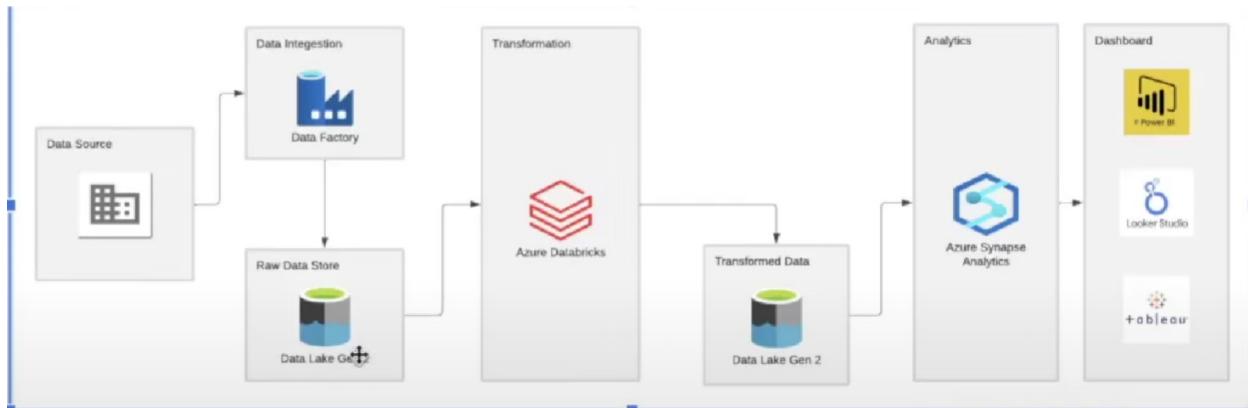


1. Project Goal

The primary objective of this data engineering project is to systematically traverse through the key phases of the Extract, Transform, Load (ETL) process and culminate in the development of an insightful dashboard. The project leverages various applications, including Azure Storage Account, Azure Data Factory, Databricks, Synapse Analytics, and Power BI, to accomplish this goal.



2. Data Source

Details

This contains the details of over 11,000 athletes, with 47 disciplines, along with 743 Teams taking part in the 2021(2020) Tokyo Olympics. This dataset contains the details of the Athletes, Coaches, Teams participating as well as the Entries by gender. It contains their names, countries represented, discipline, gender of competitors, name of the coaches and medals(gold, silver, bronze).

Source: Tokyo Olympics 2020 Website

3. Creating Azure storage account

The screenshot shows the Microsoft Azure Storage Explorer interface. The top navigation bar includes 'Home >', a search bar, and user information ('olegbudachov@gmail.c... DEFAULT DIRECTORY'). Below the navigation is a toolbar with icons for Upload, Add Directory, Refresh, Rename, Delete, Change tier, Acquire lease, Break lease, and Give feedback. The main area displays the 'tokyo0-olympic-data' container. On the left, a sidebar shows 'Overview', 'Diagnose and solve problems', 'Access Control (IAM)', 'Settings' (Shared access tokens, Manage ACL, Access policy, Properties, Metadata), and a 'Search' bar. The main content area shows a table of blobs:

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
raw-data					-	---
transformed-data					-	---

Within the container, we have established separate divisions: one designated for 'raw data' and the other for 'transformed data' to facilitate organization and accessibility.

4. Using Azure data factory for importing data source to the Data Lake Storage

The screenshot shows the Microsoft Azure Data Factory portal. At the top, there's a navigation bar with 'Microsoft Azure | Data Factory > tokyo-olympic-df-budachov'. Below the navigation is a search bar labeled 'Search factory and documentation'. On the right side of the header, there's a user profile with the email 'olegbudachov@gmail.com' and a link to 'DEFAULT DIRECTORY'. The main content area is titled 'Data factory' and shows the name 'tokyo-olympic-df-budachov'. Below the title, there's a 'New' button. To the right of the title, there's a large blue industrial building icon representing data processing. Underneath the title, there are four service cards: 'Ingest' (Copy data at scale once or on a schedule), 'Orchestrate' (Code-free data pipelines), 'Transform data' (Transform your data using data flows), and 'Configure SSIS' (Manage & run your SSIS packages in the cloud).

