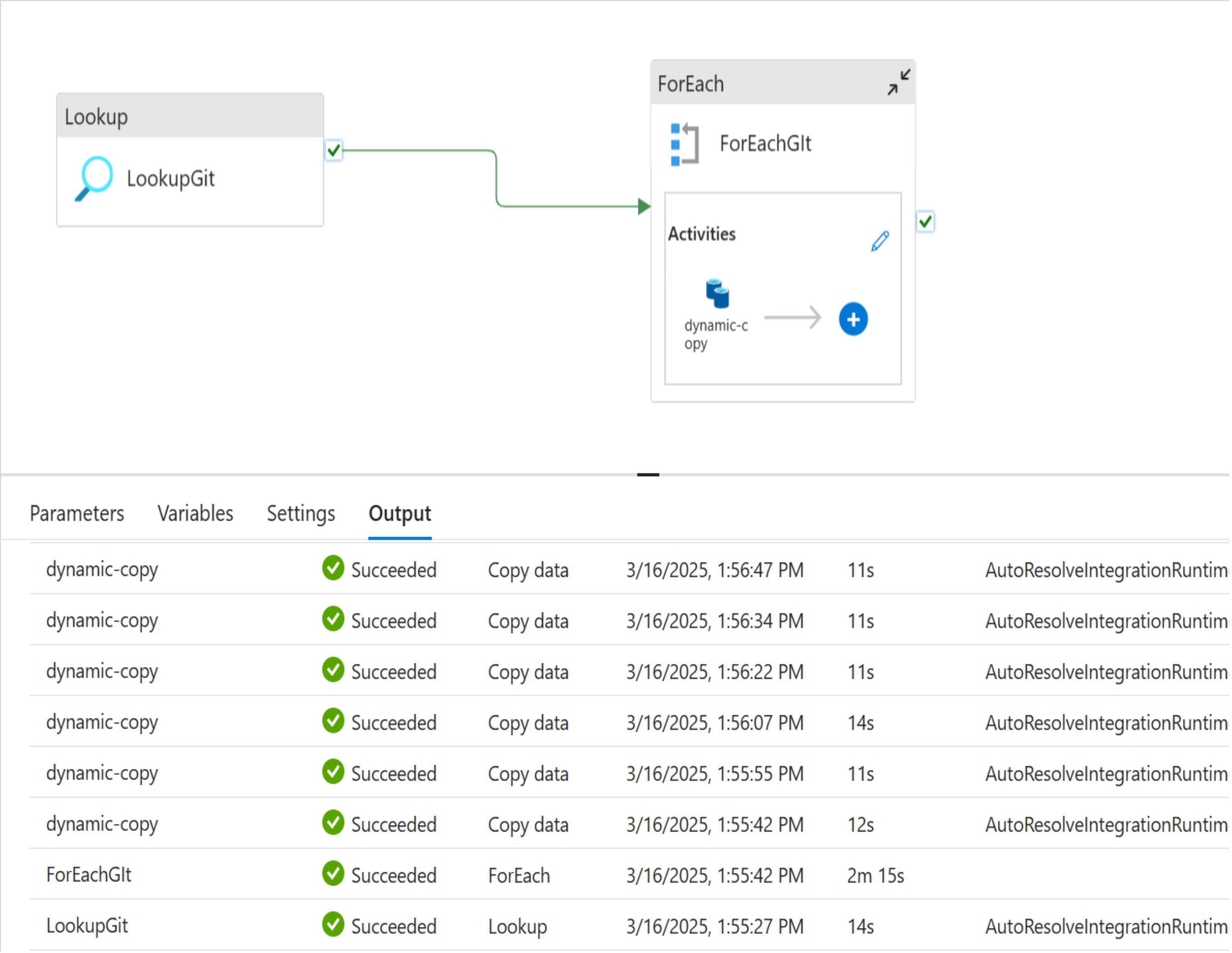
Azure-DataEngineering-BusinessAnalytics / Data /

