Milestone 1: Project Proposal and Data Selection/Preparation

Step 1: Preparing for your Proposal

1. Which client/dataset did you select and why?
    - My client/dataset is SportsStats. It is a sports analysis firm partnering with local news and elite personal trainers to provide “interesting” insights to help their partners. Insights could be patterns/trends highlighting certain groups/events/countries, etc. for the purpose of developing a news story or discovering key health insights.
2. Describe the steps you took to import and clean the data.
    - I downloaded the datasets from Coursera. The data is in a csv file. I used Databricks to import and clean my data. The data was already pretty clean, so I didn’t have much to do.
3. Perform initial exploration of data and provide some screenshots or display some stats of the data you are looking at.
    - Determining the overall athlete ranking

```
1  select Name, Medal, count(Medal)
2  from delta_athlete_events
3  where Medal != 'No medal'
4  group by Name, Medal
5  order by count(Medal) desc
```

(2) Spark Jobs

Table +

	Name	Medal	count(Medal)
1	Michael Fred Phelps, II	Gold	23
2	Raymond Clarence "Ray" Ewry	Gold	10
3	Paavo Johannes Nurmi	Gold	9
4	Mark Andrew Spitz	Gold	9
5	Frederick Carlton "Carl" Lewis	Gold	9
6	Larysa Semenivna Latynina (Diriy-)	Gold	9
7	Ole Einar Birndalen	Gold	8

10,000 rows | Truncated data | 3.29 seconds runtime

- Determining the overall Team ranking

```
select Team, count(Medal) as gold_medal
from delta_athlete_events
where Medal = 'Gold'
group by Team
order by gold_medal desc
```

(2) Spark Jobs

Table +

	Team	gold_medal
1	United States	2474
2	Soviet Union	1058
3	Germany	679
4	Italy	535
5	Great Britain	519
6	France	455
7	Sweden	451

242 rows | 2.60 seconds runtime

- Determining the number of athletes by Team

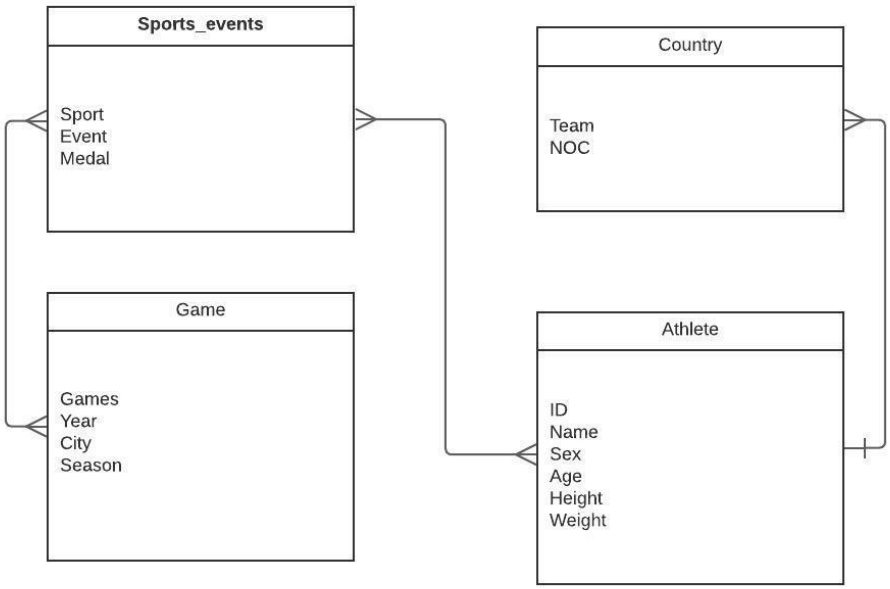
```
select Team, count(distinct Name) as athlete_count
from delta_athlete_events
group by Team
order by athlete_count desc
```

(3) Spark Jobs

Table +

	Team	athlete_count
1	United States	9114
2	France	5777
3	Great Britain	5758
4	Italy	4688
5	Germany	4569
6	Canada	4546
7	Japan	3981

4. Create an ERD or proposed ERD to show the relationships of the data you are exploring.



Step 2: Develop Project Proposal

Description

My project analyses Olympic games performances overtime by athlete. It gives a better understanding on what athlete, what country won a medal in a given Olympic game. The goal of this project is to identify key patterns on the winning element at the Olympics. The audience can be country sport federations, journalists, CIO, bookmakers, Olympics enthusiasts...

Questions

- Which team/country has the most chance of getting a medal at a given sport?
- What kind of athlete has the most chance of getting a medal at a given sport?
- Does more athletes mean more medals?

Hypothesis

- The more athletes your team have, the higher chances you have to a win a medal
- Tall athletes have a higher of winning at sports like basketball, swimming...

Approach

I am going to filter and group by sports to see how the other columns are correlated or not. I will group by team as well to see if there are any key patterns. Also, I will look at the historical performances of teams and athletes to see if there are signs for future great performances.