School of Computer Science

COMP47470

Project 1
Intro to Big Data (CLI, data management and MapReduce)

Teaching Assistant:	Vsevolods Caka
Coordinator:	Dr Anthony Ventresque
Date:	Wednesday 19th February, 2020
Total Number of Pages:	4

General Instructions

- You are encouraged to collaborate with your peers on this project, but all written work must be your own. In particular we expect you to be able to explain every aspect of your solutions if asked.
- We ask you to hand in an archive (zip or tar.gz) of your solution: code/scripts, README.txt file describing how to run your programs, a short pdf report of your work (no need to include your code in it).
- The report should list your answers and should also contain a short introduction and conclusion. The report should also contain (see Section 3 of this document):
  - a section discussing the difference between relational and non relational (NoSQL) systems (1 page max)
  - a short (2 pages max) description of one of the research papers you can find on Brightspace.
- The report should not be longer than 10 pages.
- The breakdown of marks for the project will be as follows:
  - Exercise 1: 25%
  - Exercise 2: 20%
  - Exercise 3: 30%
  - Exercise 4: 25%
- **Due date: 22/03/2020**

1 Bash for (Big) Data Analysis

Download the file crimedata-australia.csv from Brightspace. Answer the following questions using Bash scripts.

1. how many lines of content (no header) is there in the file? (tail, wc)

2. create a bash command (using, among other things, sed and wc) to count the number of columns

3. For a given city (given as a column number, e.g., 10=Sydney), what is the type of crime on top of the crime list (cat, cut, sort, head)?

4. Find the number of crimes for a given city (given as a column number, e.g., 10=Sydney): create a bash script that reads all the rows (see previous question) and sums up the crime values

5. Same question with the average – look at question 1 to get the number of rows & use tr to remove the empty white spaces and get the number

6. Get the city with the lowest average crime. Create a bash scripts that goes through all the cities, compute the average crime rate and keep only the city with the lowest value. Finally display the city and the average number of crimes.

2 Data Management

Download the Players.csv and Teams.csv datasets from Brighspace.

1. Describe in a short paragraph the two datasets.

2. Create a database in MySQL and a few tables to represent the dataset. Give the conceptual design and the relational model associated with your database

3. Populate the database using a Bash script

4. Create MongoDB collections with every line in the CSV files as an entry/document. Populate the collections with the content of the CSV files (using a script).

Now answer the following questions using SQL and MongoDB queries/searches.

1. What player on a team with “ia” in the team name played less than 200 minutes and made more than 100 passes? Return the player surname.

2. Find all players who made more than 20 shots. Return all player information in descending order of shots made.

3. Find the goalkeepers of teams that played more than four games. List the surname of the goalkeeper, the team, and the number of minutes the goalkeeper played.

4. How many players who play on a team with ranking <10 played more than 350 minutes? Return one number in a column named ’superstar’.

5. What is the average number of passes made by forwards? By midfielders? Write one query that gives both values with the corresponding position.
6. Find all pairs of teams who have the same number of goalsFor as each other and the same number of goalsAgainst as each other. Return the teams and numbers of goalsFor and goalsAgainst. Make sure to return each pair only once.
7. Which team has the highest ratio of goalsFor to goalsAgainst?
8. Find all teams whose defenders averaged more than 150 passes. Return the team and average number of passes by defenders, in descending order of average passes.

3 Reflection

In this section, you are asked to

1. compare relational and NoSQL database management models and in particular give your impressions on why one is better than the other and in which context. In particular we would like you to run run performance evaluations of the two models based on some of the queries you wrote in the previous section - and we would like some discussion on your experience writing these queries from a usability perspective. You can use some external content (papers, blogs, etc.) to support some of your ideas. This section should not be more than two pages.
2. write a short report (1 or two pages) on one of the research papers that are available on Brighspace. They are all different (one is older and pre-date the NoSQL era as such, one is less formal etc.) but they all contain some important information about the NoSQL world. The following list of sections is an indication of how to write your paper. Some of the items might not be relevant for all papers, and you might want to add some sections in your report (e.g., evaluate the posterity of a solution etc.). We will have an open mind when reading your report and we just want to see how you analyse a research paper and are able to discuss it - in short there is no one single perfect report, everything that shows you made an effort to understand and focus on the important parts (research methodology, hypotheses, etc.) will be welcome.
   - identify the question/challenge the paper addresses. Explain in your own words what the motivation for the research is.
   - describe briefly the related work, i.e., the other (related) solutions that the authors compare themselves to. Show the limitations of these related solutions
   - Give an outline of the solution proposed by the authors (no need to go into details) showing the main components
   - describe their scientific method: what are the research questions they evaluate, how do they evaluate
   - describe briefly their results
   - give your impression on the idea, what you liked about the paper and whether you see any limitations etc.

4 Hadoop

In this section you will run Hadoop MapReduce jobs on a dataset of taxi trips. Download the Yellow Taxi tripd Records for January 2019 (you can explore other files obviously but there is no need to downlaod and use more than this file). Answer the following questions with MapReduce jobs as much as you can (use counters, etc.) and complement with

1. What is the average number of passengers per trip in general and per day of the week?
2. What is the average trip distance in general and per day of the week?
3. What are the most used payment types? Create an ordered list using Hadoop MapReduce or a Bash script.
4. Create a graph (using the output of a MapReduce job) showing the average number of passengers over the day (per hour). Create a version for work days and another for weekend days.
5. Create a graph showing the average trip distance over the day (per hour). Create a version for work days and another for weekend days.
6. Create a graph showing the average number of passengers over the day (per hour). Create a version for work days and another for weekend days.