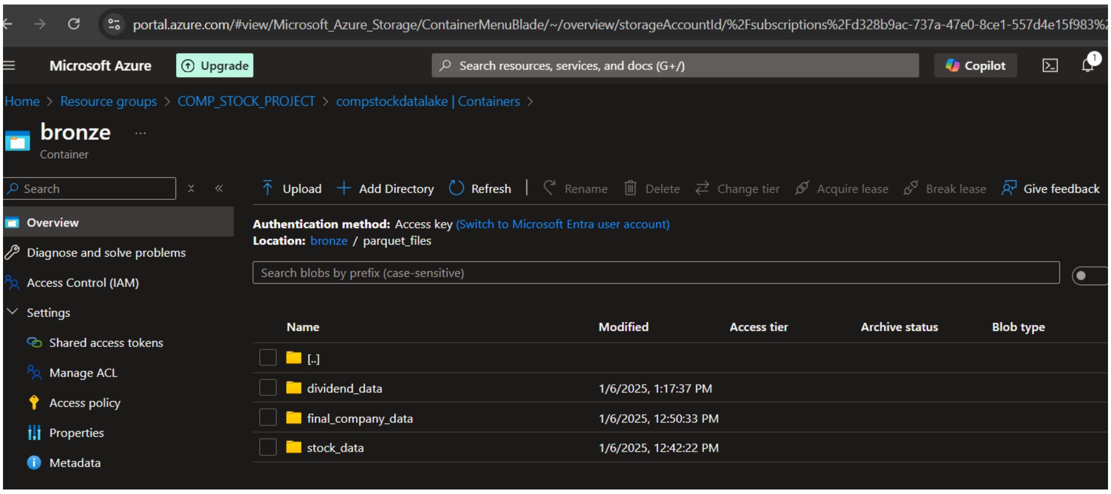
Financial Data Engineering in Azure- Guide

Data extract and transformation

Initially, I intended to use an Azure Function to deploy my data fetching code, which retrieves data from yfinance. However, yfinance implemented a limitation on company data requests, so I had to introduce a time delay in the code, waiting one second between each company data retrieval. As a result, I exceeded the 5-minute limit allowed in the Azure Function package. Therefore, I switched to using a Databricks notebook to run the Python script instead. I created a service principal through app registration, which I assigned to my Data Lake to load into bronze container



portal.azure.com/#view/Microsoft_Azure_Storage/ContainerMenuBlade/~/overview/storageAccountId/%2Fsubscriptions%2Fd328b9ac-737a-47e0-8ce1-557d4e15f983%2

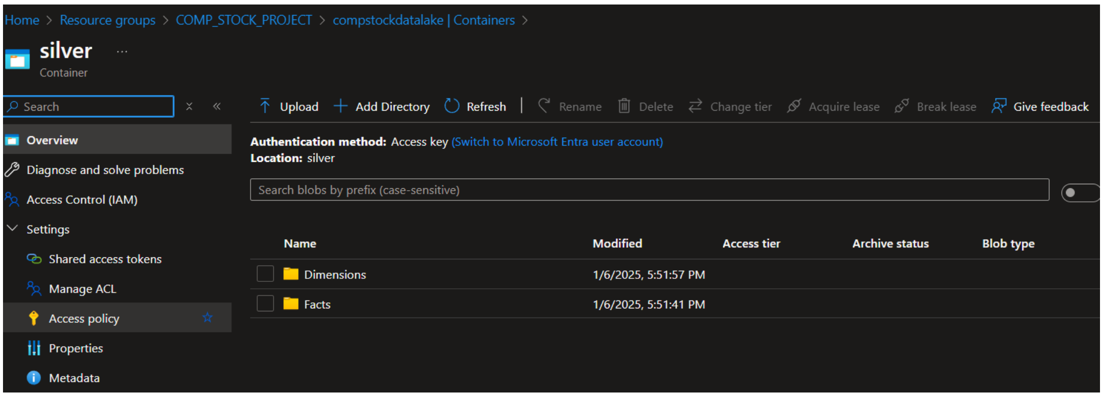
Home > Resource groups > COMP_STOCK_PROJECT > compstockdatalake | Containers >

bronze
Container

Authentication method: Access key (Switch to Microsoft Entra user account)
Location: bronze / parquet_files

