COMP4057 Distributed and Cloud Computing

Assignment 1

Due date: 28 March, 2023 (18:00)

Write MapReduce programs that run on the Big Data Cluster to solve the following problems.

Hints: Labs 0, 1, and 2 for a warm-up exercise of programming on the cluster. See Appendix A (at the end of this manual) to prepare for your deliverables. A possible marking scheme is presented in Appendix B.

1. In the file vaccination-rates-over-time-by-age.txt on Moodle, each row has the following format:

   date, age-group, gender, Sinovac 1st dose, Sinovac 2nd dose, Sinovac 3rd dose, Sinovac 4th dose, Sinovac 5th dose, Sinovac 6th dose, BioNTech 1st dose, BioNTech 2nd dose, BioNTech 3rd dose, BioNTech 4th dose, BioNTech 5th dose, BioNTech 6th dose

   A row records the number of a gender in an age group who took the *i*th dose of Sinovac and the *j*th BioNTech vaccines on the date, where $i = 1,\ldots, 6$ and $j = 1,\ldots, 6$.

   For example, consider "2021-02-22,30-39,M,1,0,0,0,0,0,0,0,0,0,0". It states that on 2021-02-22, one male of the age 30-39 took the 1st dose of Sinovac vaccine.

   Because of the data entry problem, the age group "12-19" was transcribed as "19-Dec" in the txt file; some age groups "80 and above" were transcribed as "80+". Write a MapReduce program to replace all "19-Dec" with "12-19" and replace all "80+" with "80 and above" in vaccination-rates-over-time-by-age.txt. Output the result to a new file named vaccination-rates-over-time-by-age-v2.txt for the subsequent questions.

   Hints: Use a Hadoop command to output the content of an output directory to the screen and use a Unix command to redirect the screen to a file:

   ```
   hadoop fs –cat output-dir/* > vaccination-rates-over-time-by-age-v2.txt
   ```

   Then, you will transfer vaccination-rates-over-time-by-age-v2.txt to HDFS via the copyFromLocal command for other questions.

   (10 marks)

2. It is generally known that the fifth wave of COVID-19 in Hong Kong is from 31 December 2021 to 23 March 2022. Compute the total number of the *i*th dose of Sinovac ($i = 1,\ldots, 6$) for each age group two months before the fifth wave, i.e., **from 2021-10-31 to 2021-12-30**. Similarly, compute the total number of the *j*th dose of BioNTech ($j = 1,\ldots, 6$) for each age group two months before the fifth wave.

(30 marks)

Additional question: Why were there so few people taking the vaccinations?

(0 marks)

3. To study the vaccination around the fifth wave of COVID-19 in Hong Kong, compute the total number of people in each age group who received Sinovac (regardless of the number of doses) in 31-12-2021 to 23-03-2022, respectively. Similarly, compute the total for BioNTech in those months.

   Which age group received the most number of Sinovac (regardless of the number of doses) during that period?

   Which age group received the most number of BioNTech (regardless of the number of doses) during that period?

   (20 marks)

   Additional question: From the newspapers, which age group attributed to exceedingly high death numbers in the fifth wave of COVID-19 in Hong Kong?

   (0 marks)

4. Compute the total number of vaccinations (either Sinovac or BioNTech) taken (regardless of the dose) from 2021-2-22 to 2021-12-30.

   (20 marks)

   Additional question: The Hong Kong population in 2021 was approximately 7.5 million. What is the average number of vaccinations per Hong Kong person taken during 2021-2-22 and 2021-12-30?

   (0 marks)

5. Compute the difference in the number of vaccinations between each month in 2021 and 2022 (except for January and February). For example, suppose the total number of vaccinations in 04-2021 is 100,000 and that in 04-2022 is 80,000. The difference is -20,000.

   (20 marks)

   Additional question: After the fifth wave (Mar 2022), did Hong Kong people continue to take vaccinations?

   (0 marks)

Appendix A. Deliverables

1. For each question, you must submit the java source, jar file, and a README containing the instructions/commands to run your program.

2. A brief pdf report that contains (i) screenshots of the input and output of the program and (ii) an explanation of the major steps of your program.

3. The *output* files of your program.

4. Put ALL your deliverables into ONE zip file to Moodle on or before the deadline.

Appendix B. Marking scheme of each question:

Completeness of the deliverables	10%
Correct outputs	10%
Runnable jar	10%
Clear mapper logic	30%
Clear reducer logic	30%
Efficient key-value pair	10%

**Smart solutions can be awarded ten marks bonus.*