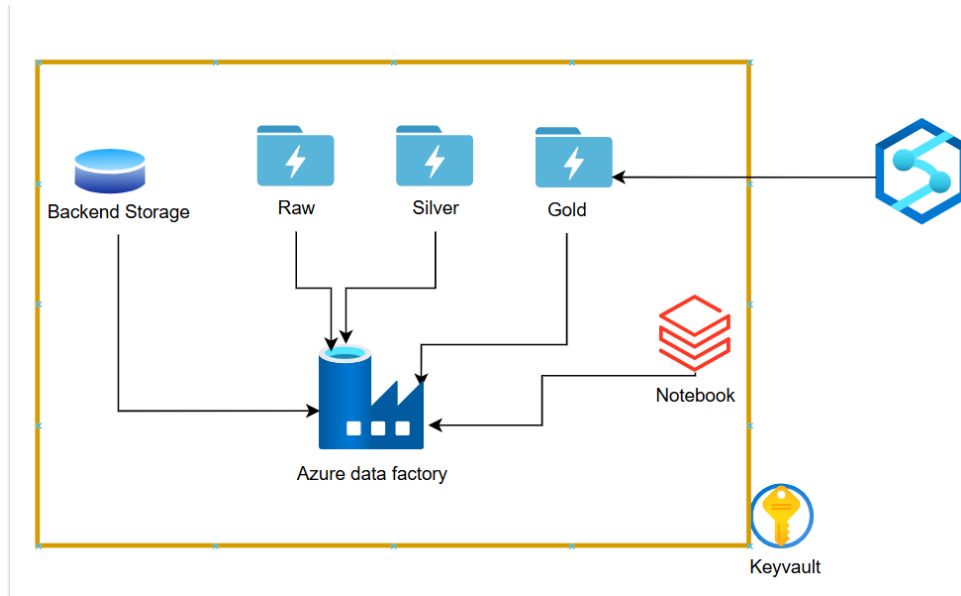


Title: Data Pipeline Documentation

1. Introduction

Overview: The data pipeline is designed to streamline data storage, transformation, and accessibility for accounts, customers, loans, loan payments, and transactions. It automates the daily ingestion, transformation, and storage of data, ensuring that only new files are processed and appended to the existing datasets for analysis and reporting.



2. Data Ingestion

Source Systems:

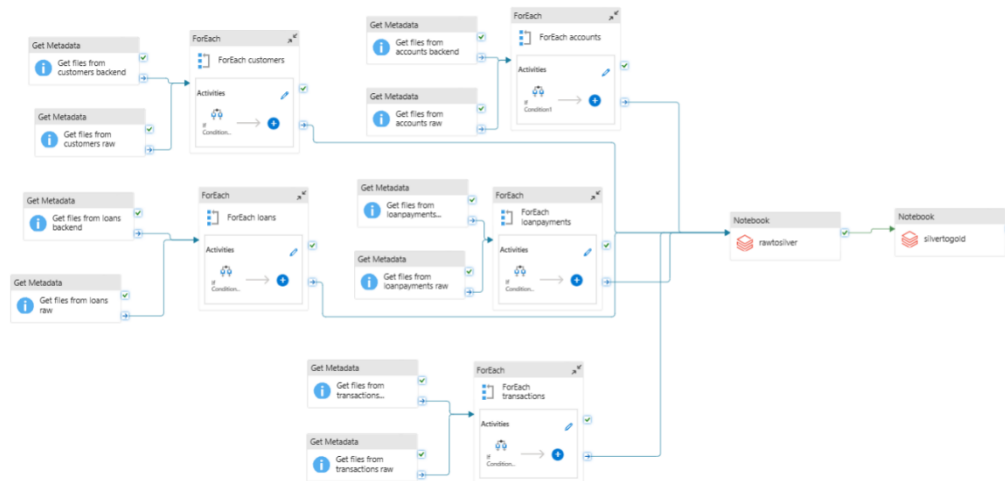
- **Backend Storage:** Contains multiple CSV files for accounts, customers, loans, loan payments, and transactions in separate folders.

Ingestion Method:

- **Pipeline Activities:**
 - **Get Metadata Activity:** Checks for new files in the backend storage.
 - **Foreach Activity:** Iterates over the list of new files.
 - **If Condition Activity:** Moves only new files to raw storage.

Tools Used:

- Azure Data Factory



3. Data Storage

Raw Storage:

- Stores newly ingested CSV files in respective folders daily.

Curated (Silver):

- Merges new files from raw storage.
- Removes duplicates, checks schema, enforces data types.
- Appends data to existing Delta tables.

Refined (Gold):

- Joins new data from accounts and customers Delta tables.
- Appends merged data to the Total balance Delta table.