SiddhantPrakashMore Add files via upload 5db46b3 · 4 days ago

Name	Last commit message	Last commit date
..		
AdventureWorks_Calendar.csv	Add files via upload	4 days ago
AdventureWorks_Customers.csv	Add files via upload	4 days ago
AdventureWorks_Product_Categories.csv	Add files via upload	4 days ago
AdventureWorks_Product_Subcategories.csv	Add files via upload	4 days ago
AdventureWorks_Products.csv	Add files via upload	4 days ago
AdventureWorks_Returns.csv	Add files via upload	4 days ago
AdventureWorks_Sales_2015.csv	Add files via upload	4 days ago
AdventureWorks_Sales_2016.csv	Add files via upload	4 days ago
AdventureWorks_Sales_2017.csv	Add files via upload	4 days ago
AdventureWorks_Territories.csv	Add files via upload	4 days ago
Data	Create Data	4 days ago

DATA INGESTION:

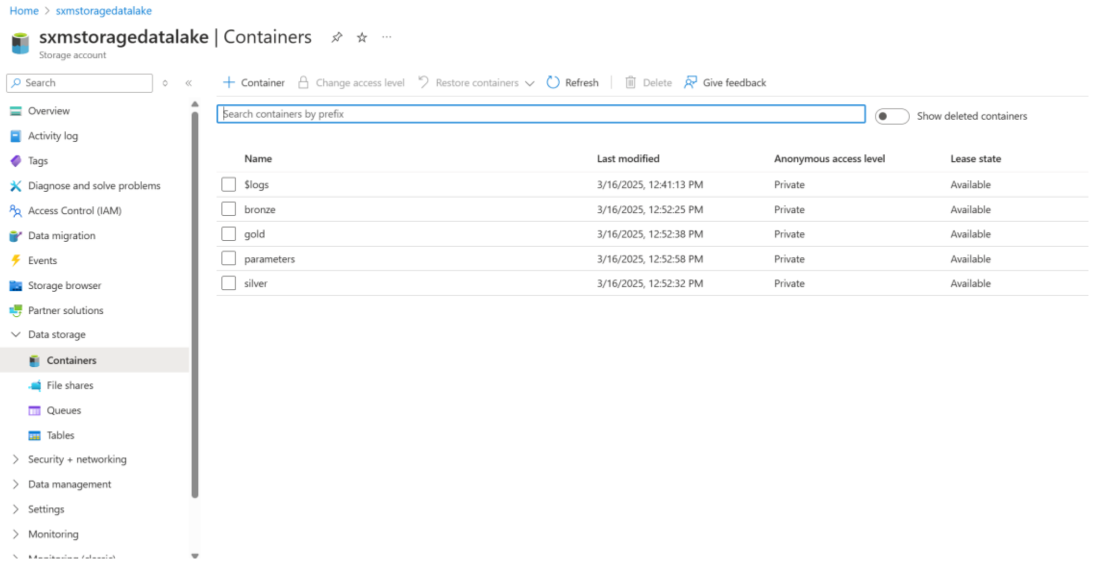
Using **Azure Data Factory (ADF)** to fetch data from the **GitHub API** via HTTP



dynamic-copy	✓ Succeeded	Copy data	3/16/2025, 1:56:47 PM	11s	AutoResolveIntegrationRuntim
dynamic-copy	✓ Succeeded	Copy data	3/16/2025, 1:56:34 PM	11s	AutoResolveIntegrationRuntim
dynamic-copy	✓ Succeeded	Copy data	3/16/2025, 1:56:22 PM	11s	AutoResolveIntegrationRuntim
dynamic-copy	✓ Succeeded	Copy data	3/16/2025, 1:56:07 PM	14s	AutoResolveIntegrationRuntim
dynamic-copy	✓ Succeeded	Copy data	3/16/2025, 1:55:55 PM	11s	AutoResolveIntegrationRuntim
dynamic-copy	✓ Succeeded	Copy data	3/16/2025, 1:55:42 PM	12s	AutoResolveIntegrationRuntim
ForEachGIt	✓ Succeeded	ForEach	3/16/2025, 1:55:42 PM	2m 15s	
LookupGit	✓ Succeeded	Lookup	3/16/2025, 1:55:27 PM	14s	AutoResolveIntegrationRuntim

