

Big Data Analytics Lab

PMDS507P

Name: **Tufan Kundu** Registration number: **24MDT0184**

Slot: L29+L30

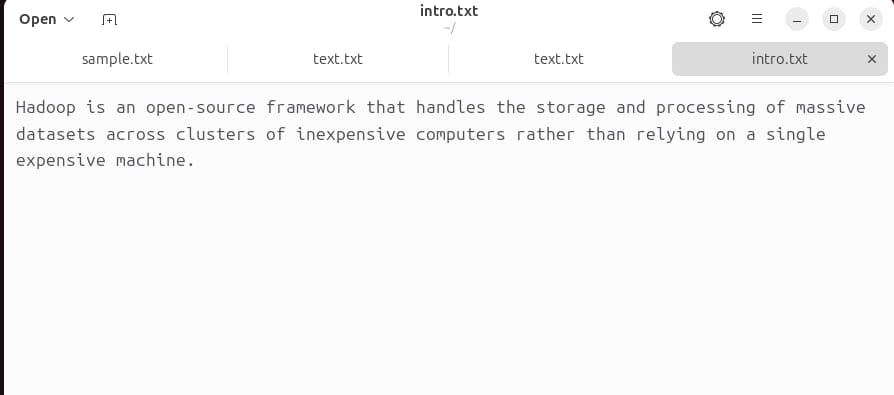
Digital Assignment 3

# 

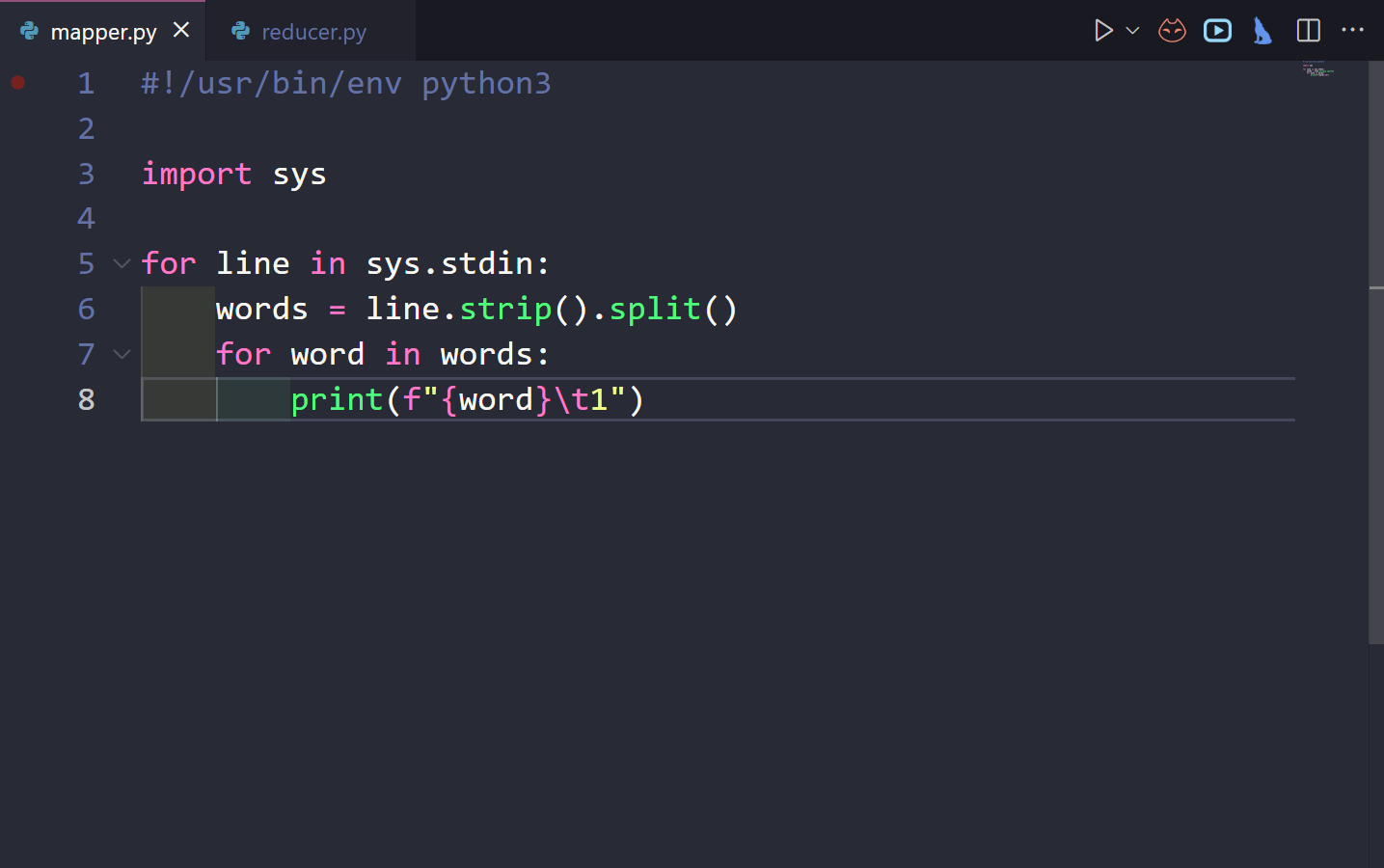
## 

## Explain the steps involved in performing a Word Count program using Hadoop Streaming with Python mapper and reducer scripts.

