COSC 2637/2633 Big Data Processing
Assignment 3 – Matrix Operation

Assessment Type	– Individual assignment. – Submit online via Canvas → Assignment 3. – Marks awarded for meeting requirements as closely as possible. – Clarifications/updates may be made via announcements or relevant discussion forums.
Due Date	23:59, 5 Oct, 2022
Marks	25

Overview

Write an advanced MapReduce program which develops your skills to solve complex problems on Hadoop execution platform and evaluate the performance in the context of various computing resources and data sizes.

Learning Outcomes

The key course learning outcomes are:

- CLO 1: model and implement efficient big data solutions for various application areas using appropriately selected algorithms and data structures.
- CLO 2: analyse methods and algorithms, to compare and evaluate them with respect to time and space requirements and make appropriate design choices when solving real-world problems.
- CLO 3: motivate and explain trade-offs in big data processing technique design and analysis in written and oral form.
- CLO 4: explain the Big Data Fundamentals, including the evolution of Big Data, the characteristics of Big Data and the challenges introduced.
- CLO 6: apply the novel architectures and platforms introduced for Big data, i.e., Hadoop, MapReduce and Spark.

Assessment Details

Given three matrixes $\mathbf{M}$ $(\boldsymbol{m} \times \mathbf{k})$, $\mathbf{N}$ $(\boldsymbol{k} \times \boldsymbol{n})$ and $\mathbf{X}$ $(\mathbf{m} \times \boldsymbol{n})$, matrix operation $\mathbf{X} - \boldsymbol{MN}$ is a fundamental problem in many machine learning algorithms such as collaborative filtering. $\boldsymbol{MN}$ is the matrix multiplication and $-$ is the matrix subtraction. Here is an example:

$$\boldsymbol{M} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \quad \boldsymbol{N} = \begin{bmatrix} 7 & 8 \\ 9 & 10 \\ 11 & 12 \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 120 & 86 \\ 140 & 210 \end{bmatrix}$$

$$\boldsymbol{MN} = \begin{bmatrix} 1\times7+2\times9+3\times11 & 1\times8+2\times10+3\times12 \\ 4\times7+5\times9+6\times11 & 4\times8+5\times10+6\times12 \end{bmatrix} = \begin{bmatrix} 58 & 64 \\ 139 & 154 \end{bmatrix}$$

$$\boldsymbol{X} - \boldsymbol{MN} = \begin{bmatrix} 120 & 86 \\ 140 & 210 \end{bmatrix} - \begin{bmatrix} 58 & 64 \\ 139 & 154 \end{bmatrix} = \begin{bmatrix} 62 & 22 \\ 1 & 56 \end{bmatrix}$$

Task 1 – Code Development

Write a MapReduce program with python to implement $\mathbf{X} - \boldsymbol{MN}$. The matrix is supposed to be very large. So, it is not allowed to use any data structure of matrix in your code. You are not allowed to use any python MapReduce library such as mrjob.

As the input, each matrix must be saved in a .txt file:

(1) You must feed the size of each matrix, i.e., the number of rows and the number of columns, as the arguments of mapper.

(2) For each matrix, it should be represented in a txt file, where the 1st column specifies the matrix #, 2nd column specifies the row #, then the row of the matrix.

For example:

M.txt	N.txt	X.txt
1, 0, 1, 2, 3 1, 1, 4, 5, 6	2, 0, 7, 8 2, 1, 9, 10 2, 2, 11, 12	3, 0, 120, 86 3, 1, 140, 210

For the output matrix, it must show "*row# column# value*" for every row and column. For example:

```
0  0  62
0  1  22
1  0   1
1  1  56
```

Note the format of input and output must comply with the requirement strictly. Failure to do so leads to 0 marks of assignment.

Task 2 - Performance Analysis

Use the developed code in Task 1 to conduct a series of tests. For each test, you need create matrices **M**, **N** and **X** of same size, e.g., 6×6. Task 2 ask you to test 5 different sizes including 6×6, 20×20, 50×50, 100×100, and 200×200, respectively. For each test, execute Task 1 on **M**, **N** and **X** with different numbers of reducers (i.e., 1, 3, 6, 9). Report the test results in a PDF file with the following information.

A. For each test, the results should include:

Map input records
Map output records
CPU time spent (ms)

It is a good practice to organize the test results in a table and a line chart. The clear and concise presentation will lead to the higher mark.