STORAGE: AZURE DATA LAKE

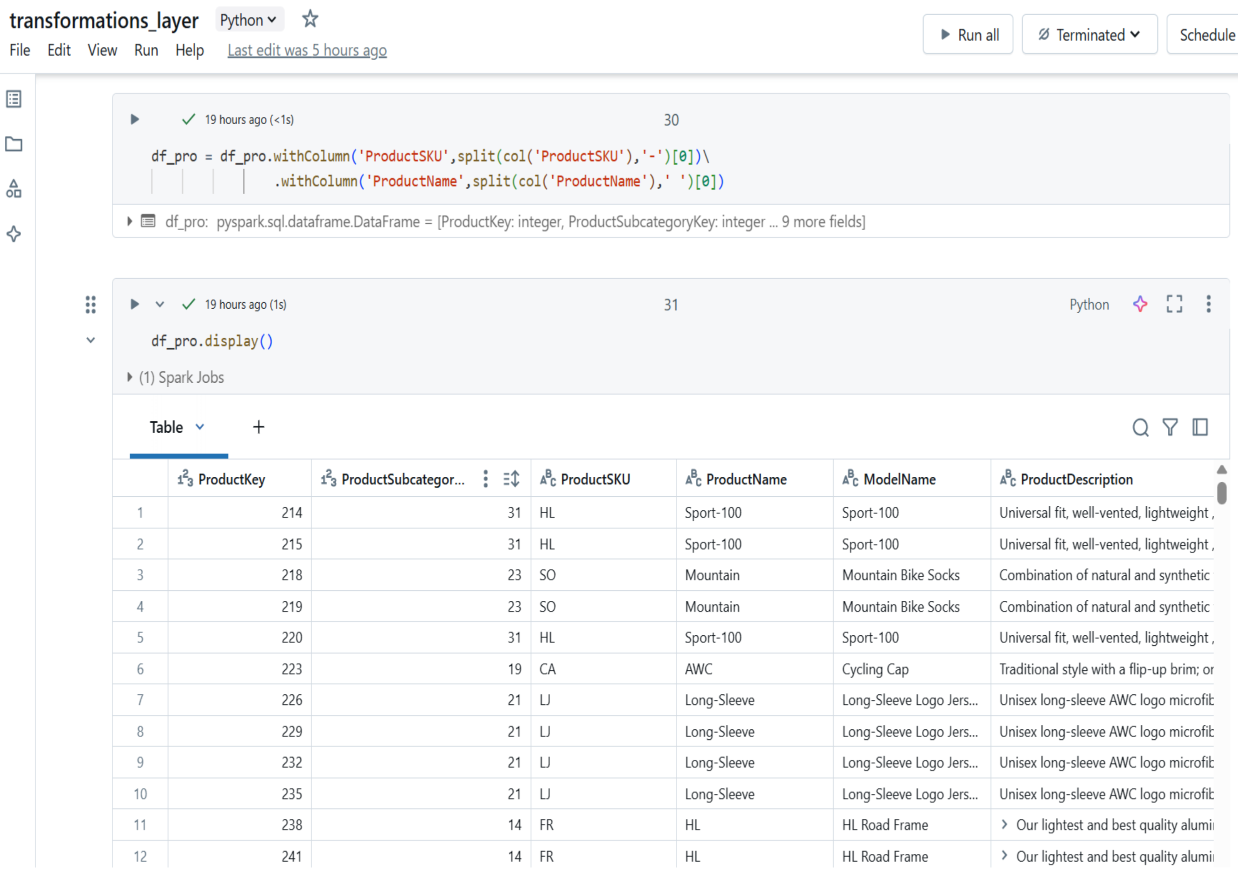
- The storage architecture follows a **medallion design** using **Azure Data Lake**, ensuring efficient data management across three layers. The **bronze layer** stores raw, unprocessed data directly ingested from the source. The **silver layer** contains cleaned and transformed data, making it structured and ready for analysis. The **gold layer** holds the final, optimized dataset, designed for efficient querying and reporting in **Azure Synapse Analytics** and **Power BI**.



