Project Report: Advanced Data Management

Participants
4709433: Sanket Rajeev Sabharwal
4714800: Humberto Enrique Della Torre Escobar

Introduction
The aforementioned project for the subject Advanced Data Management entails the analysis of information related to the Olympic games – particularly covering historical data of 120 years, from 1896 to 2016.

Through the project, we have gone through multiple iterations, and have documented every phase with supplementing our submission with python files that were originally used as well.

Also, we with our files we have uploaded a presentation that we intend to present during the submission on the specific date. For working on this project, we used github folder, where the codes were worked upon together, and can be found **here**.

We have uploaded two files, ADM_Project_Final.py and ADM_Project_Final-UniversityCluster.py. The file 'ADM_Project_Final-UniversityCluster' has been programmed to provide adequate results on the server used for our course, since we observed that a few libraries we have used in our project were not operational on the course cluster. Nevertheless, we have provided two files to ensure that they work in the desired manner offline, and in the appropriate method online (the University Cluster).

Dataset
Original (Kaggle): **Click Here**
Modified: **Click Here**

Workload
The workload for our project has been identified as following:
Query 1: Enlist the number of medals won by each country by each session
Query 2: Highlight which countries dominate each of the two sessions
Query 3: Total medal distribution over the years
Query 4: Highest and lowest medal count so far by country
Query 5: Highlight the ratio of female to male participation over the years
Query 6: Highlight the all time ratio of female to male participation amongst top participating countries
Query 7: Representation of every country in each sport
Query 8: Top 10 participating countries for each session
Query 9: Top 10 participating countries all time
Query 10: Highest participation over the years, and participation over the years
Query 11: Highest and lowest 10 (by country) average weights and heights
Query 12: Top 10 athletes by medal count, and medal breakdown
Query 13: Top 10 gold medal winners
Query 14: Top 10 athletes with most years participated
Query 15: Youngest and oldest athletes of all time
Query 16: Top 10 athletes with most number of sports participated
Query 17: Swimmers with higher than the average height who have won medals
Query 18: Weightlifters with higher than the average weight who have won medals
Query 19: Cities where the olympics have been hosted the most number of times
Query 20: Average height and weight of athletes over 10 year intervals
Query 21: Various details on olympic art competitions
Query 22: Summer olympics sport and event diversity through the years
Query 23: Winter olympics sport and event diversity through the years

The above mentioned workload helps observe interesting patterns and trends pertaining to the Olympic games.

Instructions
Shareable drive link with the dataset can be found **here**.

For **ADM_Project_Final.py**:
Code utilizes findspark and pyspark and requires environmental variables for spark and java homes. Developed on our local machines to utilize python libraries such as matplotlib.

Run ADM_Project_Final.py as all the required code is contained in this python file.

Enter the filepath of the given dataset 'DataSetFinal.csv' when requested.
If issues concerning filepath format arise in the python file try adding the filepath as shown in the spark.read.csv function in line 487 and then running line 488.

Upon succesful filepath entry and dataframe creation when requested enter a query #(1-23) to execute the query, 'h' to list the queries, or 'q' to quit.

For **ADM_Project_Final-UniversityCluster.py**:
Code runs in the university cluster utilizing pyspark.

The filepath is set in line 361. Currently set as:
hdfs://master:9000/user/user9/olympics/input/DataSetFinal.csv.
Can be changed if desired, but be sure to use the provided 'DataSetFinal.csv' file.

To display the available queries run the following command
spark-submit --master yarn ADM_Project_Final-UniversityCluster.py h

To execute a query run the following command
spark-submit --master yarn ADM_Project_Final-UniversityCluster.py <query #>

Example command and output:
spark-submit --master yarn ADM_Project_Final-UniversityCluster.py 13

```
+--------------------------------------------------+-----------+
|Name                                              |Gold medals|
+--------------------------------------------------+-----------+
|Michael Fred Phelps, II                           |23|
|"Raymond Clarence ""Ray"" Ewry"                   |10|
|Mark Andrew Spitz                                 |9|
|Larysa Semenivna Latynina (Diriy-)                |9|
|"Frederick Carlton ""Carl"" Lewis"                |9|
|Paavo Johannes Nurmi                              |9|
|Birgit Fischer-Schmidt                            |8|
|Usain St. Leo Bolt                                |8|
|"Matthew Nicholas ""Matt"" Biondi"                |8|
|"Jennifer Elisabeth ""Jenny"" Thompson (-Cumpelik)"|8|
+--------------------------------------------------+-----------+
```