COMP4057 Distributed and Cloud Computing

Assignment 1

Due date: 17 March, 2023 (18:00)

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

Hints: Practice Labs 1 and 2 for a warm-up exercise of programming on the cluster. See Appendix A 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 of an age group taken 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, there is 1 male who take the 1st dose of Sinovac vaccine.

   Because of data entry problem, the age group "12-19" is transcribed as "19-Dec" in the txt file. Write a mapreduce program to replace all "19-Dec" with "12-19" 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
   ```

   (10 marks)

2. Compute the total number of the *i*th dose of Sinovac ($i = 1,\ldots, 6$) for each age group during 2021-02-22 to 2023-01-15. Similarly, compute the total number of the *j*th dose of BioNTech for each age group during that period.

   (20 marks)

3. To study the vaccination around the fifth wave of Covid-19 in Hong Kong, compute the total number of people who received Sinovac (regardless of the number of dose) in 12-2021, 01-2022, 02-2022 and 03-2022, respectively. Similarly, compute the total for BioNTech in those months.

(20 marks)

4. Compute the numbers of days in 12-2021, 01-2022, 02-2022 and 03-2022, respectively, that have vaccination records.

(20 marks)

5. Compute the differences of the total number of vaccinations between each two consecutive months. For example, suppose the total number of vaccinations of 03-2021 is 484,400 and that of 04-2021 is 908,693. The difference is 424,293.

(30 marks)

Appendix A. Deliverables

1. For each question, you are required to submit the java source, jar file, and a README that contains the instructions/commands to run your program.

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

3. The *output* file 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 10 marks bonus.*