	Name	Modified	Access tier	Archive status	Blob type
□	[..]				
□	dividend_data	1/6/2025, 1:17:37 PM			
□	final_company_data	1/6/2025, 12:50:33 PM			
□	stock_data	1/6/2025, 12:42:22 PM			

Then, I performed transformations using a Databricks notebook, and loaded the results into the Silver container.



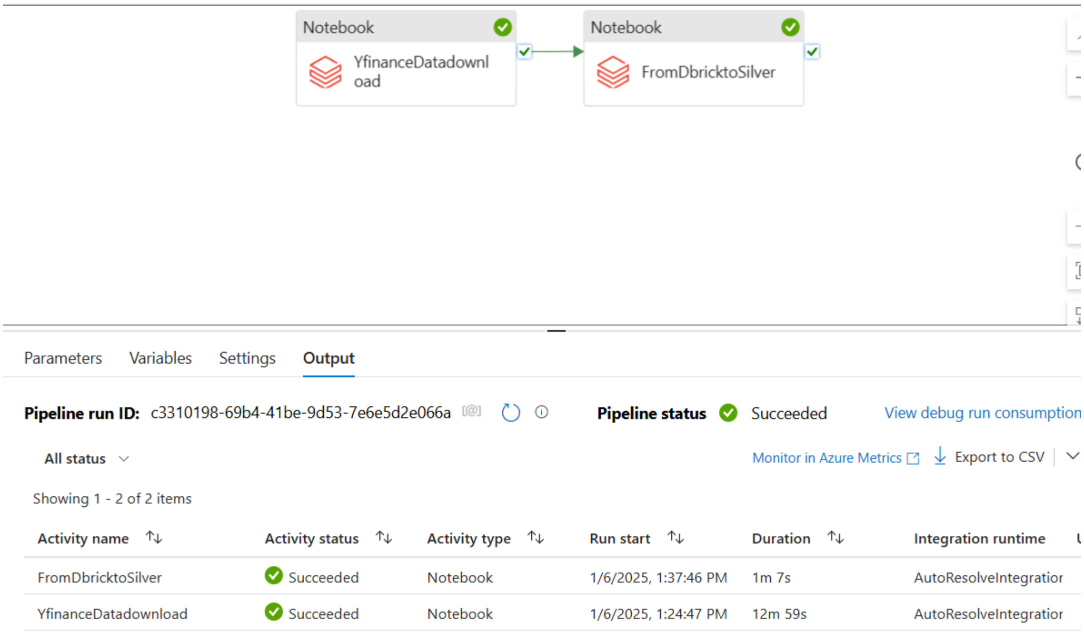
Home > Resource groups > COMP_STOCK_PROJECT > compstockdatalake | Containers >

silver
Container

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

	Name	Modified	Access tier	Archive status	Blob type
□	Dimensions	1/6/2025, 5:51:57 PM			
□	Facts	1/6/2025, 5:51:41 PM			

Pipeline



Pipeline run ID: c3310198-69b4-41be-9d53-7e6e5d2e066a **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 status	Activity type	Run start	Duration	Integration runtime
FromDbricktoSilver	✓ Succeeded	Notebook	1/6/2025, 1:37:46 PM	1m 7s	AutoResolveIntegratior
YfinanceDatadownload	✓ Succeeded	Notebook	1/6/2025, 1:24:47 PM	12m 59s	AutoResolveIntegratior

Azure Synapse

Synapse Managed Identity: Ensure the **Storage Blob Data Contributor** role is assigned to the Synapse workspace's managed identity in the storage account.

