WQD7007 Big Data Management

Lab Test (Hive/Pig)

Instructions: Work in a pair of two per group. Answer all the questions. Explain adequately how you get the answers that can include: 1) codes or process used; and 2) print screens and/or related files that can justify your outcome. You can create a readme file (either in .txt or .docx) to explain your answer. Zip all the files and submit to the spectrum "Lab test" submission page. Elaborate how you split your tasks with your team member.

Part 1: (7 marks)

1. Download two sets of data (in .csv) about student's score for their assessments from the Spectrum page. Please refer to Appendix 1 (at the end of the document) on which dataset you should download. Combine two sets of data from both members. Import the downloaded dataset to HDFS. (2 marks)

2. By using Hive or Pig, identify:
   a. 10 students that have the highest MT score.
   b. 3 male and 3 female students that have the lowest 1FE score.
   c. 5 students that shows the biggest difference from LT (LTP1, LTN1, LTS1, LTP2, LTN2, LT3) score to FE (1FE, 2FE, 3FE, 4FE, 5FE, 6FE) score.

   (6 marks)

   *Explain how you fix the datasets, if there are any issues with it.

Part 2: (8 marks)

1. Import two sets of text (in .txt) from the Spectrum page to HDFS. (1 mark)

2. Run a word count program using Hadoop MapReduce concept to count the word occurrence of the imported texts as in step 1. Save the results in HDFS. (2 marks)

3. Import the result from step 2 to Hive or Pig. Display:
   a. 5 words with 10 counts in descending alphabetical order.
   b. 10 words with highest counts in ascending alphabetical order.

   (2 marks)

4. By applying some preprocessing steps of your choice, clean the text imported in Question 1. Then, repeat the steps in Question 2 and 3. Compare both sets of results.

   (2 marks)

Appendix 1: Please download respective datasets (in .csv format, e.g. Set1.csv) based on the name list below:

No	Name	Set
1.	FARAH AMIRAH BINTI MOHAMAD RAFI	1
2.	ILI NURIZZATI BINTI HANIM	2
3.	TAN JIA YUE	3
4.	LEE JIH SHIAN	4
5.	LEE YEN WEN	5
6.	EMILY SIA ZI XUAN	6
7.	ANG QI KANG	7
8.	SARVINNAH KAJANDREN	8
9.	LEE GUANG SHEN	9
10.	DARVHIND MAGAYNDRAN	10
11.	ALOYSIUS PALLIS GERAD	1
12.	ZE YING TAN	2
13.	LIM KHAI FUNG	3
14.	VIJAYRAJ KALACHILVAM	4
15.	HAN DIE	5
16.	PRIYADARSHINI NAIR AP MUNIANDY	6
17.	DANIAL MIRZA BIN MADRAWI	7
18.	OOI MEI LING	8
19.	TIUN WAI REN	9
20.	MANG YU JIE	10
21.	WAI HONG WOO	1
22.	KHAIRUN NADZIRAH ABDUL KARIM	2
23.	RONGXUAN LEI	3
24.	MUHAMMAD SYAHIRUL KHALIQ MOHAMED AIDI SHAHRIZ	4
25.	LEONG MIN QI	5
26.	CHEN YI	6
27.	ZHENG KEXIN	7
28.	TSU HIAO PING	8
29.	VAISHNAVI A/P YUDEYELJODI	9
30.	ZHANG YU	10
31.	JINGKE TAN	1
32.	VIKTOR ANCHUTIN	2
33.	CHI HSIANG CHEW	3
34.	YUAN WEI KAM	4
35.	TAN HOOI YI	5
36.	HU LIANGLIANG	6
37.	ELAINE LI	7
38.	YUYANG SU	8
39.	YIN YEE LEE	9
40.	SHU ERN LEE	10
41.	JIE YENG LEE	1
42.	NUR IZZAH ATHIRAH ALZAHRI	2
43.	YASHWINIE SELIYAN	3
44.	PHUA LIH JANG	4
45.	WOO YONG SHEN	5
46.	LIU XIAOMENG	6
47.	CHENG WEI NG	7
48.	IZZAH ATHIRAH MOHAMAD RADZI	8
49.	YING MING TANG	9
50.	PUVANESWARI A/P POOBALAN	10
51.	XING ZHAO CHUA	1
52.	KAR HONG SAM	2
53.	CHONG HUN YEE	3
54.	SADMAN CHOWDHURY	4