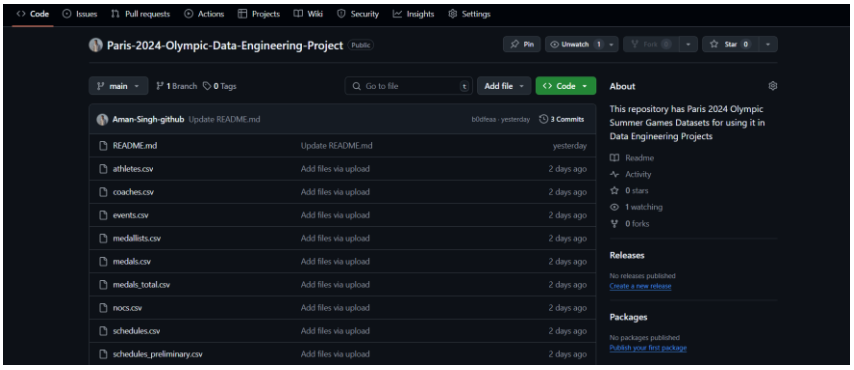
Data Engineering Project

Project: Paris 2024 Olympics - ETL Workflow for Dataset Processing with Azure

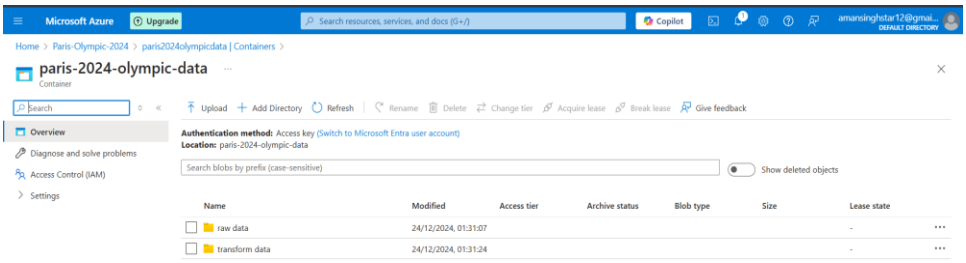
Project Architecture :

Detailed Explanation:

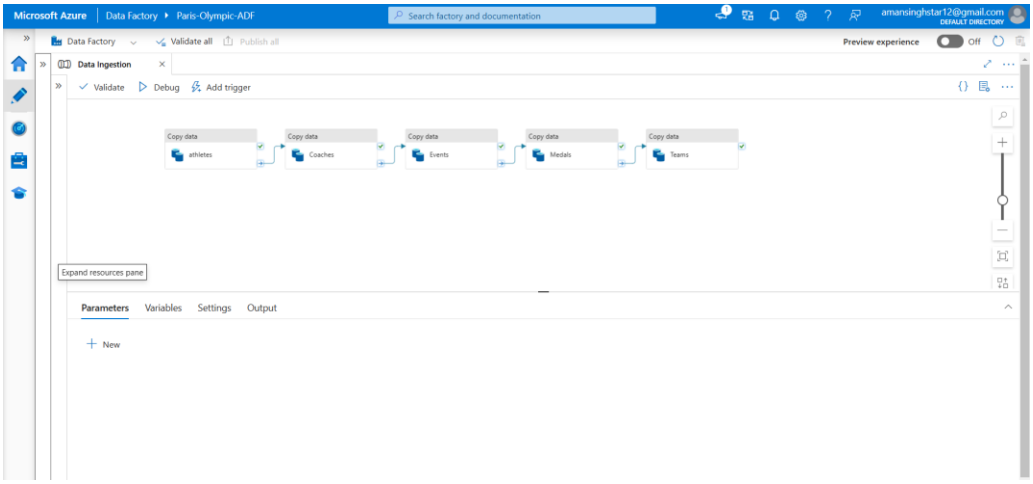
1. Downloaded the **Paris 2024 Olympic Summer Games datasets** from Kaggle.com and stored them in a GitHub repository (link) for ingestion from the web.



2. Created an **Azure Data Lake Storage Gen 2** under a designated Resource Group to facilitate data integration.

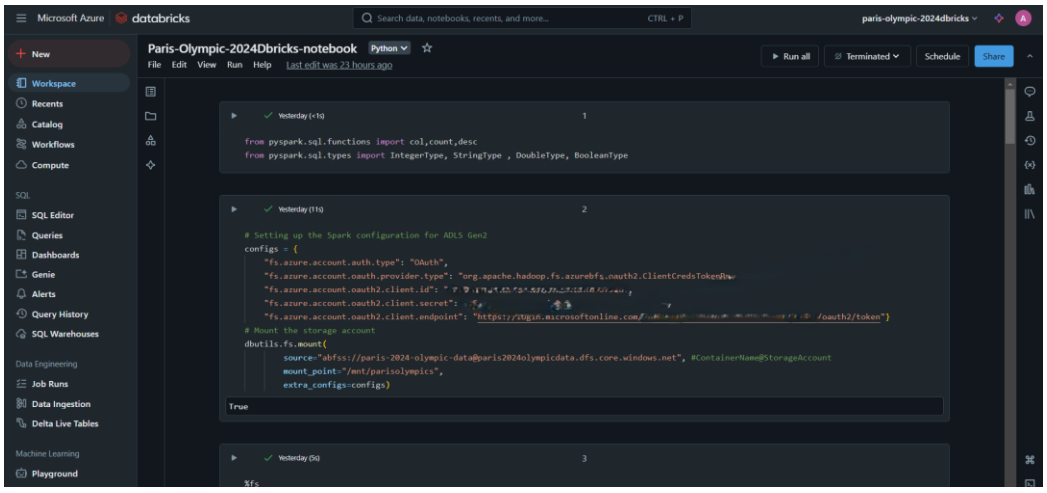


3. Within the storage account, set up a **RAW Data** directory inside a container to store unprocessed data.
4. Established another directory within the container to store **Transformed Data** after processing.
5. Provisioned an **Azure Data Factory** resource for the data ingestion process. Configured a copy activity to source data from GitHub and store it in the **RAW Data directory**.



6. Set up an **Azure Databricks** resource for data transformation. The process involved: Creating a **Spark Compute Cluster** to run Spark sessions. Developing a Python notebook for data processing, which included:

    a. Configuring Spark to connect to **Azure Data Lake Gen 2**.
    b. Mounting the storage account in Databricks.
    c. Transforming the data by adjusting column data types, handling null values, and validating headers.
    d. Loading the transformed data into the **Transformed Data directory** in the storage account.



7. Deployed an **Azure Synapse Analytics** resource for data analysis and visualization. Imported the transformed data into Synapse and created connected tables.

Performed basic analytics using SQL queries and generated visualizations for insights.

8. Connected **Power BI Desktop** to Azure Synapse Analytics using a serverless connection to build interactive dashboards for further insights and reporting.