User Identity: Assign the **Storage Blob Data Contributor** role to your user account in the storage account.

Why I Chose Synapse Serverless SQL

I opted for the serverless SQL solution in Azure Synapse because it is significantly more cost-effective than storing data directly in Synapse's dedicated SQL pools. This approach allows me to query data stored in Azure Data Lake using an abstraction layer. The serverless SQL layer acts as a logical interface, enabling on-demand querying of files (e.g., Parquet, CSV) without the need for pre-loading or duplicating the data. This flexibility reduces storage and compute costs while maintaining scalability for ad-hoc analytics.

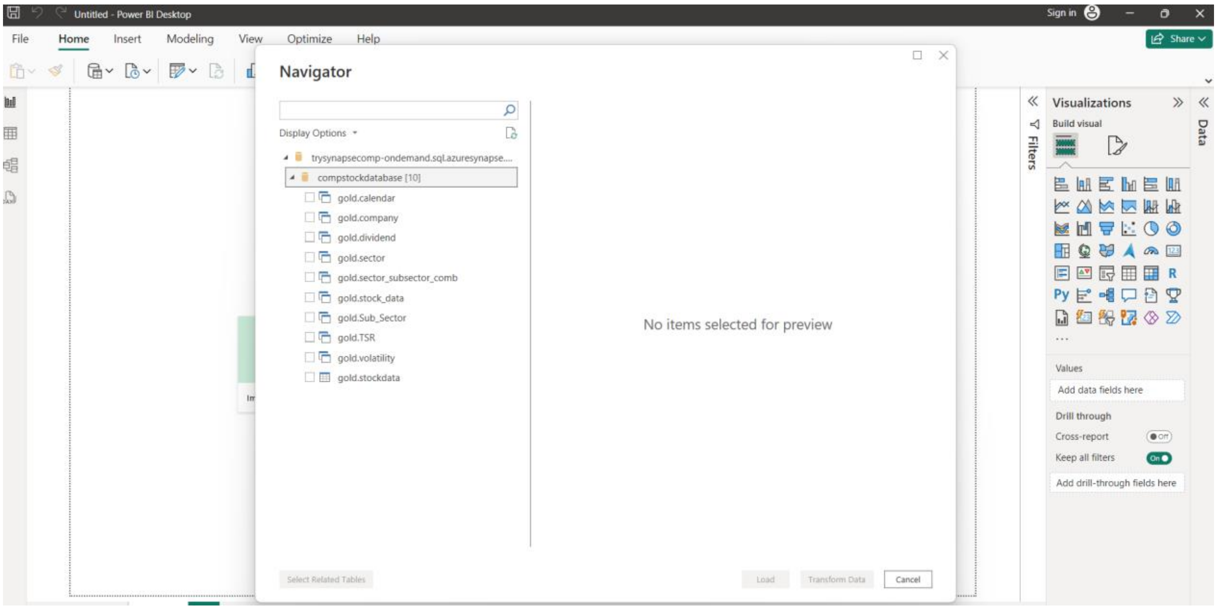
In **Synapse**, you created **views** and **external tables** to query the data directly from the Data Lake.

External tables allow you to query data stored in your Data Lake without having to move it into a database, providing efficient and scalable querying.

Views were created on top of these external tables to simplify access and provide a structured way to query the transformed data.

Loading

1. After creating the schema in Synapse, I proceeded with loading the the schema into Power BI desktop using SQL serverless endpoints.