	Name	Last modified	Anonymous access level	Lease state
□	$logs	3/16/2025, 12:41:13 PM	Private	Available
□	bronze	3/16/2025, 12:52:25 PM	Private	Available
□	gold	3/16/2025, 12:52:38 PM	Private	Available
□	parameters	3/16/2025, 12:52:58 PM	Private	Available
□	silver	3/16/2025, 12:52:32 PM	Private	Available

Data Transformation: Databricks

- Used **Databricks with PySpark** for data transformation, processing raw data from the **bronze layer** and refining it into the **silver layer**. The transformation included data cleaning, structuring, and enrichment to ensure high-quality, usable data. This process optimized the data for further analysis and integration with downstream systems.



```
df_pro = df_pro.withColumn('ProductSKU',split(col('ProductSKU'),'-')[0])\
                    .withColumn('ProductName',split(col('ProductName'),' ')[0])
```

```
df_pro:  pyspark.sql.dataframe.DataFrame = [ProductKey: integer, ProductSubcategoryKey: integer ... 9 more fields]
```

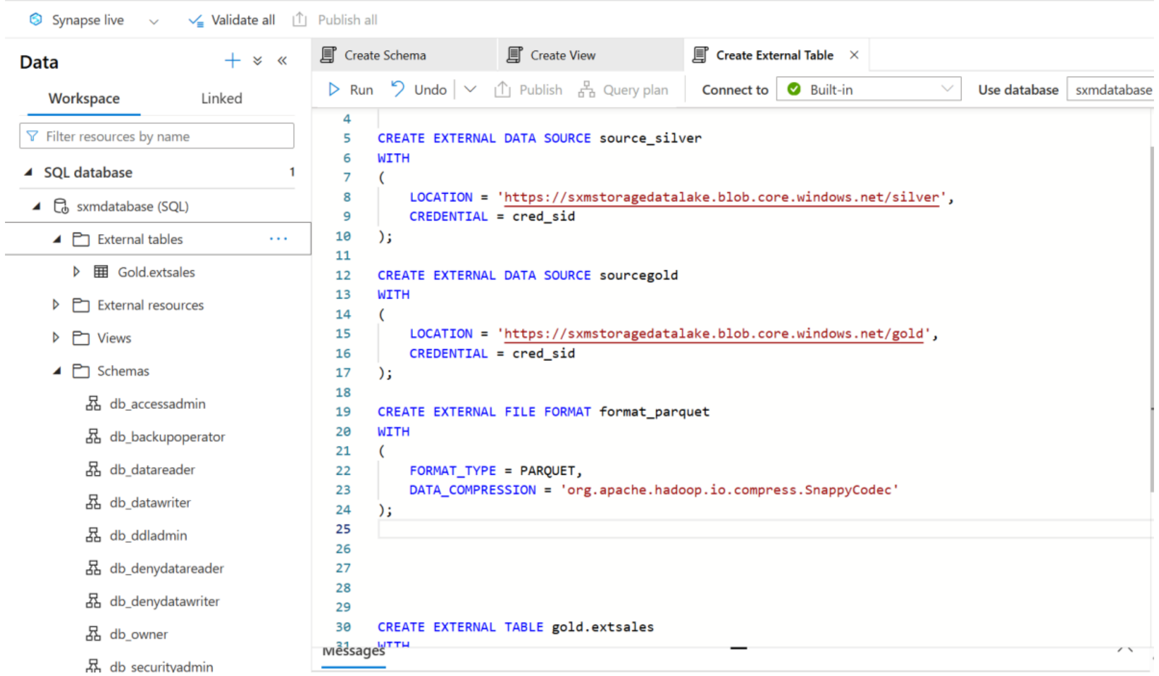
```
df_pro.display()
```

	ProductKey	ProductSubcategor...	ProductSKU	ProductName	ModelName	ProductDescription
1	214	31	HL	Sport-100	Sport-100	Universal fit, well-vented, lightweight ,
2	215	31	HL	Sport-100	Sport-100	Universal fit, well-vented, lightweight ,
3	218	23	SO	Mountain	Mountain Bike Socks	Combination of natural and synthetic
4	219	23	SO	Mountain	Mountain Bike Socks	Combination of natural and synthetic
5	220	31	HL	Sport-100	Sport-100	Universal fit, well-vented, lightweight ,
6	223	19	CA	AWC	Cycling Cap	Traditional style with a flip-up brim; or
7	226	21	LJ	Long-Sleeve	Long-Sleeve Logo Jers...	Unisex long-sleeve AWC logo microfib
8	229	21	LJ	Long-Sleeve	Long-Sleeve Logo Jers...	Unisex long-sleeve AWC logo microfib
9	232	21	LJ	Long-Sleeve	Long-Sleeve Logo Jers...	Unisex long-sleeve AWC logo microfib
10	235	21	LJ	Long-Sleeve	Long-Sleeve Logo Jers...	Unisex long-sleeve AWC logo microfib
11	238	14	FR	HL	HL Road Frame	› Our lightest and best quality alumi
12	241	14	FR	HL	HL Road Frame	› Our lightest and best quality alumi

