

## COSC 3366 – Distributed Databases Using Hadoop

### Project 4 – MapReduce Data Analysis Application

In this assignment you will write a MapReduce program to analyze web advertising data for a company. The input file for this assignment is **Project4.txt** which contains a 7-column comma separated data.

The table below explains the data in each column. Each customer can have multiple transactions.

Column	Type Of Data
1	The ID for the record
2	The timestamp of when the advertisement was shown
3	The location of the advertisement
4	The category of the advertisement
5	The number of clicks on the advertisement
6	The number of sales made from the advertisement
7	The age group in which the advertisement was shown

Write a MapReduce program to read the input file from HDFS and find the average success rate for each category by location. Success rate is calculated by dividing the number of sales made from the advertisement (**column 6**) by the number of clicks on the advertisements (**column 5**) and then multiply by 100.

For example, the average success rate for an Ecommerce advertisement for Austin is calculated as follows:

$$((13 / 39 * 100) + (5 / 281 * 100) + (9 / 30 * 100)) / 3 = 21.70\%$$

The output file in HDFS should look like this.

**Category    Location    Success Rate**

Other analysis can be performed on this data, but for this assignment, you are finding the average success rate for each category by location.

**Write a Driver (named Project4Driver.java), a Mapper (named Project4Mapper.java), and a Reducer (named Project4Reducer.java).** You may include other public non-static classes, you think, will help you complete the assignment.

**Handle all exceptions where necessary including the Driver class.**

**All classes in this project must be public, non-static and not nested in other classes.**

**Every method in your program should be limited to performing a single, well-defined task, and the name of the method should express that task effectively. All methods should be non-static unless it is absolutely necessary for it to be static.**

Compile the program into a .jar file for Hadoop.

Create a directory in HDFS and move the input text file, **Project4.txt**, there.

**Take a screenshot of this process showing the HDFS command** and save the image as JPG or PNG in a file named Project4-1.xxx (where “xxx” should indicate the file format).

Execute the Hadoop jar file.

**Take a screenshot of this process showing the Hadoop command** and save the image as JPG or PNG in a file named Project4-2.xxx (where “xxx” should indicate the file format).

Display the output file in HDFS and verify that it is correct.

**Take a screenshot of this process showing the HDFS command** and save the image as JPG or PNG in a file named Project4-3.xxx (where “xxx” should indicate the file format).

**Take a screenshot of your output** and save the image as JPG or PNG file named Project4-4.xxx (where “xxx” should indicate the file format).

Create a folder named, **<YourLastNameFirstName>\_Project4**. Copy your image files, the \*.java files, and the jar file to the folder. **Zip the folder as \*.zip file and upload it to Blackboard.**

**Before you upload your project to Blackboard:**

- Ensure that your code conforms to the style expectations set out in class and briefly discussed below.
- Make sure your variable names and methods are descriptive and follow standard capitalization conventions.

- Put comments wherever necessary. Comments at the top of each module should include your name, file name, and a description of the module. Comments at the beginning of methods describe what the method does, what the parameters are, and what the return value is. Use comments elsewhere to help your reader follow the flow of your code.
- *Program readability and elegance are as important as correctness.* After you have written your method, read and re-read it to eliminate any redundant lines of code, and to make sure variables and methods names are intuitive and relevant.

Read the assignment very carefully to ensure that you have followed all instructions and satisfied all requirements. **You will not get full credit for this project if it is not written as instructed even if it works as expected.**