1■■ Data Source

Every data pipeline begins with a data source. In this project, the source data is most likely a structured dataset like a CSV file containing Olympic data (athletes, events, medal counts, etc.). This is raw data—it hasn't been cleaned or processed yet. The data source could also be a SQL database, an API, or any storage system providing structured data.

2■■ Data Ingestion – Azure Data Factory (ADF)

The first step after identifying the data source is to automate its ingestion into the Azure environment. This is handled by Azure Data Factory (ADF), which acts like a data transport service.

Role of ADF: ADF connects to the external data source, extracts the raw data, and loads it into Azure's storage service. While ADF supports some basic data transformation (like column selection or simple filtering), its main role in this architecture is data movement.

Why Use ADF: Instead of manually uploading files, ADF automates and schedules the movement of large datasets from source to Azure storage, ensuring reliability and repeatability.

3■■ Data Storage – Azure Data Lake Storage Gen2 (ADLS Gen2)

Once the data has been extracted via ADF, it is stored in Azure Data Lake Storage Gen2 (ADLS Gen2). This is Microsoft Azure's enterprise-grade cloud storage service, optimized for big data analytics.

How ADLS Works in This Architecture: - Raw Data Storage: The raw, unprocessed data is stored in its original form. - Transformed Data Storage: After processing (described later), the cleaned/processed version of the data is also stored back into ADLS Gen2, in a separate folder or directory.

Structure of ADLS: ADLS Gen2 organizes data inside: - Storage Accounts (main container) - Containers (big folders) - Directories/Subdirectories (like subfolders) - Files (actual data files)

Why Use ADLS Gen2: It supports both structured and unstructured data and allows easy organization using folder structures (hierarchical namespace), making it scalable and efficient for large datasets.

4■■ Data Transformation – Azure Databricks (Using Apache Spark)

With the raw data stored in ADLS Gen2, the next step is data processing or transformation. This is handled using Azure Databricks, which internally uses Apache Spark.

Role of Databricks in This Architecture: Databricks reads the raw data from ADLS Gen2 and applies processing logic using Spark-based code (usually Python or SQL). This processing could involve: - Removing null or duplicate records. - Filtering unnecessary data. - Applying business rules to clean and prepare the data.

What Happens After Processing: The transformed (cleaned and structured) data is written back to ADLS Gen2. This separation between raw and processed data ensures data lineage and allows future reprocessing if needed.

Why Use Databricks: Databricks is ideal for large-scale data processing as it distributes the workload across multiple servers using Spark, enabling quick processing even for very large datasets.

5■■ Data Analytics – Azure Synapse Analytics

Once the data has been transformed and stored, the next step is analysis. This is handled using Azure Synapse Analytics.

Role of Synapse in This Architecture: Synapse connects directly to the transformed data stored in ADLS Gen2 and allows running analytical queries. Using SQL-like commands, you can analyze: - Which country won the most gold medals. - Top-performing athletes. - Overall performance summaries.

Why Use Synapse: Synapse allows running complex queries over large datasets without needing to load the data into a database first. This saves time and reduces costs.

6■■ Reporting – Power BI / Looker / Tableau

The final layer of the architecture is reporting and visualization. Processed data (from Synapse or directly from ADLS) is connected to a reporting tool such as Power BI, Looker Studio, or Tableau.

Role of Reporting Tools: These tools allow creating: - Dashboards. - Charts. - Visual reports.

This enables stakeholders to easily view insights without needing to query raw data themselves.

7■■ Complete Flow – How Everything Connects

Here’s the full sequence: 1. Raw data (CSV or similar) is extracted from its source. 2. Azure Data Factory (ADF) loads this data into ADLS Gen2. 3. Azure Databricks reads the raw data, applies transformations using Spark, and writes back the cleaned data to ADLS Gen2. 4. Azure Synapse Analytics runs analytical queries on the transformed data. 5. Reporting tools like Power BI visualize insights through dashboards.

Key Benefits of This Architecture

- Scalable: Can handle large datasets. - Modular: Each component (ingestion, storage, processing, analysis, reporting) is independent. - Cost-Efficient: Storage and compute resources are managed separately. - Enterprise-Ready: Secure and compliant with corporate standards.