Lab 01	Map Reduce Programming
IT494 Big Data Processing, Autumn '2023, DAIICT, Gandhinagar; pm_jat	

In this lab, you do some hands-on map-reduce programming. You work google colab for exercises here.

All data files used in this lab are placed under a "mr" subdirectory of a shared folder https://drive.google.com/drive/folders/1Q0sy0NlD2nkjmzxuYURQoFt5XRZpcScs

Exercise #1:

Go through the map-reduce program in the colab notebook at https://colab.research.google.com/drive/17nVGLN509DZlB_MAlA-kZS-x-Ps5c51i

Understand all solved problems here. Each problem has one Map and one Reduce function. You need to document the following for both functions in each problem:

1. Input parameters: Key and Value with one-line description about each parameter. Description typically tells what data this parameter carries.
2. Output: Key and Value with one-line description about each output. Description typically tells what data each output emits.

Exercise #2:

Process "employee.csv"

Write down map-reduce programs for performing operations that are equivalent to following SQL statements on given employee data file "employee.csv".

Each record (line) in data file has six values, separated by comma. These are employees `empno, name, department no, salary, state, gender` respectively.

1. Compute department wise total salary.
   ```
   SELECT dno, sum(salary) from employee group by dno;
   ```
2. Compute department wise maximum salary among employees from Massachusetts ('MA').
   ```
   SELECT dno, max(salary) from employee where state='MA' group by dno;
   ```
3. Compute department wise average salary.
   ```
   SELECT dno, avg(salary) from employee group by dno;
   ```
4. List all details of employees that are from dno=5 and salary is greater than 100 thousand.
   ```
   SELECT * from employee where dno = 5 and salary > 100000;
   ```
   This may be implemented as map only program.
5. Compute total number of male and female employees for each department.
   ```
   SELECT dno, gender, count(empno) from employee group by dno, gender;
   ```

Process “web access log”

A web server produces a log file that has entries for every http request it receives to a resource. Input log file is available at “web_access_log.txt”

You can understand about the content of web access log file from https://httpd.apache.org/docs/2.4/logs.html#accesslog

Perform following computation using Map Reduce on this file.

6. Computes following summaries on a Monthly Basis –

   (a) Total number of requests.

   (b) Total download size (in MegaBytes)

   It should output: <Year-Month, Number of Requests, Download Size> for every month like Dec-2016, Jan-2017, and so!

7. Create another Map Reduce program that lists Timestamp, and URL for which http response status has been 404.

Submission Required:

A document in pdf format (all in one pdf file) that contains. Using of markdown or latex for creating PDF would be better.

1. Answers to questions in exercise #1
2. All source code is pasted here. Try having the code that is formatted and colored.
3. Link to “colab notebook” that contains source code and outputs of programs for problems in exercise 2.