UNIVERSITY OF RUHUNA

Faculty of Engineering

Assignment 1 - Semester 7: May 2025

Module Name: Cloud Computing Module Number: EC7205

Deadline: - 07th June

[Answer **all questions**. This accounts for 15% marks of the module]

Assignment Title

Large-Scale Data Analysis Using MapReduce

Objective

You will implement a custom MapReduce job using Hadoop to process a real or synthetic dataset and extract meaningful insights.

Team Size

3 members

Assignment Task

1. Choose a dataset

Find a dataset that you can use to address a real-world map-reduced task. You can use:

- Public dataset (ex., tweets, book texts, product reviews, logs)
- Or generate synthetic data with at least 100,000 rows

You are encouraged to use publicly available datasets, such as those found on Kaggle. Preference will be given to assignments that use existing datasets. However, if you choose to generate your dataset, it must be sufficiently complex and realistic. In that case, you must also provide a clear explanation of how the dataset was generated, including the tools or methods used and the reasoning behind its structure.

2. Implement a MapReduce job

You can choose your own task, which can align with the dataset you have chosen. A few of the sample tasks are given as follows:

- Inverted Index: Map each word to the documents it appears in.
- Log Analysis: Extract top IPs or most frequent errors from logs.
- Product Ratings Aggregation: Average product rating from reviews.

- Hashtag Popularity: Count hashtags from tweet data.
- Sales Aggregation: Sum total sales per product, region, or store.
- Temperature Extremes: From weather data, find the hottest/coldest day per year.

Notes: No two teams should select the same dataset and the same task. Each team is supposed to implement a unique solution for a unique problem. Please make sure to update the Google Sheet with the dataset you have selected, along with the relevant dataset details.
https://docs.google.com/spreadsheets/d/1K-gweyxglLec0sQx06uyhtmIXlt-gj0xvD9uw4XFleI/edit?usp=sharing.

You can select any programming language of your choice.

3. Setup environment

Install Hadoop locally or on a cloud platform.

Show evidence of installation: (Ex, screenshot, video).

4. Test and run on real data

Run the job on the full dataset.

Save and submit a log of input/output samples.

5. Interpret the result

Provide a summary of the result (1–2 paragraphs):

What patterns or insights did you discover in the selected data?

Were there any performance or accuracy observations? Give some suggestions on how you can expand your model.

6. Documentation and submission

Include the following in the submission:

- Clear README file with steps to run
- Source code (GitHub or zip)
- Dataset used to execute
- Results summary, approach, and result interpretation
- Screenshots/logs/video as evidence of execution

Mark allocation

Map/Reduce Logic Accuracy: 30
Dataset Appropriateness: 10
Code Quality & Structure: 20
Execution Output Evidence: 10
Results Interpretation: 10
Documentation & Clarity: 10
Bonus for Creativity/Scale: 10
Total: 100