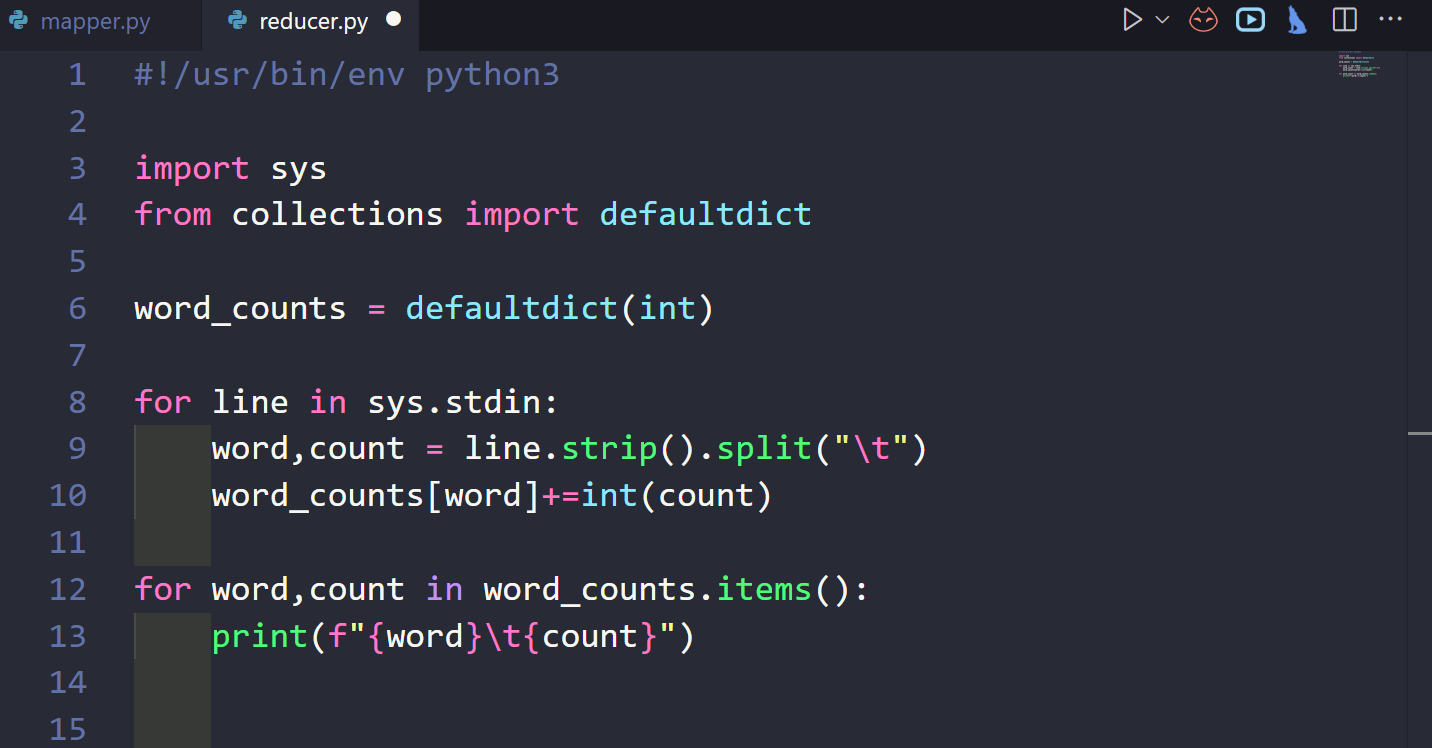
**Step 1:** Preparing the text file



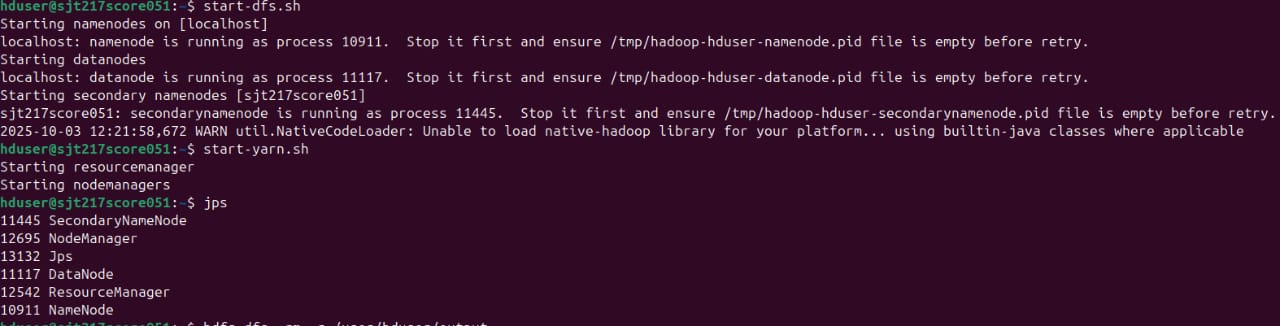
**Step 2**: [Mapper.py](http://mapper.py)



**Step 3**: [Reducer.py](http://reducer.py)



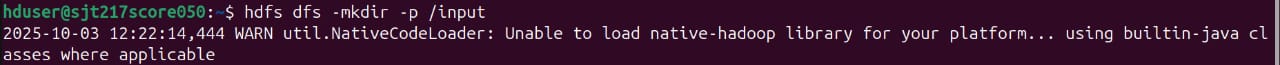
**Step 4**: Start HDFS and YARN and verify with jps



**Step 5:**

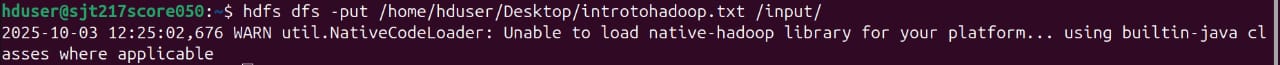
1: Create an Input Directory: An input directory was created in HDFS to store our data file.

***hdfs dfs -mkdir -p /input***



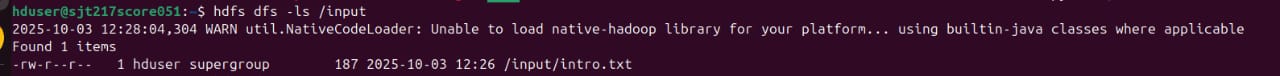
2 - Copy Data to HDFS: The introtohadoop.txt file was copied from the local filesystem to the newly created HDFS directory

***hdfs dfs -put /home/hduser/Desktop/introtohadoop.txt /input/***



3. Verify the File: We listed the contents of the HDFS directory and viewed the file's content to ensure it was uploaded correctly.

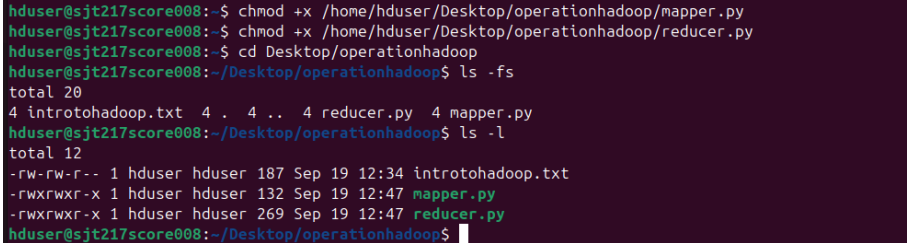
