**Olympics Data Analytics Project**

**Goals:**

1. Understand which countries are improving in their Olympics overall performance.
2. Look for country dominance in specific sports and how is that trending over time.

**Data Sources:**

1896 to 2016

<https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results>

2020/21 and 2024

<https://www.olympedia.org/editions/61>

<https://www.lemonde.fr/en/sport/jo-2024/results/>

**Data Check:** According to ChatGPT there were 33 sports and 339 medaling events in Tokyo 2020. And 329 medaling events across 45 different sports in Paris 2024 and 206 countries in Paris.

Confirm that the scraped data matches with the expected counts.

**Data Architecture:**

Storage: S3 or ??

Transformation: AWS Glue or ??

Visualization: Superset (No Cost) or ??

**Data Pipeline:**

* Raw (Kaggle\_CSV + Scraped\_2020\_2024\_CSV)
* Aggregated Data for Visualization (may not be needed)

**Tasks:**

1. Create web scraper for the Olympics Tokyo 2020 and Paris 2024 data
   1. Store locally first then integrate S3 upload
2. Setup S3 buckets and store raw CSV data
3. Brainstorm dashboard visualizations
   1. Define KPIs
   2. Which charts will tell the story?
4. Design any transformations needed and tables
   1. Do we need ETL or is the raw data structured properly for the necessary visualizations?
5. Build dashboard