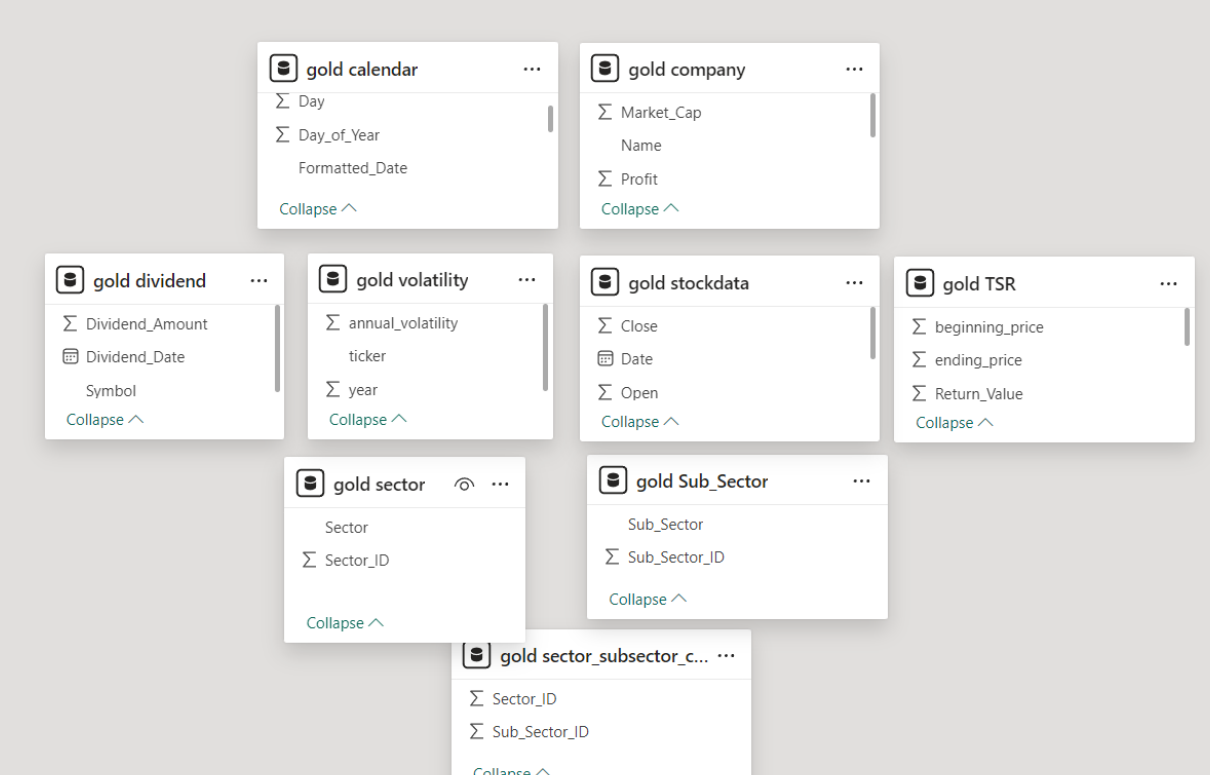
Navigator

Display Options

- trysynapsecomp-ondemand.sql.azuresynapse....
  - compstockdatabase [10]
    - □ gold.calendar
    - □ gold.company
    - □ gold.dividend
    - □ gold.sector
    - □ gold.sector_subsector_comb
    - □ gold.stock_data
    - □ gold.Sub_Sector
    - □ gold.TSR
    - □ gold.volatility
    - □ gold.stockdata

No items selected for preview

Here is the schema:



