CS 5433 - Big Data Management
Spring 2022

Spark Programming Assignment #1 (24 Points)

Your source code should run on the Hadoop cluster in the department. Instructions to log in and use Spark are outlined in the document named "Spark Instructions.pdf".

Collaboration Policy:

Any doubts/clarification about the questions should be directed to either the instructor or the TA. Make sure you acknowledge web & other resources that you have used in your work.

Submissions

One submission per group

1. README File for each part [Group_y_README_x]. 'y' is the group number, 'x' is the part number. The readme file will give instructions to run your code and list the relevant files.
2. All the source code you submit should be well commented [Penalty for not commenting adequately 25%]. Commented source code for each part [Group_y__Program_x]. y' is the group number and 'x' is the part number.
3. Report
    a. A maximum of one page per part that describes your approach (2 pages in total)
    b. Up to two pages showing for each question screenshots of results. Include the part number and the figure number as well as a caption that will explain what a figure refers to. For example: 'Figure 2,3' refers to figure 3 in part 2.
4. All the source files zipped as a single zip file [Group_y__Program_x _Code.zip]. Each file should be named Code_x_i where' x' is the part number and 'i' is the i^{th} file for part x.

Datasets

1. The HDFS path `/user/kaggle/kaggle_data/Group_y` where 'y' is the group number, points to the data in the cluster which is used for this assignment.
2. The command to view data: `hdfs dfs -ls /user/kaggle/kaggle_data`

Note: You have read and execute permission.

Individual Deliverables (one per individual) [4 marks]

To be submitted separately by each individual. The file should be named [Group_y_Name_x]. 'y' is the group number, 'x' is the last name.

1. Evaluation: The following table should be completed by each student
   50% of actual evaluation score will be deducted if you do not submit an evaluation of your team members.

Evaluation of team members (do not evaluate yourself)	**Name of team member 1** **……**	**Marks for team member 1**	**Name of team member 2** **……**	**Marks for team member 2**
Communication (out of 50):	Explain reason for marks	(out of 50)	Explain reason for marks	(out of 50)
Contribution (out of 100)	Explain reason for marks	(out of 100)	Explain reason for marks	(out of 100)

2. Summary of your contribution

Summary of your contribution:

Deadline: Friday April 15th, 2022

Submit all your deliverables on Canvas

Task 1:
Using the SQL 'GROUP' clause, each group should accomplish the tasks listed below. [10 marks]

Task 2:
Using Python or the language of your choice in Spark, each group should accomplish the tasks listed below. [10 marks]

Breakdown of marks

Task 1	10
Task 2	10
Evaluation	4
TOTAL	24

Tasks
Use the GROUP construct.

Group 1
Obtain the average numbers of undergraduate students enrolled for each of the 4 semesters

Group 2
Show the average number of bachelor degrees holders by males and females

Group 3
Calculate the average closing price of each company is indicated by its ticker symbol

Group 4
Calculate the average number of institutes or universities in each state

Group 5
Calculate the average house price by location_area

Group 6
Calculate the average TOEFL scores by major

Group 7
Calculate the average number of universities by country

Group 8
Calculate the average prices (or median house value) of houses based on ocean proxmity

Group 9
Calculate the average income by marital status

Note: if you run into cluster related technical problems contact the GTLAs Sankirth and Ankit at sankirth.paladugu@okstate.edu and ankaush@okstate.edu respectively.