Verifying the successful upload of data within Data Factory

This screenshot shows the Microsoft Azure Data Factory portal focusing on the 'data-ingestion' pipeline activity. The left sidebar lists various activities: Move and transform (Copy data, Data flow), Synapse, Azure Data Explorer, Azure Function, Batch Service, Databricks, Data Lake Analytics, General, HDInsight, Iteration & conditionals, Machine Learning, and Power Query. A red arrow points to the 'Validate all' button next to the pipeline name. The main area is titled 'Preview data' and shows a table of data from the 'AthletesHTTP' linked service. The table has columns: PersonName, Country, and Discipline. The data includes 10 rows of athletes from different countries and disciplines. Below the preview, there's a 'Request body' section. At the bottom of the screen, there's a table of uploaded files in the Data Lake Storage:

Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
[...]						...
athletes.csv	11/7/2023, 4:25:48 PM	Hot (Inferred)		Block blob	408.68 KiB	Available
coaches.csv	11/7/2023, 4:26:03 PM	Hot (Inferred)		Block blob	16.49 KiB	Available
entriesGender.csv	11/7/2023, 4:26:18 PM	Hot (Inferred)		Block blob	1.1 KiB	Available
medals.csv	11/7/2023, 4:26:33 PM	Hot (Inferred)		Block blob	2.36 KiB	Available
teams.csv	11/7/2023, 4:26:49 PM	Hot (Inferred)		Block blob	34.44 KiB	Available

A red arrow also points to the 'athletes.csv' file in the list.

5. Leveraging Databricks for exploratory analysis and data transformation processes

The screenshot shows two Databricks notebooks running on the 'tokyo-olympic-db' cluster.

Notebook 1: Tokyo Olympic Transformation

```

1 configs = {"fs.azure.account.auth.type": "OAuth",
2 "fs.azure.account.oauth.provider.type": "org.apache.hadoop.fs.azurebfs.oauth2.ClientCredsTokenProvider",
3 "fs.azure.account.oauth2.client.id": "REDACTED",
4 "fs.azure.account.oauth2.client.secret": "REDACTED",
5 "fs.azure.account.oauth2.client.endpoint": "https://login.microsoftonline.com/REDACTED/oauth2/v2.0/token"}
6
7
8 dbutils.fs.mount(
9 source = "abfs://tokyo-olympic-data@tokyoolympicdatabudachov.dfs.core.windows.net", # contrainer@storageacc
10 mount_point = "/mnt/tokyoolympic",
11 extra_configs = configs)
12
13
Out[1]: True
Command took 12.45 seconds -- by olegbudachov@gmail.com at 11/7/2023, 5:40:28 PM on Oleg Budachov's Cluster

```

Notebook 2: Tokyo Olympic Transformation

```

1 %fs
2 ls "/mnt/tokyoolympic"
Table + New result table: OFF
+-----+-----+-----+-----+
| path | name | size | modificationTime |
+-----+-----+-----+-----+
1 dbfs:/mnt/tokyoolympic/raw-data/ raw-data/ 0 1699380787000
2 dbfs:/mnt/tokyoolympic/transformed-data/ transformed-data/ 0 1699380807000
+-----+-----+-----+-----+
2 rows | 9.09 seconds runtime
Command took 9.09 seconds -- by olegbudachov@gmail.com at 11/7/2023, 5:48:34 PM on Oleg Budachov's Cluster

```

Notebook 3: Tokyo Olympic Transformation

```

1 entriesGender.printSchema()
root
|-- Discipline: string (nullable = true)
|-- Female: string (nullable = true)
|-- Male: string (nullable = true)
|-- Total: string (nullable = true)

Command took 0.05 seconds -- by olegbudachov@gmail.com at 11/7/2023, 6:06:59 PM on Oleg Budachov's Cluster

```

Notebook 4: Tokyo Olympic Transformation

```

1 entriesGender = entriesGender.withColumn("Female", col("Female").cast(IntegerType()))\
2 .withColumn("Male", col("Male").cast(IntegerType()))\
3 | .withColumn("Total", col("Total").cast(IntegerType()))
entriesGender: pyspark.sql.dataframe.DataFrame = [Discipline: string, Female: integer ... 2 more fields]
Command took 0.23 seconds -- by olegbudachov@gmail.com at 11/7/2023, 6:14:22 PM on Oleg Budachov's Cluster

```

Notebook 5: Tokyo Olympic Transformation

```

1 entriesGender.printSchema()
root
|-- Discipline: string (nullable = true)
|-- Female: integer (nullable = true)
|-- Male: integer (nullable = true)
|-- Total: integer (nullable = true)

Command took 0.12 seconds -- by olegbudachov@gmail.com at 11/7/2023, 6:14:36 PM on Oleg Budachov's Cluster

```

6. Synapse Analytics

We utilize Synapse Analytics after the data has been transformed through Databricks and uploaded to the transformed data storage for its comprehensive analytical capabilities. Synapse Analytics offers advanced querying, data warehousing, and scalable processing, enabling in-depth analytics, performance optimization, and seamless integration of transformed data for further insights and reporting.

