**PROJECT – 2**

**CLOUD COMPUTING**

**Below are the file names and significance:**

*Task1*: Cleaning of data is performed using Spark(filename: *task1-2-spark.py*)

*Task 2*: We take the data from spark as input and perform Map reduce to reduce the values.   
 This task is performed in spark as well

[Filename: mapreduce*- ProjectTask2.java*

Spark*- task1-2-spark.py*

*Task 3:* This task is performed in both java and python

Filename- java- *ProjectTask3.java* //takes input from Map-reduce in task2

Python-*Task3-spark.py* //takes input from Spark task2

**Output files: *outputspark.txt*** ---output from spark file

***outputmp.txt***---output from java file

Please perform below steps to execute the project

Note: We are using input file without header

**Executing the project using spark**

Note: In spark we perform task1 and task2 in same file, but it will generate two different output files

Output:/user/kasiresa/resultsparktask1-Folder which contains output from task1

/user/kasiresa/resultspark - Folder which contains outputs from task2

*>>cleaning the folders to avoid any failures with error files/directory already exists*

*hadoop fs -rm -R /user/kasiresa/resultspark\**

*hadoop fs rm -R /user/kasiresa/output\**

*hadoop fs -rm -R /user/kasiresa/resultsparktask1*

*rm -R /home/kasiresa/resultspark/*

*hadoop fs -rm -R* **/**user/kasiresa/output

*rm -R* **/**home/kasiresa/inputtask

Task 1 and 2:

*>> Running the spark command to perform task1 and task2*

spark-submit --master yarn --deploy-mode client --executor-memory 1g --name zip --conf "spark.app.id=zip" task1-2-spark.py /user/data/nypd/NYPD\_Motor\_Vehicle\_WithOutHeader.txt

This will generate two output folders with results from task 1 and task2

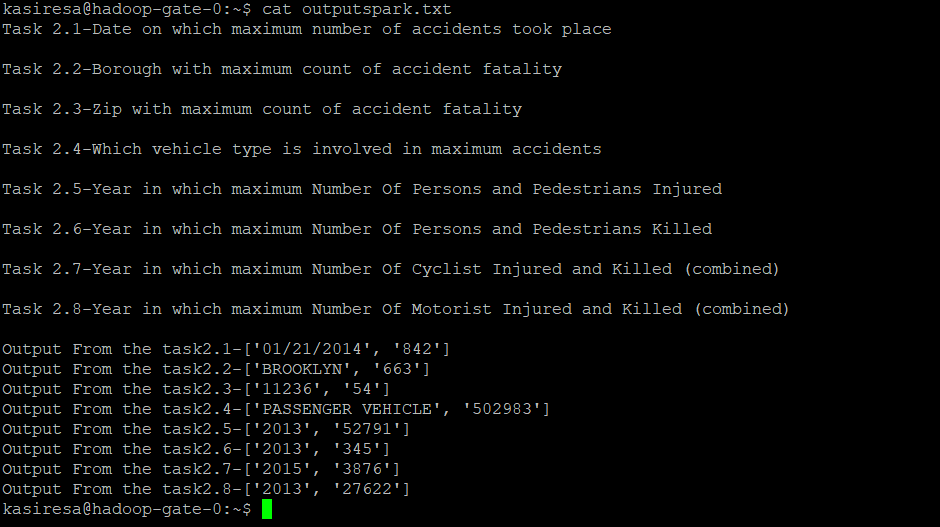
Task 3:

*>> Running python file to perform task3 display the output(this file will get the data from hdfs and copies output to a file and deletes them once action is done)*

python Task3-spark.py

*>>Output is stored in a text file outputspark.txt*

cat outputspark.txt



**Executing the project using Map Reduce in Java**

Note: We perform task2 and task3 using Map reduce in java. We take input to task2 from Spark task1 output

Task 1: Executed in spark and results are used as input for executing Task 2 in Map reduce

Task 2 : Executing below commands to reduce the data using Map reduce

*>>compiling :* **hadoop com.sun.tools.javac.Main ProjectTask2.java -d .**

*>> creating jar file :* **jar -cvf ProjectTask2.jar -C /home/kasiresa/ .**

*>> executing the file :*

**hadoop jar ProjectTask2.jar ProjectTask2 /user/kasiresa/resultsparktask1/\* /user/kasiresa/output1**

Task 3: We need to get output of task2 to feed as input to task3

*>>Getting output from hdfs to local*

**hadoop fs -get /user/kasiresa/output /home/kasiresa/inputtask**

*>>compiling the task 3 file:*

**javac ProjectTask3.java**

*>>executing task3*

**java ProjectTask3**

>>output is saved to a text file *outputmp.txt*

**cat outputmp.txt**

