Olympic Data Analytics Using Azure

Pipeline Architecture Document

Architecture Overview

The Olympic Data Analytics pipeline is designed to process structured datasets related to athletes, coaches, medals, genders and teams. The pipeline leverages Azure-native services to implement a scalable and automated data workflow, consisting of five major layers:

Ingestion Layer:

- **Tool:** Azure Data Factory
- **Purpose:** Automatically fetch raw Olympic datasets (CSV/JSON) from local storage or public sources like Kaggle.
- **Output:** Stores raw files in the Raw Zone of Azure Data Lake Storage Gen2.

Transformation Layer:

- **Tool:** Azure Databricks (using PySpark)
- **Purpose:** Cleans, filters, and aggregates the raw data into usable structured data.
- **Output:** Transformed datasets stored in the transformed zone of Data Lake Gen2.

Storage Layer:

- **Tool:** Azure Data Lake Storage Gen2
- **Zones:**
    - **Raw Zone:** Stores original unmodified datasets.
    - **Transformed Zone:** Stores clean, analysis-ready data.
- **Purpose:** Durable and secure storage for both raw and processed data.

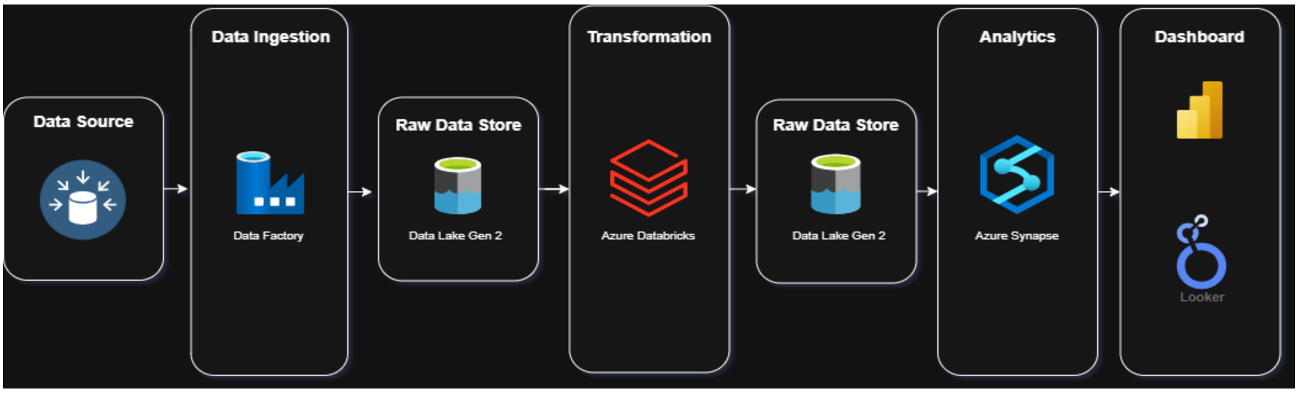
Analytics Layer:

- **Tool:** Azure Synapse Analytics
- **Purpose:** Run SQL queries and aggregate metrics on the transformed data for insights.

Visualization Layer:

- **Tool:** Power BI
- **Purpose:** Create dynamic dashboards to visualize insights.

Detailed Diagram of the Pipeline



Key Components Description

Azure Data Factory (Ingestion):

- Used for orchestrating data movement.
- Easy scheduling and automation.
- Chosen for its seamless integration with Azure services and support for batch ingestion.

Azure Databricks (Transformation):

- Used for large-scale data processing.
- PySpark allows parallel processing and complex transformation logic.
- Chosen for performance, scalability, and tight Azure integration.

Azure Data Lake Storage Gen2 (Storage):

- Stores both raw and cleaned data.
- Hierarchical namespace support.
- Chosen for cost-effective, scalable storage with fine-grained security controls.

