COMP6210 - Big Data Assignment 1

MapReduce

Semester 2, 2024

School of Computing, Macquarie University

Dataset:

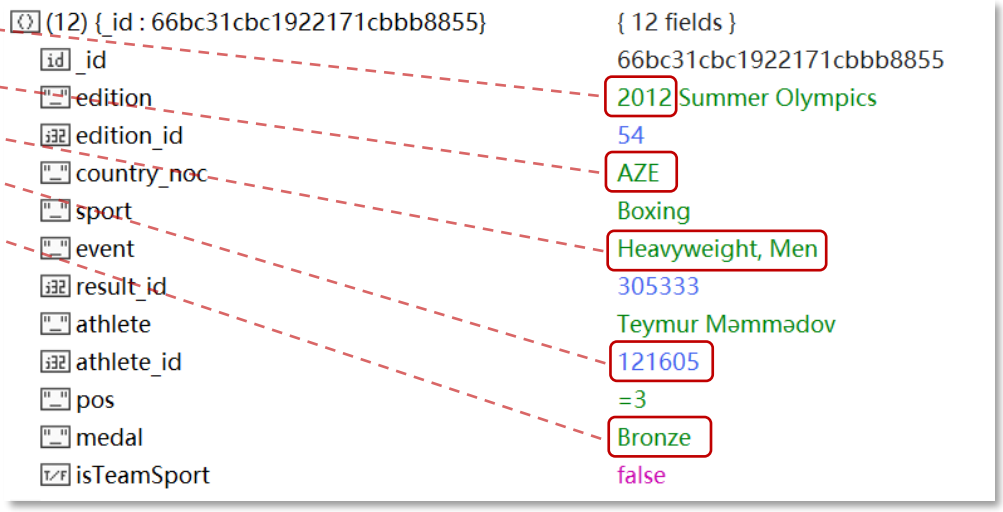
The Olympic historical dataset "Olympic_Athletes.zip" is available on iLearn. This dataset contains information about athletes who participated in the Summer and Winter Olympic Games from 1896 to 2022.

Programming Environment:

- **MongoDB & Studio 3T:** Used for creating databases and importing datasets into collections.
- **Pymongo:** Used for connecting to MongoDB and extracting information from documents within collections.
- **Mrjob:** Used for implementing MapReduce programs.

Task 1: Data Curation (20 marks)

- **Task 1.1 - Data Extraction (10 marks)**: Extract information about medal-winning athletes in the Summer Olympics from the 1980 edition to the 2020 edition. Note that you need to verify whether the athlete participated in the Summer Olympics and whether they won a medal (gold, silver, or bronze). Then, extract the following values:
  - year
  - country
  - event
  - athlete id
  - medal



For each qualified athlete, create an entry in the format: <id, country, year, event, medal>. Store these entries in a text file named "athletes.txt". Refer to the following screenshot for formatting (note that the screenshot is for reference purpose only; actual results may vary).

```
132627  AZE 2016    Welterweight, Greco-Roman, Men  Bronze
132628  AZE 2016    Super-Heavyweight, Greco-Roman, Men Bronze
132626  AZE 2016    Featherweight, Freestyle, Men   Bronze
121429  AZE 2016    Welterweight, Freestyle, Men    Silver
121430  AZE 2016    Middleweight, Freestyle, Men    Bronze
121432  AZE 2016    Light-Heavyweight, Freestyle, Men   Bronze
117109  AZE 2016    Heavyweight, Freestyle, Men Silver
116984  AZE 2016    Flyweight, Freestyle, Women Silver
117004  AZE 2016    Featherweight, Freestyle, Women Bronze
140915  AZE 2020    Light-Heavyweight, Men  Bronze
121656  AZE 2020    Heavyweight, Women  Bronze
140924  AZE 2020    Kumite, ≤75 kg, Men Silver
140906  AZE 2020    Kumite, >61 kg, Women   Silver
140928  AZE 2020    Middleweight, Greco-Roman, Men  Bronze
132626  AZE 2020    Welterweight, Freestyle, Men    Silver
116984  AZE 2020    Flyweight, Freestyle, Women Bronze
65090   BAR 2000    100 metres, Men Bronze
```

athlete id

You can also use other delimiters, such as commas, semicolons, underscores, or quotation marks, to separate each line's values. The "**athletes.txt**" text file will then serve as the input for the subsequent MapReduce programs.

