Tokyo Olympics 2021

Requirement Understanding Document

**Objective:**

* A company wants to use the Azure cloud platform to store big data in different formats in a cost-effective way.
* This data will be stored in different formats which need to be consumed by business users for multiple purposes like descriptive analytics.
* This pipeline should pick data from a source and ingest it into the data lake.
* The data lake should contain a landing zone, bronze zone (raw data), silver zone (curated data), and gold zone (aggregated data).
* The business user should be able to effectively & easily use this data using simple & popular tools like Power BI.
* As a Data Engineer, you need to design and implement an end-to-end solution that can be scalable, cost-effective, and can be maintained easily.

**Problem Statement :**

This project focuses on processing and analysing data related to the 2021 Tokyo Olympics. The dataset consists of over 11,000 athletes across 46 disciplines and 743 teams. It includes detailed information about: Athletes, Coaches, Teams, Disciplines, Medals, Entries by Gender, Team Performance, etc.

The goal is to build a data pipeline that ingests this dataset into the Azure Data Lake landing zone and processes the data through the bronze, silver, and gold zones, and makes it accessible for useful descriptive analytics. Business users should be able to access key insights—such as medal distribution by country, athlete performance, and gender participation—via tools like Power BI.

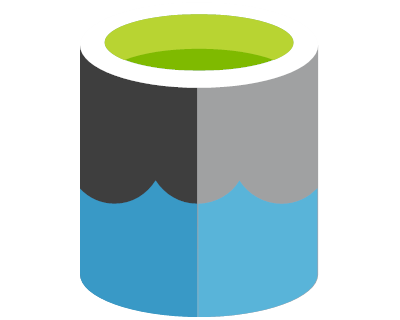
**Source Data Realization:**

The source Data is present in the On-premises SQL server in the form of tables. The tables are :

* **Athletes** : This table contains individual records for each athlete participating in the event, capturing essential attributes:
* PersonName: The full name of the athlete.
* Country: The country the athlete represents.
* Discipline: The sport or event in which the athlete is competing.
* **Coaches** : This table provides details about the coaches involved in the competition, focusing on their roles and the events they oversee:
* Name: The full name of the coach.
* Country: The country represented by the coach.
* Discipline: The sport or discipline the coach is involved in.
* Event: The specific event or competition the coach is responsible for.
* **EntriesGender** : This table highlights the gender breakdown of participants across different disciplines:
* Discipline: The sport or event being measured.
* Female: The number of female participants.
* Male: The number of male participants.
* Total: The total number of participants, both male and female.
* **Medals** : This table focuses on the medal count for each country, providing a summary of performance across various events:
* Rank: The rank of the country based on gold medals.
* Team\_Country: The name of the country or team.
* Gold: The number of gold medals won.
* Silver: The number of silver medals won.
* Bronze: The number of bronze medals won.
* Total: The total number of medals won.
* Rank by Total: The rank of the country based on the total number of medals.
* **Teams** : This table shows the teams participating in the event, focusing on their composition and roles:
* TeamName: The official name of the team.
* Discipline: The sport or discipline the team competes in.
* Country: The country represented by the team.
* Event: The specific event in which the team competes.

**Resources Required :**

**Azure Data Lake Storage Gen2**:

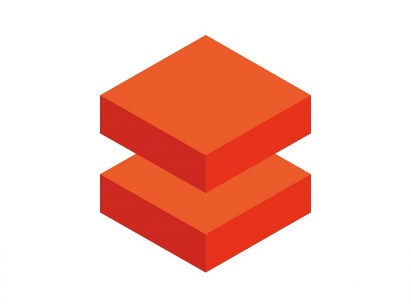
* **Use Case**: Acts as the primary storage system for data at different stages. Initially, it stores raw data extracted from the on-premises SQL Server in the Landing Zone. This data is then copied as-it-is to the Raw Zone (Bronze Layer). Subsequently,   
  Holds cleaned or processed and transformed   
  data in the Silver and Gold Zones respectively.

