Big Data – cs522

Lab 8 – Day 16 (Thursday)

This is a practice lab for Spark and so there's nothing to submit.

Spark Word Count in local mode

1. Need to create a Maven project in Eclipse for spark word count.
2. Spark Word Count Java program (JDK 1.7) is given to you, download it and add it to your project.
3. Your pom.xml file should look like the one which is given to you. It'll add all the required dependencies to your project.
4. Create input folder in your project and add a test input file there on which you'll run the word count.
5. Run the WordCount.java program as Java application.
6. You can check the part file in the output folder after refreshing the project.

Spark Word Count in pseudo-distributed mode

1. Make sure your input folder is present under "/user/cloudera" directory and it has the input file in it.
2. From the path of your project, execute "`mvn package`" command in the terminal. After successful execution, it'll create WordCount.jar file inside the "target" directory of your project structure. Verify this by looking at the project structure in eclipse.
3. Now you need to submit this jar to Spark using the following command. This command must be executed from the path of your project.
   ```
   spark-submit --class "cs522.spark.WordCount.WordCount" --master local[4] target/WordCount.jar
   ```
4. After successful execution, it'll create an output folder in "/user/cloudera" . Check it and verify the part file.