Lab 04 Programming `Spark RDD`

[IT488, Big Data Processing, Autumn'23]

Instructor: PM Jat (pm_jat@daiict.ac.in)

In this lab, we do some hands-on "Programming `Spark RDD`".

Do following. Use RDD only and do not use data frame.

1. Project and print (empno, name) of employees that are from state 'TX' from employee.csv
2. Generate List of (empno, name, salary, dep_avg_sal) of employees who have salary > 1.5 times of department average from `employee.csv`
3. Compute state wise count of employees from employee.csv
4. Compute the Standard Deviation of salary
5. Compute Department wise Standard Deviation of Salary
6. List (empno, dno, name, salary, dept_sal_sd) for employee that are having salary > 1.5 times of SD.
7. Compute how much offset each department's average salary from the overall average.
8. Computes monthly summary on `web access` log of Lab01, and compute:
   (a) Total number of requests.
   (b) Total download size (in Mega Bytes).
   It should output: <Year-Month, Number of Requests, Download Size> for every month like Dec-2016, Jan-2017, and so!
9. List `Timestamp`, `URL` of requests in `web access` for which http response status is `404`.

Submission Required:

A document in pdf format (all in one pdf file) that contains the following:

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

Using of markdown or latex for creating PDF would be better.