Coursework assigned: 3 February 2017.

Coursework deadline: 4:00pm, 17 February 2017.

Late submission deadline (capped at 50%): 4:00pm, 18 February 2017.

Overview: The coursework aims to make you familiar with the following concepts: (i) Big Data characteristics and analytics, (ii) Big Data collection, and (iii) programming using the MapReduce framework.

This coursework is formally assessed and is worth 10% of your final mark.

You will receive feedback as part of the marking of the coursework.

Submission: Include:

(i) A file, **Coursework1.PDF**, containing your answers. For tasks that require writing code, write your code as part of the answer.

(ii) A file, **Coursework1_code.ZIP**, containing, for each program, the code of the program and a file containing the output of applying the program to the required dataset. Name the code and output to indicate the task it corresponds to (e.g., task3.py for the code and task3.out for the output of Task 3).

Evaluation: The maximum number of marks (out of 100) for each task is given in square brackets [] next to each question.

Plagiarism: “Plagiarism is passing off someone else’s work as your own, or submitting a piece of your own work that you have already submitted as part of a different programme, module or at a different institution. The penalties for plagiarising by the College can be severe. Uploading work to KEATS is regarded by the Department as a statement by the student concerned, confirming that the work has not been plagiarised.”

Late submission: "If you are submitting your coursework after the deadline, you must submit an Extension Request Form to your Programme Administrator, with evidence to justify why you have not submitted on time. If you do not do this or your reasons are not acceptable, your coursework may be given a mark of zero."

Task 1. Big Data characteristics

Choose one application domain (e.g., healthcare, transportation, finance, marketing, web).

(a) Describe why the data in this domain typically possess each of the characteristics of Big Data (i.e., 5Vs). [10]

(b) Describe the challenges entailed by each characteristic of 1(a). [10]

Task 2. Big data collection using Apache Sqoop.

(a) Discuss what happens when the following command is executed:

scoop import --connect jdbc:mysql://localhost/hadoop --username U --password P --table adult -m8 --columns "age, gender"

Your answer should explain step by step how the database table, client, and MapReduce cluster interact during the execution of the command. [6]

(b) Describe three features of Apache Sqoop that help import data into a distributed file system efficiently. [9]

Task 3. Top-3 frequent values in MapReduce.

Download the Adult dataset (file adult.data) from

https://archive.ics.uci.edu/ml/machine-learning-databases/adult/

Write a Python program based on the MapReduce framework, using mrjob, which outputs the 3 most frequent values in the Age attribute of the Adult dataset and their frequency. That is, the three values in Age that appear the largest number of times in the dataset and the number of times each value appears. In addition, provide the output file (result) from applying the program to the Adult dataset. [15]

Task 4. Join in MapReduce.

Download the datasets id_age_occ.csv and id_educ_marital.csv from KEATS.

(a) Write pseudocode for the map and reduce functions to perform a join between these two datasets. [5]

(b) Write a Python program program based on the MapReduce framework, using mrjob, which performs a join between these two datasets. [10]

Example input:

(i) sample of id_age_occ.csv

1, 39, State-gov

2, 50, Self-emp-not-inc

3, 38, Private

4, 53, Private

(ii) sample of id_educ_marital.csv

1, Bachelors, Never-married

2, Bachelors, Married-civ-spouse

3, HS-grad, Divorced

4, 11th, Married-civ-spouse

Example output:

"1" [["39", " State-gov"], ["Bachelors", "Never-married"]]

"2" [["50", " Self-emp-not-inc"], ["Bachelors", "Married-civ-spouse"]]

"3" [["38", " Private"], ["HS-grad", "Divorced"]]

"4" [["53", " Private"], ["11th", "Married-civ-spouse"]]

(c) Execute the program you wrote in (b) on Amazon EMR with 4 mappers and 4 reducers. Show the command to execute the program and provide a file with the output. [10]

Task 5. Inverted index in MapReduce.

An inverted index is a data structure commonly used to map symbols into their location. Many search engines utilize this data structure to efficiently process user queries. In this task, you will implement a simple inverted index and apply it to a web dataset coming from msn.com. Download the dataset msnbc.seq from KEATS. Each record (line) in the dataset contains one or more symbols which occur one or more times.

Write a Python program based on the MapReduce framework, using mrjob. Given the dataset, your program must output each symbol of the dataset along with a set of line ids in which the symbol appears.

Example input:

1 1

2

3 2 2 4 2 2 2 3 3

5

1

6

1 1

6

6 7 7 7 6 6 8 8 8 8

6 9 4 4 4 10 3 10 5 10 4 4 4

Example output:

"1" [1, 5, 7]

"10" [10]

"2" [2, 3]

"3" [10, 3]

"4" [10, 3]

"5" [10, 4]

"6" [10, 6, 8, 9]

"7" [9]

"8" [9]

"9" [10]

Observe, for example, that the symbol 1 is mapped to the set [1,5,7], because 1 appears in the first, fifth, and seventh line of the example input. Note that we are interested in the set of lines in which a symbol appears. Therefore, multiple occurrences of a symbol in a line do not affect the output. In addition, provide the output file (result) from applying the program to the Adult dataset. [25]

[END of Coursework 1]