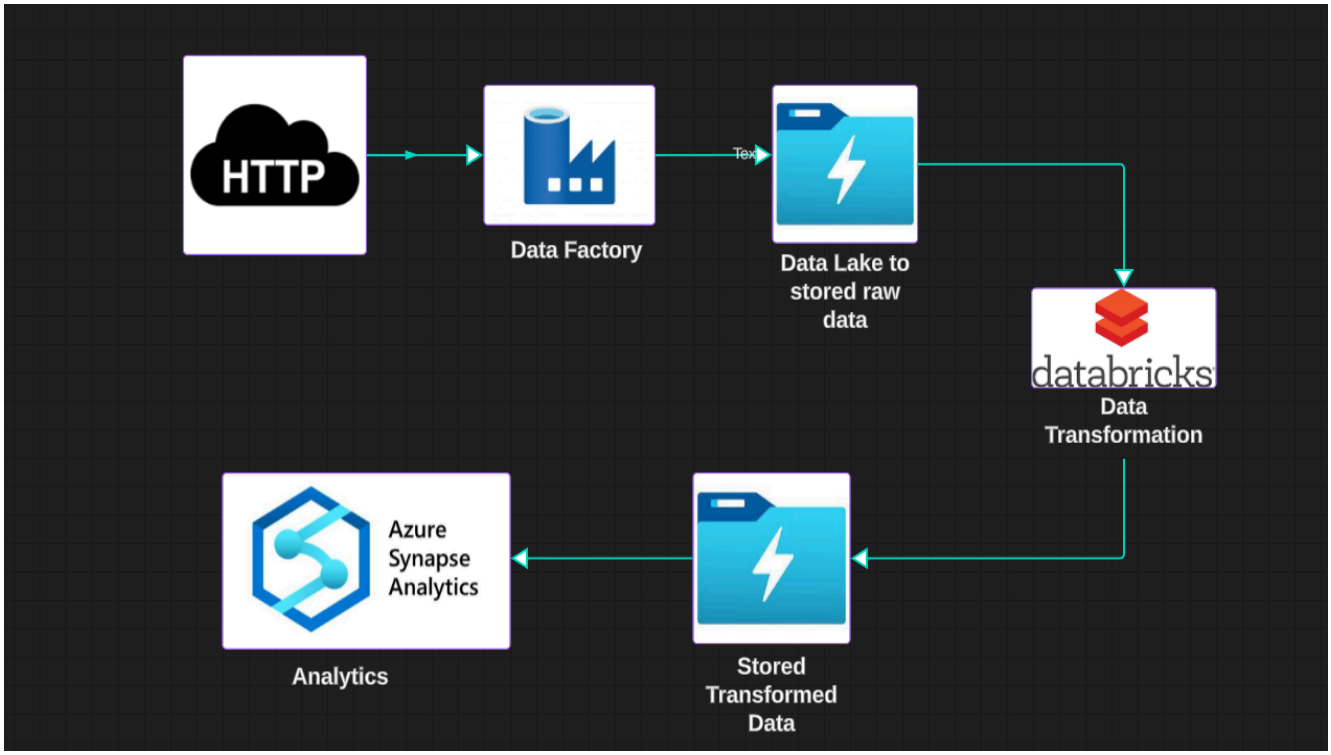
Olympic Data Analysis Project Documentation

Introduction

This documentation outlines the process and architecture for an Olympic data analysis project utilizing various tools and technologies available on the Microsoft Azure platform. The project involves building a data pipeline with the help of Data Factory to connect multiple services and easily flow the data. However, the Olympic data is extracted from HTTP sources and stored the raw data on Data Lake storage, transforming it using Azure Databricks, and stored the transformed data into Azure Data Lake Gen 2. Furthermore, it demonstrates the usage of Azure Synapse Analytics for performing analytical queries on the transformed data using SQL.

Architecture Overview

The architecture of the project involves the following key components:



Tools and Technologies Used

Azure Data Factory: Used for orchestrating and automating data movement and data transformation workflows.

Azure Data Lake: Provides a scalable and secure data lake storage solution for storing large amounts of data.

Azure Databricks: Used for data engineering and data analytics tasks, providing a collaborative environment for Apache Spark-based analytics.

Azure Synapse Analytics: Used for performing data warehousing and analytics tasks, offering insights through SQL-based queries.

Steps Involved

Data Extraction and Loading Olympic data

Extracted from HTTP sources. The extracted data is loaded into a raw data container in Azure Data Lake Gen 2.

Data Transformation

The raw data is transformed using PySpark on Azure Databricks.

Transformation includes tasks such as data cleaning, schema manipulation, and calculations.

Data Storage

The transformed data is stored into a transformed data container in Azure Data Lake Gen 2.

Each dataset (e.g., athletes, coaches, entriesgender, medals, teams) is stored as CSV files within the container.

Analytics

Analytical queries are performed on the transformed data using Azure Synapse Analytics.

SQL queries are executed to gain insights into various aspects of the Olympic data, such as medal counts, athlete demographics, and team performances.

Conclusion

This documentation provides an overview of the Olympic data analysis project architecture and the usage of various Azure tools and technologies, including Azure Data Factory, Data Lake Gen 2, Azure Databricks, and Azure Synapse Analytics. By following the outlined steps, users can extract, transform, store, and analyze data effectively on the Azure platform.