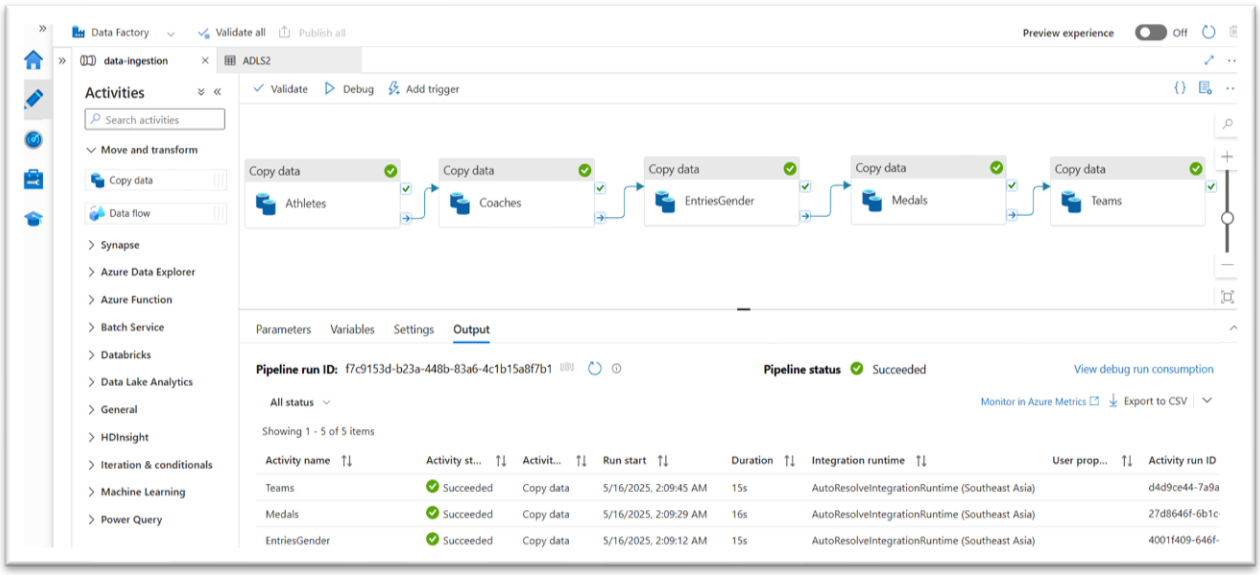
Olympic Data Analytics Using Azure

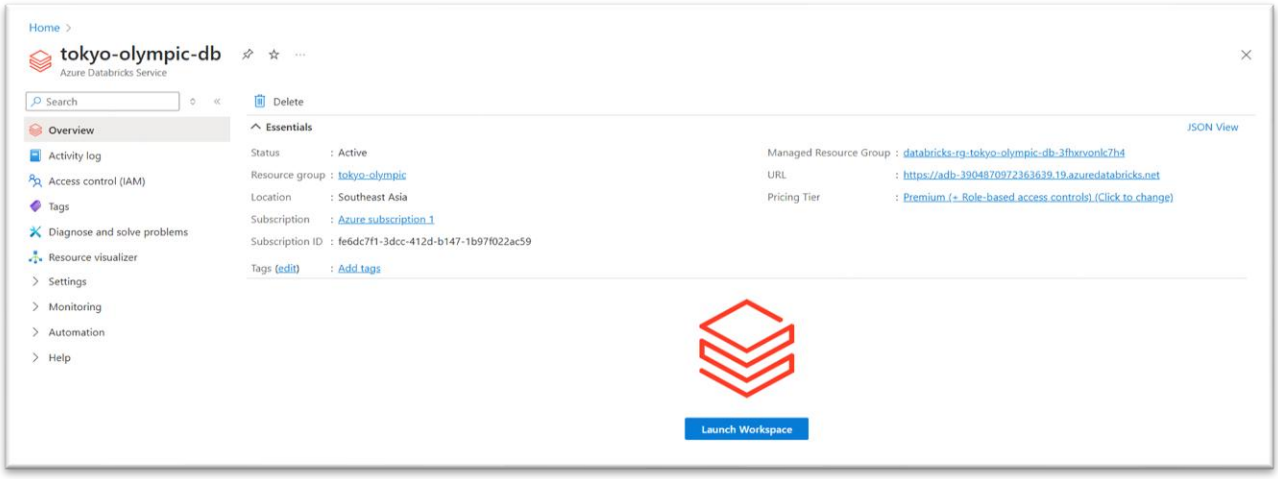


Azure Data Factory – Ingestion Layer Description:

Azure Data Factory was used as the ingestion layer to automate the loading of raw Olympic data (CSV format) into Azure Data Lake Storage Gen2. The pipeline was designed to:

- Connect to data sources (e.g., local uploads or public repositories)
- Ingest multiple CSV files.
- Store raw data into a structured folder hierarchy under the *Raw Zone* of the Data Lake
- Schedule and orchestrate batch ingestion workflows

Each file was loaded using a dedicated Copy Data activity, ensuring the data was efficiently transferred and logged for monitoring. This modular design allows easy scaling and reusability for future Olympic datasets.