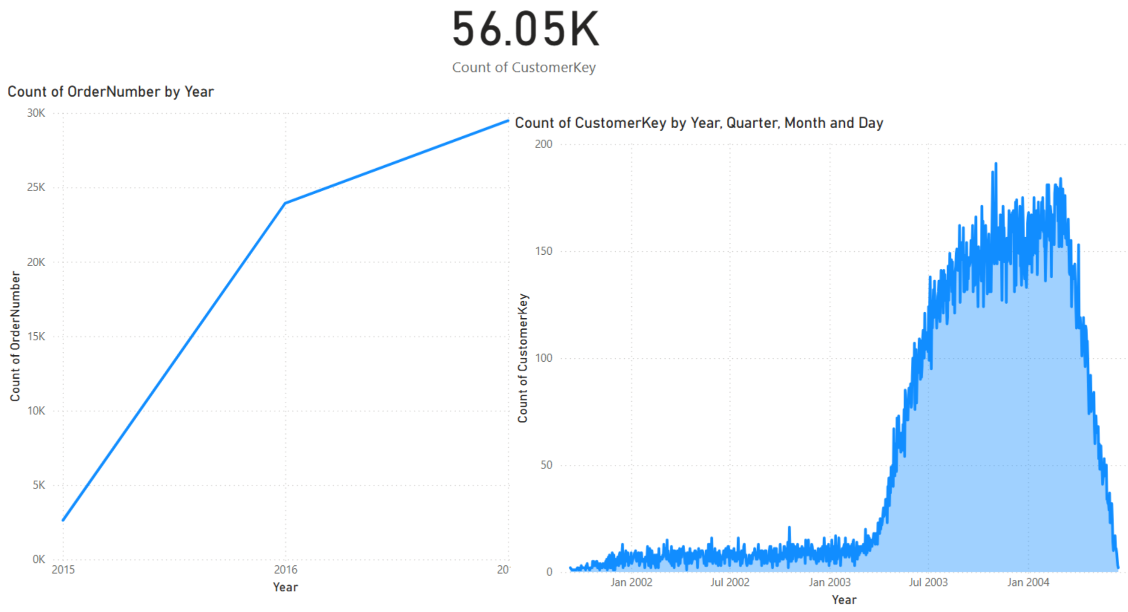
Analytics & BI:

- **Azure Synapse Analytics** – Used for running analytical queries on transformed data and integrating with BI tools.

Power BI – Utilized for creating dashboards and visual reports based on the processed data.



```

CREATE EXTERNAL DATA SOURCE source_silver
WITH
(
    LOCATION = 'https://sxmstoragedatalake.blob.core.windows.net/silver',
    CREDENTIAL = cred_sid
);

CREATE EXTERNAL DATA SOURCE sourcegold
WITH
(
    LOCATION = 'https://sxmstoragedatalake.blob.core.windows.net/gold',
    CREDENTIAL = cred_sid
);

CREATE EXTERNAL FILE FORMAT format_parquet
WITH
(
    FORMAT_TYPE = PARQUET,
    DATA_COMPRESSION = 'org.apache.hadoop.io.compress.SnappyCodec'
);





CREATE EXTERNAL TABLE gold.extsales
WITH
```


56.05K
Count of CustomerKey
Count of OrderNumber by Year
30K
25K
20K
15K
10K
5K
0K
Count of OrderNumber
Year
Count of CustomerKey by Year, Quarter, Month and Day
Count of CustomerKey
Jan 2002
Jul 2002
Jan 2003
Jul 2003
Jan 2004
Year