gold calendar
- Day
- Day_of_Year
- Formatted_Date

gold company
- Market_Cap
- Name
- Profit

gold dividend
- Dividend_Amount
- Dividend_Date
- Symbol

gold volatility
- annual_volatility
- ticker
- year

gold stockdata
- Close
- Date
- Open

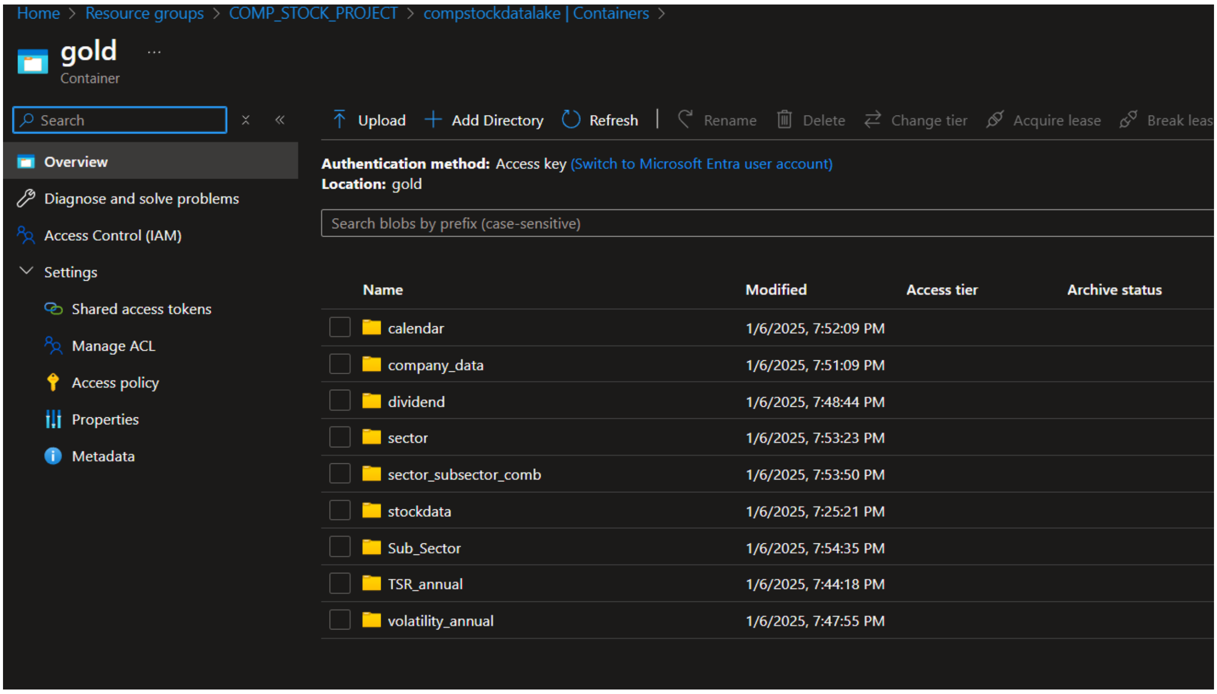
gold TSR
- beginning_price
- ending_price
- Return_Value

gold sector
- Sector
- Sector_ID

gold Sub_Sector
- Sub_Sector
- Sub_Sector_ID

gold sector_subsector_c...
- Sector_ID
- Sub_Sector_ID

2. I also got it loaded into the gold container

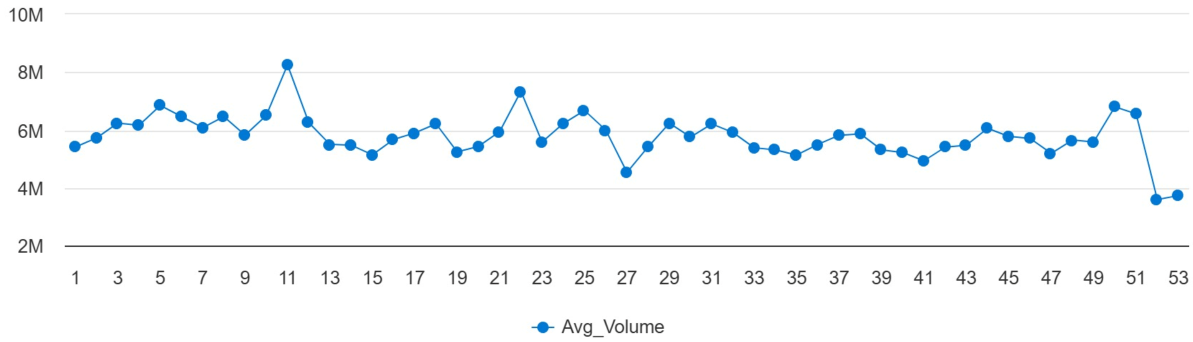


Authentication method: Access key (Switch to Microsoft Entra user account)
Location: gold