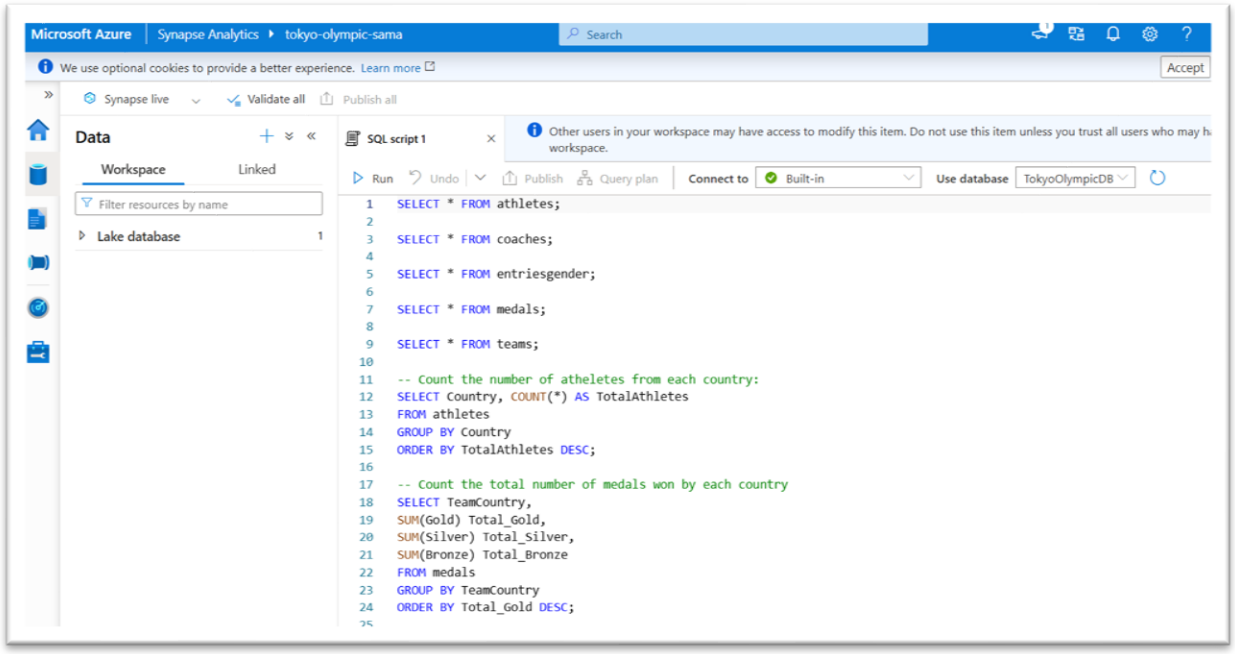
Azure Databricks – Transformation Layer Description:

Azure Databricks served as the transformation layer, where raw Olympic datasets ingested into Azure Data Lake were cleaned, structured, and enriched using PySpark.

Key transformations performed include:

- **Schema enforcement** and null value handling for consistency
- **Filtering and deduplication** of athlete and coach records
- **Creating calculated columns** (e.g., total medals, athlete counts)
- **Writing clean data** to the *Transformed Zone* of Azure Data Lake Gen2 in CSV format for optimized analytics

This layer ensured that only high-quality, analysis-ready data moved forward into the analytics and visualization layers.



Azure Synapse Analytics – Analytics Layer Description:

Azure Synapse Analytics was used as the analytics layer to query and analyze the transformed Olympic datasets stored in Azure Data Lake Gen2. By creating SQL-on-demand (serverless) views, Synapse enabled:

- **Fast querying** of CSV files in the Transformed Zone
- **Data exploration** across athletes, coaches, and medal statistics
- **Aggregation and filtering** for key metrics like medals by country, athlete participation by sport, and gender distribution

These queries served as the backend for Power BI dashboards, delivering efficient, scalable insights without the need to move or copy data.

Power BI Dashboard:

Power BI was used to build an interactive dashboard that visualizes key insights from the Olympic dataset, including medal counts by country, athlete distribution by sport, and gender participation. The dashboard connects directly to transformed data for real-time reporting.