The figure consists of four vertically stacked screenshots of the Microsoft Azure Synapse Analytics portal. The top two screenshots show the 'Synapse live' workspace for the 'tokyo-olympic-sa-bud' project. The left side shows a dashboard with 'Ingest', 'Explore and analyze', and 'Visualize' options. The right side shows a query editor with a SQL script window containing:

```
-- Count the number of athletes in each country
SELECT Country, COUNT(*) AS Total_Athletes
FROM athletes
GROUP BY Country
ORDER BY Total_Athletes DESC;
```

The results table shows the total number of athletes per country:

Country	Total_Athletes
United States of America	615
Japan	586
Australia	470
People's Republic of China	401
Germany	400
France	377
Canada	368

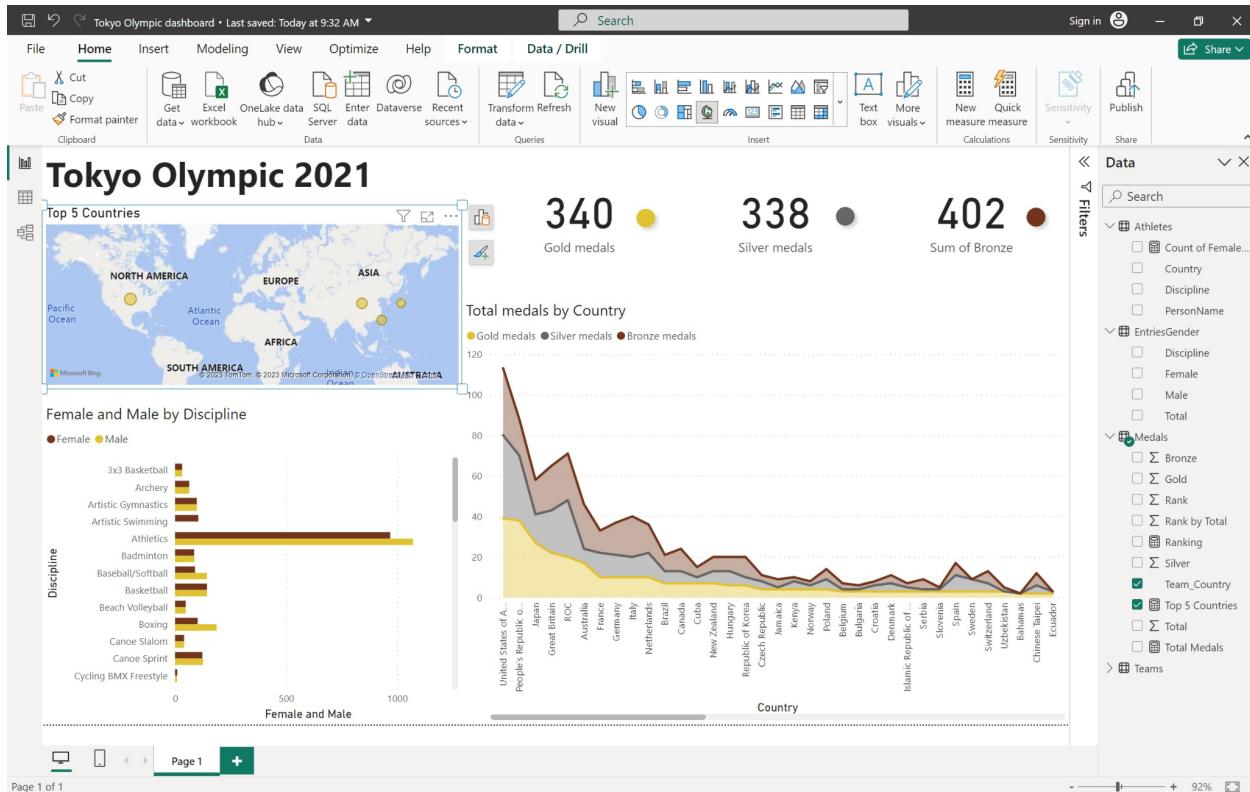
The bottom two screenshots show the same workspace and query editor, but the results pane now displays a bar chart titled 'TotalGold' showing the total number of gold medals by country. The chart includes configuration settings on the right:

- Chart type: Column
- Category column: Team_Country
- Legend (series) columns: TotalGold
- Legend position: bottom - center
- Legend (series) label: TotalGold
- Legend (series) minimum value: 0

At the bottom of each screenshot, a message indicates the query was executed successfully.

7. Creating Power BI Dashboard

We meticulously shape the data in Power Query Editor within Power BI by executing various tasks: splitting and removing old columns, crafting new columns, adjusting data types, establishing a data model with appropriate relationships. Subsequently, we design a dashboard showcasing key highlights concerning the Tokyo Olympic 2021, offering a comprehensive overview of pertinent information.



8. Conclusion

The project culminated in an extensive data transformation process involving various stages. Initially, the data underwent cleansing, reformatting, and structuring within Databricks, followed by uploading the transformed data to storage facilitated by Data Factory. The curated dataset was then further refined and shaped using Power Query Editor within Power BI. Throughout these stages, the old columns were modified, new columns were created, data types were adjusted, and a robust data model was established, ensuring appropriate relationships for comprehensive insights. This diligent process paved the way for the creation of an informative and visually compelling dashboard, focusing on the major highlights of the Tokyo Olympic 2021.