Paris 2024 Olympics Data Engineering Project

Project Overview

This project is an end-to-end data engineering solution designed to perform detailed analysis on the Paris 2024 Olympic data. Utilizing various Azure Cloud Computing Services, the project handles the complete data pipeline from data ingestion, storage, transformation, and analysis, providing valuable insights into the Olympic events, athletes, and performances.

Technologies Used

- **Azure Data Factory**: Orchestrates the data ingestion and ETL (Extract, Transform, Load) processes.
- **Azure Data Lake Storage Gen2**: Stores raw and processed data efficiently and securely.
- **Azure Databricks**: Performs data transformations, cleansing, and advanced analytics.
- **Azure Synapse Analytics**: Provides a powerful data warehouse for querying and analyzing large datasets.

Project Architecture

1. **Data Ingestion**: Azure Data Factory is used to ingest raw data from various sources, including APIs, CSV files, and databases, into Azure Data Lake Storage Gen2.
2. **Data Storage**: Raw data is stored in Azure Data Lake Storage Gen2, organized in a hierarchical structure that supports various data formats like CSV, JSON, and Parquet.
3. **Data Transformation**: Azure Databricks is employed to clean, transform, and enrich the data. The transformed data is then stored back in Azure Data Lake Storage Gen2.
4. **Data Loading**: The processed data is loaded into Azure Synapse Analytics for further analysis and reporting.
5. **Data Analysis**: Azure Synapse Analytics is used to run complex queries and perform detailed analysis on the data, generating insights into the Paris 2024 Olympic Games.

Features

- **Scalable Data Pipeline**: Leverages Azure services to handle large volumes of data efficiently.
- **Automated ETL Process**: Azure Data Factory automates the extraction, transformation, and loading of data.
- **Advanced Analytics**: Azure Databricks enables data scientists and engineers to perform sophisticated data analysis and machine learning tasks.
- **Comprehensive Data Insights**: Azure Synapse Analytics allows for the exploration of data with SQL queries and integrates with Power BI for rich data visualization.

Prerequisites

Before you can run this project, ensure you have the following:

- An active Azure subscription.
- Azure services configured:
    - Azure Data Factory
    - Azure Data Lake Storage Gen2
    - Azure Databricks
    - Azure Synapse Analytics
- Basic knowledge of Python, SQL, and Spark (used in Azure Databricks).