- **Task 1.2 - Data Organization (10 marks):** Using the generated "athletes.txt" file as input, implement a MapReduce program to sort the data in ascending order based on the athlete id. The partial results of Task 1.2 are similar to the following screenshot (note that the screenshot is for formatting reference only; actual results may vary).

```
132562  ["AUS","2020","Doubles, Mixed","Bronze"]
132581  ["AUT","2020","Discus Throw, Men","Bronze"]
132593  ["AUT","2016","Multihull, Mixed","Bronze"]
132608  ["AZE","2016","Middleweight, Men","Bronze"]
132609  ["AZE","2016","Light-Welterweight, Men","Silver"]
132622  ["AZE","2016","Welterweight, Men","Bronze"]
132623  ["AZE","2016","Heavyweight, Men","Gold"]
132626  ["AZE","2016","Featherweight, Freestyle, Men","Bronze"]
132626  ["AZE","2020","Welterweight, Freestyle, Men","Silver"]
132627  ["AZE","2016","Welterweight, Greco-Roman, Men","Bronze"]
132628  ["AZE","2016","Super-Heavyweight, Greco-Roman, Men","Bronze"]
```

athlete id ascending order

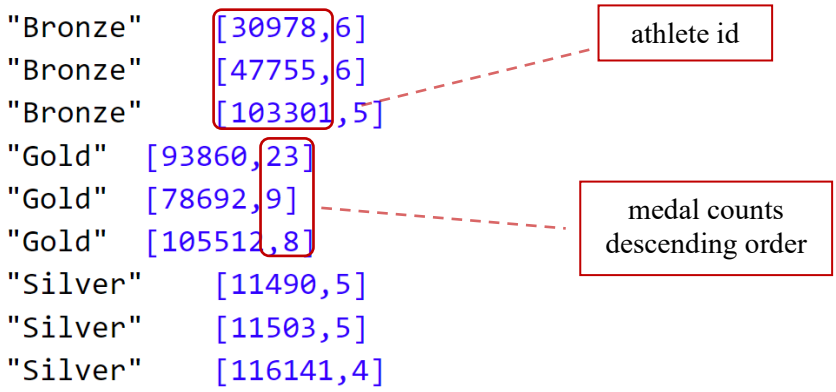
Note that for records with the same athlete ID, there is no specific requirement regarding their order.

Task 2: Data Analysis with MapReduce (60 marks)

Using the generated "**athletes.txt**" file as input, implement three MapReduce programs to complete the following analysis tasks.

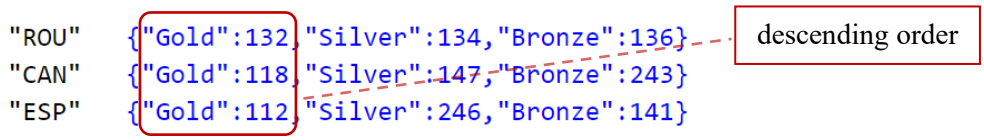
- **Task 2.1 (20 marks)** Find the top three athletes who won the most number of medals in each category (gold, silver, and bronze) in 1980-2020. Firstly, you need to calculate the total number of medals each athlete has earned in gold, silver, and bronze categories, respectively. Next, sort the athletes in descending order based on their medal counts for each category. Finally, for each medal category, output the top three athletes along with their respective medal counts.

Note that there is no specific requirement regarding the order of medal categories. The partial results of Task 2.1 are similar to the following screenshot (the screenshot is for formatting reference only; actual results may vary).



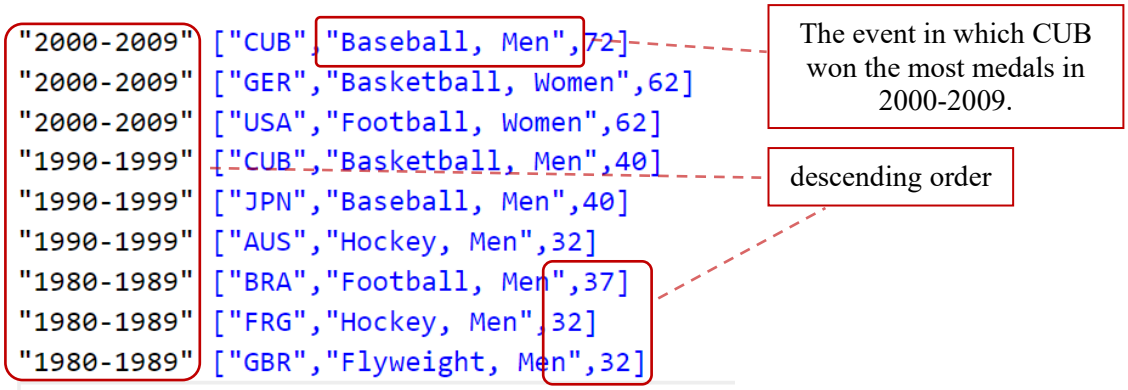
- **Task 2.2 (20 marks)** Find the top three countries with the most number of gold medals in 1980-2020. First, you need to count the total number of each medal type (gold, silver, and bronze) for each country. Then, sort the countries in descending order based on their gold medal count. Finally, output the top three countries along with their medal counts for all medal types (gold, silver, and bronze).

