CS 5433

Bigdata Management
Programming Assignment 3
ReadMe for Task1

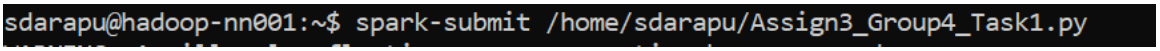
Group 4

Task 1: Data Correction

Steps to run the code:

1) Login to Hadoop cluster and create a python file “Assign3_Group4_Task1.py”.

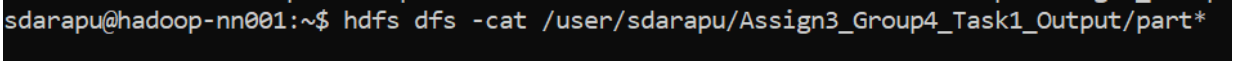
2) Execute the below command to run the file



```
sdarapu@hadoop-nn001:~$ spark-submit /home/sdarapu/Assign3_Group4_Task1.py
```

3) Once file gets executed, first the output will display on the terminal and the output gets saved in the specified path in the program.

4) To view the data stored in specified folder under HDFS, execute the below command.



```
sdarapu@hadoop-nn001:~$ hdfs dfs -cat /user/sdarapu/Assign3_Group4_Task1_Output/part*
```

Note: We have given permissions to username “aanmol” for all our input and output file under username “sdarapu”. So that, you can access all my files. If there is any issue with permissions, then please let us know.