***hdfs dfs -ls /input***



**Step 6:** Making Scripts Executable For Hadoop Streaming. To run our Python scripts, they must have execute permissions. The chmod +x command was used to grant these permissions.

***chmod +x /home/hduser/Desktop/operationhadoop/***[***mapper.py***](http://mapper.py)

***chmod +x /home/hduser/Desktop/operationhadoop/reducer.py***



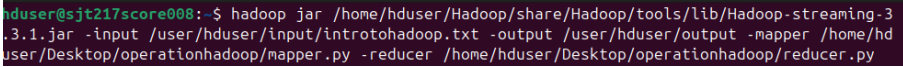
**Step 7:** Displaying the text file

***hdfs dfs -cat /input/intro.txt***



**Step 8:** Running the Hadoop streaming jobs

***hadoop jar /home/hduser/Hadoop/share/Hadoop/tools/lib/Hadoop-streming-3.3.1.jar -input /user/hduser/input/introtohadoop.txt -output /user/hduser/output -mapper /home/hduser/Desktop/operationhadoop/***[***mapper.py***](http://mapper.py) ***-reducer /home/hduser/Desktop/operationhadoop/***[***reducer.py***](http://reducer.py)



**Step 9:** Display the results

***hdfs dfs -cat /output/part-r-00000***