  The partial results of Task 2.2 are similar to the following screenshot (note that the screenshot is for formatting reference only; actual results may vary).



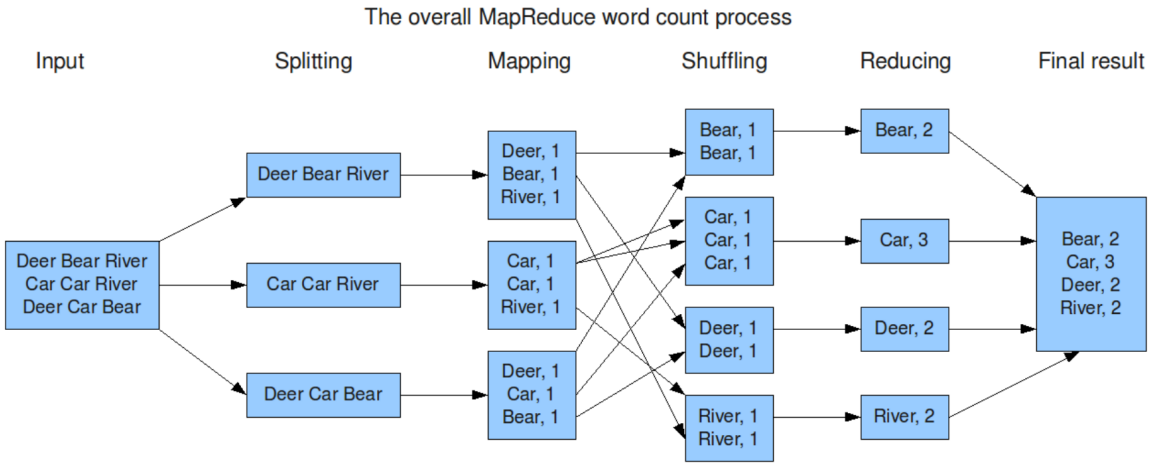
- **Task 2.3 (20 marks)** Find the top three events with the highest medal counts for each decade in 1980-2020. Firstly, for each decade (e.g., 2010-2020, 1980-1989, etc.), you need to find the event with the greatest number of medals for each country, i.e., calculate the total medal count by summing gold, silver, and bronze medals. Then, within each decade, sort the events by their total medal count in descending order, and output the medal counts of the top 3 events (Note: The decades should be listed in descending order, and the medal counts for the top 3 events within each decade should be also sorted in descending order).

  The partial results of Task 2.3 are similar to the following screenshot (note that the screenshot is for formatting reference only; actual results may vary).



Task 3: MapReduce Flowcharts (20 marks)

For the three MapReduce programs in Task 2, create a 3-4 page Word or PDF document that includes flowcharts to illustrate the process of each MapReduce program. You can use specific data examples to clarify the processes. Note that your flowcharts should be consistent with your code. You can refer to the following example provided in the Week 5 lecture notes as an example.



6 steps: input, split, map, shuffle, reduce, output

To create your diagrams, we recommend using the online diagramming tool https://miro.com/. If you choose to use online tools, you can include the diagram link in your document.

Programming Environment:

You are required to use Python to complete all code-related parts of this assignment. The use of other programming languages is strictly prohibited and will result in a loss of all marks.

For Task 1.2 and Tasks 2.1-2.3, you are required to use only MapReduce model to complete the corresponding tasks. The use of any additional Python programs is not allowed. Even if correct results are obtained, failing to use MapReduce will result in 0 mark.

Submission:

Submit a zip file named 'FirstName_LastName_Assignment1.zip' via iLearn. The submission should include the following items:

- Source code for Task 1: 'task1_1.py' and 'task1_2.py'.
- Source code for Task 2: 'task2_1.py', 'task2_2.py', and 'task2_3.py'.
- Output files for Task 1: 'athletes.txt' and 'output1_2.txt'.
- Output files for Task 2: 'output2_1.txt', 'output2_2.txt', and 'output2_3.txt'.
- Flowchart documentation for MapReduce programs in Tasks 2.