4. Data Transformation

Transformation Process:

1. **Merge New Files:** Combine new files in the raw container.
2. **Remove Duplicates:** Ensure no duplicate records are present.
3. **Check and Enforce Schema:** Validate and apply the required schema.
4. **Append to Silver:** Add the transformed data to Delta tables in the Silver container.

Tools Used:

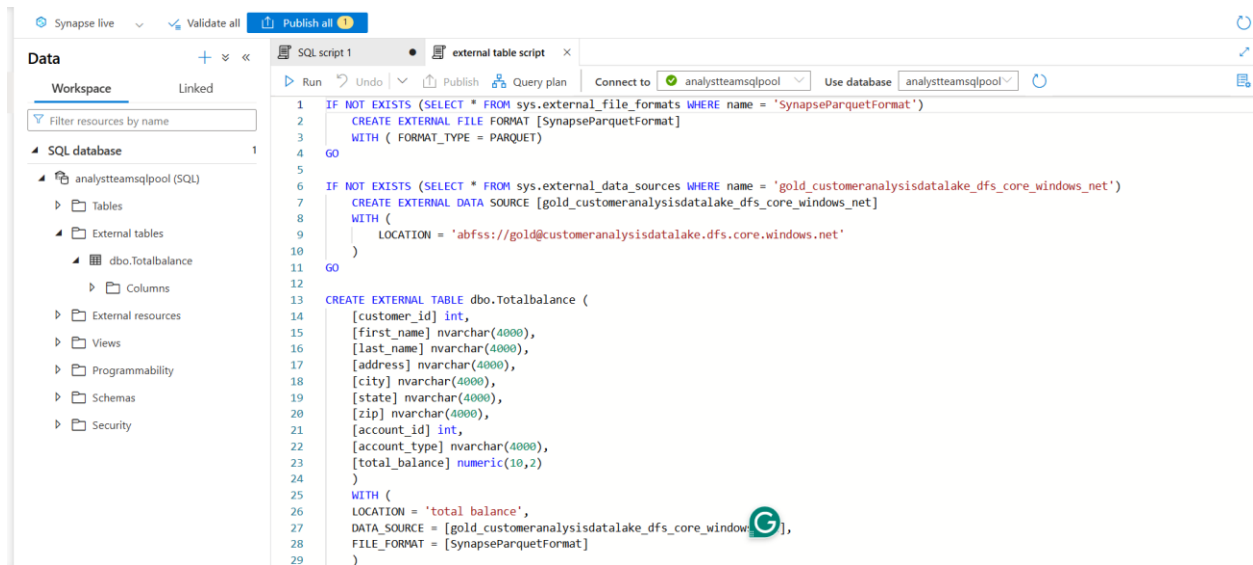
- Databricks Notebooks
- PySpark

Link : <https://github.com/RahulG005/cxanalysis/tree/main/notebooks>

5. External Tables in Synapse Analytics

Create External Tables:

- Steps to create external tables in Azure Synapse Analytics.



The screenshot displays the Azure Synapse Analytics interface. On the left, the 'Data' pane shows a tree view of resources under 'analyststeamsqlpool (SQL)', including 'Tables', 'External tables', and 'Columns'. The 'External tables' folder is expanded, showing a table named 'dbo.Totalbalance'. The main pane shows a SQL script editor with the following code:

```

1 IF NOT EXISTS (SELECT * FROM sys.external_file_formats WHERE name = 'SynapseParquetFormat')
2 CREATE EXTERNAL FILE FORMAT [SynapseParquetFormat]
3 WITH ( FORMAT_TYPE = PARQUET)
4 GO
5
6 IF NOT EXISTS (SELECT * FROM sys.external_data_sources WHERE name = 'gold_customeranalysisdatalake_dfs_core_windows_net')
7 CREATE EXTERNAL DATA SOURCE [gold_customeranalysisdatalake_dfs_core_windows_net]
8 WITH (
9     LOCATION = 'abfss://gold@customeranalysisdatalake.dfs.core.windows.net'
10 )
11 GO
12
13 CREATE EXTERNAL TABLE dbo.Totalbalance (
14     [customer_id] int,
15     [first_name] nvarchar(4000),
16     [last_name] nvarchar(4000),
17     [address] nvarchar(4000),
18     [city] nvarchar(4000),
19     [state] nvarchar(4000),
20     [zip] nvarchar(4000),
21     [account_id] int,
22     [account_type] nvarchar(4000),
23     [total_balance] numeric(10,2)
24 )
25 WITH (
26     LOCATION = 'total balance',
27     DATA_SOURCE = [gold_customeranalysisdatalake_dfs_core_windows_net],
28     FILE_FORMAT = [SynapseParquetFormat]
29 )

```