B. What is the impact of the matrix size to the performance? Explain your answer based on the test results

C. Can more reducers always benefit the performance? Explain your answer based on the test results.

Submission

Your assignment should follow the requirement below and submit via Canvas > Assignment 3. Assessment declaration: when you submit work electronically, you agree to the assessment declaration:

Format Requirements

Failure to follow the requirements incurs up to 6 marks penalty

- If your student ID is s1234567, then please create a zip file named s1234567_BDP_A3.zip with the following files without sub-folders.
    a. All Python files you have developed.
    b. run.sh: a bash script to run your MapReduce job on the EMR master node.
    c. report.pdf: a PDF file for task 2.
    d. README: a text file that includes your student's name, student ID, and how to run your code.
- Do NOT submit the Hadoop Streaming jar file.
- Do NOT submit the given input files.
- Any path in the shell scripts must be specified as follows:
    - -file ./mapper.py
    - -mapper ./mapper.py
    - -file ./reducer.py
    - -reducer ./reducer.py
    - -input /input
    - -output /output
- Please assume the Hadoop Streaming jar file and all your Python files are in the same folder on the EMR master node.

Functional Requirements

Failure to follow the requirements incurs up to 5 marks penalty

- The code must be well written using good coding style.
- The code must include sufficient comments which can clearly explain the major logic flow of the program.

Academic integrity and plagiarism (standard warning)

Academic integrity is about honest presentation of your academic work. It means acknowledging the work of others while developing your own insights, knowledge and ideas. You should take extreme care that you have:

- Acknowledged words, data, diagrams, models, frameworks and/or ideas of others you have quoted (i.e., directly copied), summarized, paraphrased, discussed or mentioned in your assessment through the appropriate referencing methods
- Provided a reference list of the publication details so your reader can locate the source if necessary. This includes material taken from Internet sites.

If you do not acknowledge the sources of your material, you may be accused of plagiarism because you have passed off the work and ideas of another person without appropriate referencing, as if they were your own.

RMIT University treats plagiarism as a very serious offence constituting misconduct. Plagiarism covers a variety of inappropriate behaviours, including:

- Failure to properly document a source
- Copyright material from the internet or databases
- Collusion between students

For further information on our policies and procedures, please refer to
https://www.rmit.edu.au/students/student-essentials/rights-and-responsibilities/academic-integrity

Marking Guide

- Late submission of the assignment results in penalty of 2 marks for (up to) each 24 hours being late. Submissions more than 5*24 hours late results in zero marks.
- If unexpected circumstances affect your ability to complete the assignment, you can apply for special consideration.
  - Requests for special consideration of within 7*24 hours please can be via emailing the course coordinator directly with supporting evidence.
  - Request for special consideration of more than 7*24 hours must be via the University Special consideration: https://www.rmit.edu.au/students/student-essentials/assessment-and-exams/assessment/special-consideration.

Task 1 - 1 Code Development – mapper	0 marks - cannot run on AWS EMR or - no/unreasonable output or - not follow input format or - >1 major logic error in code	1 mark output incorrect due to 1 major logic error in code	3 marks output incorrect due to >1 minor logic error in code	4 marks output incorrect due to 1 minor logic error in code	5 marks output correct and no code error
Task 1 - 2 Code Development - reducer	0 marks - cannot run on AWS EMR or - no/unreasonable output or - not follow input format or - >1 major logic error in code	1-3 marks output incorrect due to 1 major logic error in code	4-5 marks output incorrect due to >1 minor logic error in code	6 marks output incorrect due to 1 minor logic error in code	7 marks output correct and no code error
Task 1 - 3 Code Development – iteration	0 marks - cannot run on AWS EMR or - no/unreasonable output or - not follow input format or - >1 major logic error in code	1 mark Logic incorrect due to 1 major logic error in code	2-3 marks Logic incorrect due to >1 minor logic error in code	4 marks Logic incorrect due to 1 minor logic error in code	5 marks Logic correct and no code error
Task 2 – A Performance Analysis – test results	0 marks - test results not reported or - < 8.5 marks in Task 1 or - unreasonable test results	1 mark - <50% test results correct or - >50% test results are missing - < 11 marks in Task 1	2 marks >= 50% test results correct	3 marks most test results correct	4 marks all test results correct
Task 2 – B Performance Analysis – analysis report	0 marks - < 8.5 marks in Task 1 - no/unreasonable report “explain in which situation MapReduce is more preferrable”.	1 mark - answer correctly in general but there are incorrect statements in report - < 11 marks in Task 1	1.5 marks answer correctly in general but verbose	2 marks correctly and concisely	
Task 2 – C Performance Analysis – analysis report	0 marks - < 8.5 marks in Task 1 no/unreasonable answer “can more reducers always benefit the performance? Explain your answer”.	1 mark - answer correctly in general but there are incorrect statements in report - < 11 marks in Task 1	1.5 marks answer correctly in general but verbose. e.g., clear, well written, thorough, complete, etc.	2 marks correctly and concisely	
Functional requirement	Failure penalty on functional requirements detailed in specification				
Format requirement	Failure penalty on format requirements detailed in specification				