Lab 4 - MapReduce

In this lab exercise, you will be utilizing the MapReduce solution to solve a given task. It is recommended, although not mandatory, to read the MapReduce Paper before beginning this lab.

Requirements

1. Choose one of the tasks mentioned in Section 2.3 of the MapReduce Paper to implement a MapReduce solution. The tasks include **Distributed Grep**, **Count of URL Access Frequency**, **Reverse Web-Link Graph**, **Term-Vector per Host**, **Inverted Index**, or **Distributed Sort**.
2. Generate test data for the selected task and validate the correctness of your implementation. You can compare the output of your MapReduce solution with the result obtained from a brute force solution.
3. It is recommended to use the `pyspark` framework for implementing MapReduce using Python.

Submission

Submit a single report for this lab exercise named `<Student_ID>_lab4.pdf` to Cool. The report should contain the following details:

- Description of the task you are solving.
- Format of your input and output.
- Explanation of your Map and Reduce functions.
- Verification method used to validate the result.

Also, include your code in the report. If the code implementation is short, you can directly paste it into the report. Otherwise, provide a link to your code on platforms like GitHub, Gist, etc.

Note: Generating comprehensive test data for the task is not required. Simple hardcoded test data can be used in your code. Late submissions will not be accepted.