Name	Modified	Access tier	Archive status
□ calendar	1/6/2025, 7:52:09 PM		
□ company_data	1/6/2025, 7:51:09 PM		
□ dividend	1/6/2025, 7:48:44 PM		
□ sector	1/6/2025, 7:53:23 PM		
□ sector_subsector_comb	1/6/2025, 7:53:50 PM		
□ stockdata	1/6/2025, 7:25:21 PM		
□ Sub_Sector	1/6/2025, 7:54:35 PM		
□ TSR_annual	1/6/2025, 7:44:18 PM		
□ volatility_annual	1/6/2025, 7:47:55 PM		

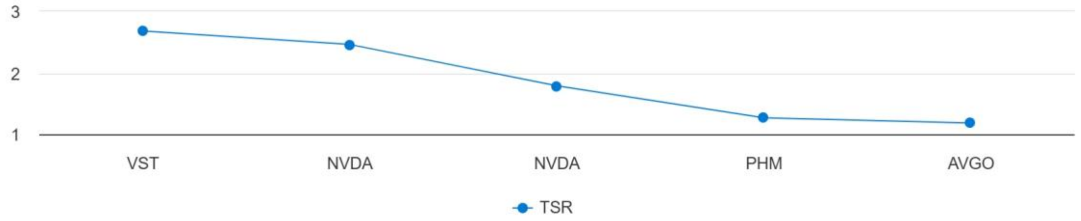
Analytics

```
SELECT
    DATEPART(WEEK, Date) AS Week_Number,
    AVG([Close]) AS Avg_Close,
    AVG(Volume) AS Avg_Volume
FROM gold.stock_data_transf
GROUP BY DATEPART(WEEK, Date)
ORDER BY Week_Number;
```

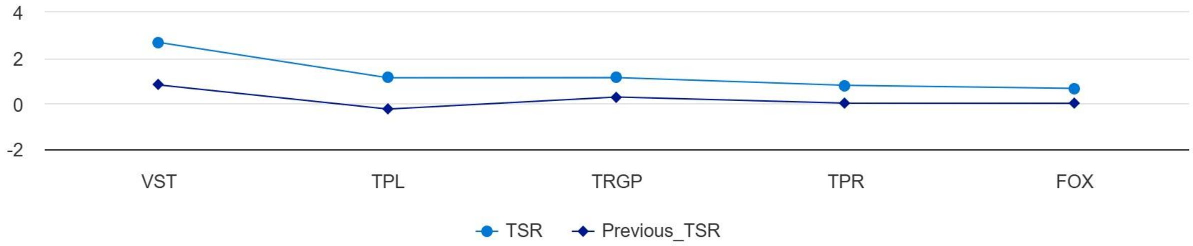


```
--- Top 5 companies with the largest TSR
SELECT TOP 5
    ticker,
    year,
    TSR
```

```sql
FROM gold.TSR_annual
ORDER BY TSR DESC;
```



```sql
---TOP 5 companies with the best yearly TSR change --------
WITH TSR_Changes AS (
    SELECT
        ticker,
        year,
        TSR,
        LAG(TSR) OVER (PARTITION BY ticker ORDER BY year) AS Previous_TSR
    FROM gold.TSR_annual
)
SELECT TOP 5
    ticker,
    year,
    TSR,
    Previous_TSR,
    (TSR - Previous_TSR) AS TSR_Change
FROM TSR_Changes
WHERE Previous_TSR IS NOT NULL
ORDER BY TSR_Change DESC;
```


VST
TPL
TRGP
TPR
FOX
TSR
Previous_TSR