**Azure Data Factory**:

* A blue pipe with white squares and a black background

  Description automatically generated**Use Case**: An ETL tool that extracts data from on-premises SQL Server to Azure Data Lake Storage (ADLS). It creates pipelines for data movement, processing, and transformation, and integrates with services like Azure Databricks and SQL for scalable, automated data handling and analysis preparation.

**Azure Databricks Notebook**:

* **Use Case**: Utilized for data cleansing and transformation. In this workflow, it processes the raw data from the Bronze Layer, removing null values and corrupted data and stores it in the Silver Layer. It then transforms the cleansed data into a more refined format, suitable for analysis, which is stored in the Gold Layer.

**Azure Delta Table**:

* A blue triangle with black text

  Description automatically generated**Use Case**: Delta format enables efficient   
  storage of cleansed and transformed data with ACID guarantees. It supports real-time and batch processing, incremental updates, and data versioning, ensuring optimal performance. Delta tables allow Power BI and other tools to query and analyze data seamlessly, providing quick access to aggregated insights for   
  reporting and analysis.

**Azure Logic Apps**:

* A blue and green symbol with a square and square rectangle

  Description automatically generated**Use Case**: Automates the workflow notifications and updates. It is   
  configured to send email alerts about the status of data pipeline activities,   
  such as the completion of data ingestion, cleansing, and transformation   
  processes, ensuring admins or analysts   
  are informed in real-time.

**Azure Key Vault & Microsoft Entra ID**:

* **Use Case**: Ensures the security and integrity of the data throughout the workflow. Azure Key Vault securely stores secrets, keys, and credentials required by various services, while Microsoft Entra ID manages user authentication and access control, ensuring only authorized personnel   
  can access sensitive data and services.

**Power BI**:

* **Use Case**: Connects to the Gold Layer of the Azure Delta Lake table to visualize and generate reports from the useful, transformed data. It provides interactive dashboards and analytical insights, facilitating data-driven decision-making based on the processed data.

**Workflow :**

The workflow for this project includes several steps :

1. Configuring Resources – Configuring and setting up the required resources such as creation of suitable and feasible Data pipelines, activities, Linked Services, Integration Runtimes, Datasets in pipelines, Azure Data Lake Storage, Azure databricks notebooks, Azure delta lake table, Logic Apps, Azure Key vault and Power Bi.
2. Extraction of Source Data – The source initial data is extracted from the On-premises SQL Server location and ingested to Azure Data Lake Storage Gen2 in the Landing Zone using the Self-Hosted Integration Runtime service.
3. Adding CreatedTime and ModifiedTime columns – Adding the CreatedTime and ModifiedTime columns in the ingested data tables.
4. Creating the Metastore/Control Table – Creating the metadata or control table in the Azure SQL Server that will contain the data table names and related information. This will act as the metadata or control table to keep track of the pipeline run and data processing.
5. Landing zone to Raw zone (Bronze Layer) – The data in the landing zone is copied ‘as-it-is’ to the raw zone (Bronze layer) for making the landing zone available to store the upcoming incremental data.
6. Cleansing the Data – The data present in the bronze layer or raw zone is cleansed and curated to remove the null and corrupted data using the Azure Databricks. The clean and curated data is then stored in the silver layer in the form of delta format and as external delta table.
7. Transformation of curated data into useful data – The cleansed and curated data present in the silver layer is then transformed using the Azure Databricks and Spark SQL into useful data that would be used for Analysis and Report Generation using Analytics and Visualization tools like Power Bi , etc. The useful data is then stored in the gold layer in delta format and as external delta table.
8. Visualization and Report Generation – Visualization of the useful data using Power Bi tool and generating useful reports. Connecting the Power Bi with the Azure Databricks Delta table to retrieve the useful data present in the gold layer delta lake table.
9. Privacy and Security – Maintaining Privacy and Security of the useful data by configuring Microsoft Entra ID and Azure Key Vault services.
10. Email and Notifications – Configuring Azure Logic Apps to notify and update about the status of the Data pipelines and activities through email notifications.