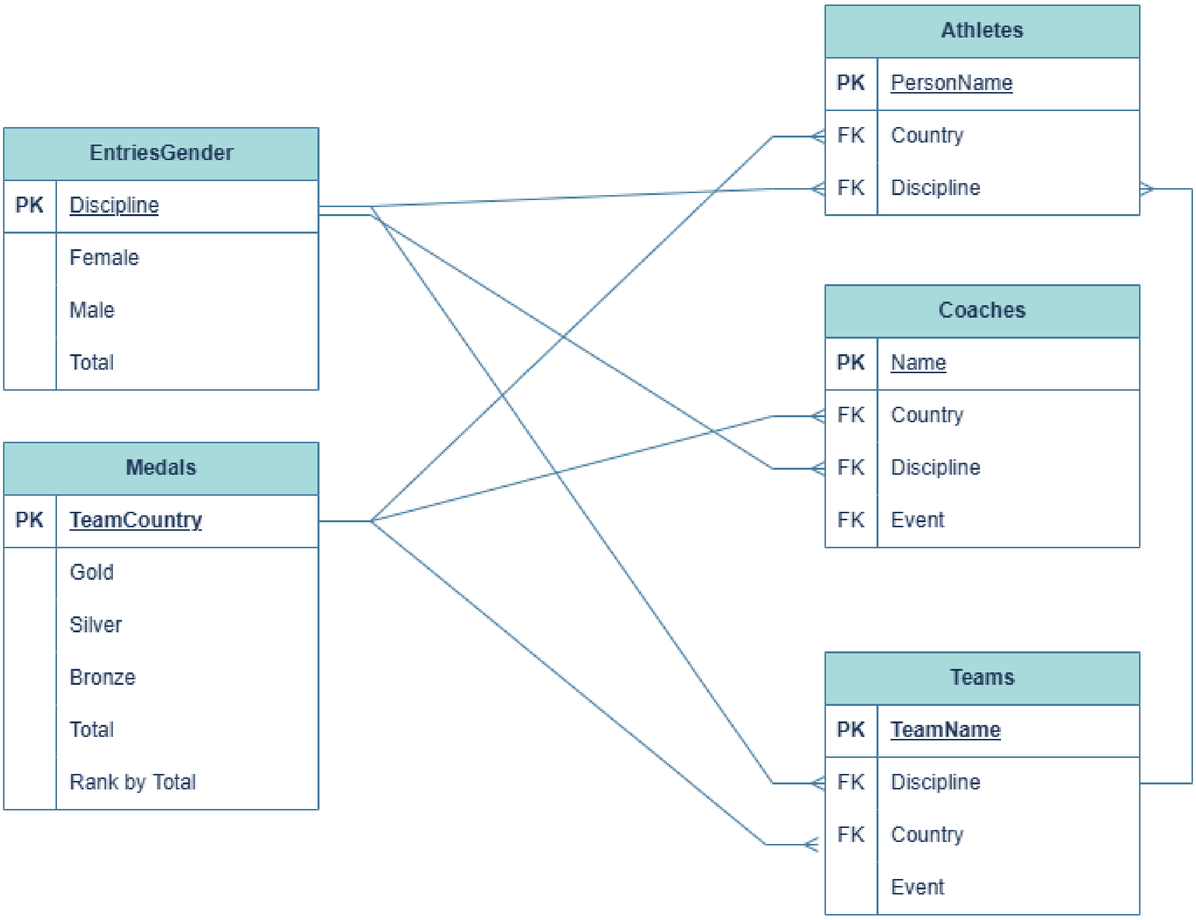
Azure Synapse Analytics (Analytics):

- Performs querying and aggregation on transformed datasets.
- Chosen for its high-performance SQL engine and easy connection to Power BI.

Power BI (Visualization):

- Used for reporting and insights sharing.
- Chosen for its rich visualizations, interactivity, and real-time refresh capabilities.

Entity Relationship Diagram (ERD)



Relationships:

- One country has many athletes
- One country has many coaches
- One country has many teams
- One discipline can have may athletes
- One discipline can have may coaches
- One discipline can have may teams
- One team/discipline can have many athletes

Tool and Technology Reflection

Reflection on Initial Tool Selection:

In Week 1, the initial plan was to use a combination of Azure Data Factory, Azure Data Lake Gen2, Azure Databricks (PySpark), and Power BI for the end-to-end Olympic data pipeline. At that time, the selections were based on their general popularity and seamless integration within the Azure ecosystem.

As the project progressed:

- I validated that ADF works well for batch ingestion.
- I explored how Azure Databricks excels at large-scale transformation using PySpark.
- Synapse Analytics was considered initially but later limited due to dataset simplicity.
- Power BI was confirmed as the primary dashboard tool for insights.

Overall, the tools remained mostly unchanged, but the design matured — such as optimizing transformation logic in Databricks and restructuring data zones in Data Lake for better organization.

Final Tool and Technology Justification:

Layer	Tool	Justification
Ingestion	Azure Data Factory	Automates importing of raw CSV data from local storage or cloud locations. Offers scalable, scheduled batch ingestion with minimal code.
Transformation	Azure Databricks (PySpark)	Ideal for distributed data processing. PySpark enables efficient filtering, joining, and aggregation of Olympic data (athletes, medals, coaches).
Storage	Azure Data Lake Gen2	Supports hierarchical namespace, allowing Raw and Transformed zones. Cost-effective, secure, and scalable.
Visualization	Power BI	Allows creation of interactive, shareable dashboards. Easy integration with Azure data sources and supports filtering, slicing, and drill-downs.

Testing and Validation

To ensure the correctness of the pipeline, I used the following methods at each stage:

1. **Data Quality Checks (Manual Validation):**
   - Checked row counts before and after transformation.
   - Verified joins between Athletes, Coaches, and Medals using unique values like Country and Discipline.
   - Ensured no nulls or unexpected values were introduced.
2. **Schema Validation:**
   - Verified the schema in each notebook cell during transformation (e.g., column names, data types).
   - Used. printSchema() and. describe() methods in PySpark to confirm expected structures.
3. **Power BI Validation:**
   - Cross-checked total medal counts with official Olympic records.
   - Validated athlete counts by country and sport.
   - Tested slicers and filters to ensure visuals updated correctly.