Coursework assigned: 8 February 2019.

Coursework deadline: 4:00pm, 22 February 2019.

Feedback: 22 March 2019.

Late submission deadline (capped at 50%): 4:00pm, 23 February 2019.

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 questions that require writing code, write your code as part of the answer. For questions that require output of a program provide the output (or a small part- first few lines – if the output is too large).

(ii) A file, **Coursework1_code.ZIP**, containing, for each program, the code of the program (.py file) 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." Please check your handbook and contact your personal tutor for more information.

Task 1. Big Data characteristics

(a) Can data from the transportation domain be classified as Big Data? Justify your answer by examining whether such data possess each of the characteristics of Big Data (i.e., 5Vs). [10]

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

Note: Refer to Lecture 1 for discussion of the characteristics and an example of Big Data from a different domain (game industry).

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. [10]

(b) Describe three features of Apache Sqoop that help import data into a distributed file system efficiently. [10]

Note: Refer to Lecture 2 for the process and also to the pdf of Lab 2 for details.

Task 3. MapReduce Combiners

(a) Provide an example of a function ***f*** that cannot be used in a combiner. Your example should explain why the function ***f*** cannot be used by referring to suitable properties of the function. [5]

(b) Write a program task3_b1.py using mrjob, which applies ***f*** to a small input file (5-10 lines of text would suffice) without using a combiner. Comment your code appropriately to explain what each step does. [10]

(c) Write another program program task3_b2.py using mrjob, which applies ***f*** to the same input file and it uses a combiner. Comment your code appropriately to explain what each step does. [10]

(d) Provide the output of both programs (in files program_task3_b1.out and program_task3_b2.out and in your report) and explain why the output of the second program is incorrect**.** [10].

Task 4. Join in MapReduce.

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

Write a Python program program_task4.py based on the MapReduce framework, using mrjob, which performs a join between these two datasets. Please comment your code appropriately to explain what each step does. Provide the output of that program on the datasets in a file program_task4.out. Your report should also contain a small part of program_task4.out. [25]

Notes:

- Your are asked to join two files. Solutions that generalize to more than two files are not needed.
- You can use two input files in your mrjob program as shown in the following example, which creates two input files and then applies wordcount.py to the files. Wordcount.py is a program the measures how many times each word appears in the files.

  [cloudera@quickstart Desktop]$ cat file1.txt
  one two three

  [cloudera@quickstart Desktop]$ cat file2.txt
  one four five

  [cloudera@quickstart Desktop]$python3 wordcount.py file1.txt file2.txt
  “five” 1
  “four” 1
  “one” 2
  “three” 1
  “two’ 1

- The join attribute is the id (it is included in the files and it is not something you need to calculate). You can see its function in the join from the example output on the next page. I expect to see the example output, based on the example input.

- The order of the attributes must be maintained. That is, every record in the joined table has id, then the attributes age and occupation of id_age_occ.csv and finally the attributes education and marital status of id_educ_marital.csv. Please see the example output.

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"]]

 [END of Coursework 1]