**Olympics Dataset**

**Data Storage**: We can start with local storage for initial testing and then move to AWS S3 when we are ready for “production”. Or we can just go ahead with AWS S3, cost will mostly depend on how much data we have, personally I don’t think so it should be a problem.

**Data Transformation**: If we move forward with AWS S3 then we should definitely go ahead with AWS Glue. If we are storing it locally like for initial testing etc then we can obviously transform data using python scripts but it depends on the size of the data we have.  
  
**Data Visualization**: We can integrate AWS Glue with Apache Superset(open source) though it isn’t direct we will have to utilize AWS Athena which can be integrated to Superset. AWS Glue processes and catalogs the data, which Athena can then query. Superset can connect to Athena, which will help us to create interactive dashboards and visualizations based on the transformed data.

**KPIs**:

Medal Trends Over Time: Track the total number of medals won by each country over different Olympic games.

Country Dominance in Specific Sports: Identify which countries dominate specific sports and how this has changed over time.

Performance Improvement: Measure the improvement in a country's performance by comparing recent Olympics to past ones.

Top Performers by Sport: Highlight the top athletes or countries in each sport.

**Visualizations**:

Line Charts: To show trends over time for medal counts.

Bar Charts: To compare the performance of different countries in a particular sport.

Heatmaps: To show the concentration of medals by country and sport over different years. Maps: To visualize geographic trends in Olympic performance.

Stacked Bar Charts:Display the breakdown of medal types (gold, silver, bronze) won by countries in different sports or events.

Scatter Plots: To explore relationships between variables, such as team size and medal count.

**Tasks**:

Create web scraper for the Olympics Tokyo 2020 and Paris 2024 data

Store locally first then integrate S3 upload

Setup S3 buckets and store raw CSV data

Brainstorm dashboard visualizations

Define KPIs

Which charts will tell the story?

Design any transformations needed and tables

Do we need ETL or is the raw data structured properly for the necessary visualizations?

* Depends on how the quality of the raw data is.

Implement ETL processes if needed using AWS Glue or Python scripts.

If we are connecting with Apache Superset we will have to set up AWS Glue Data Catalog and integrate with AWS Athena.

Build dashboard