Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Big Data Analytics Lab
PMDS507P

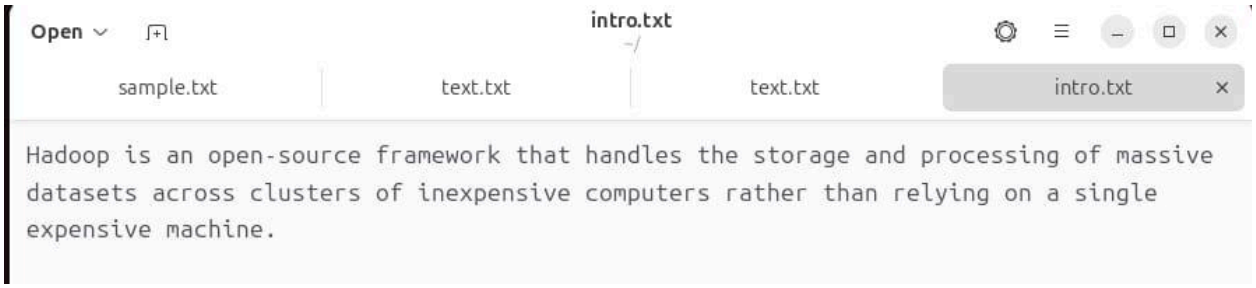
Name: **Tufan Kundu** Registration number: **24MDT0184**

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



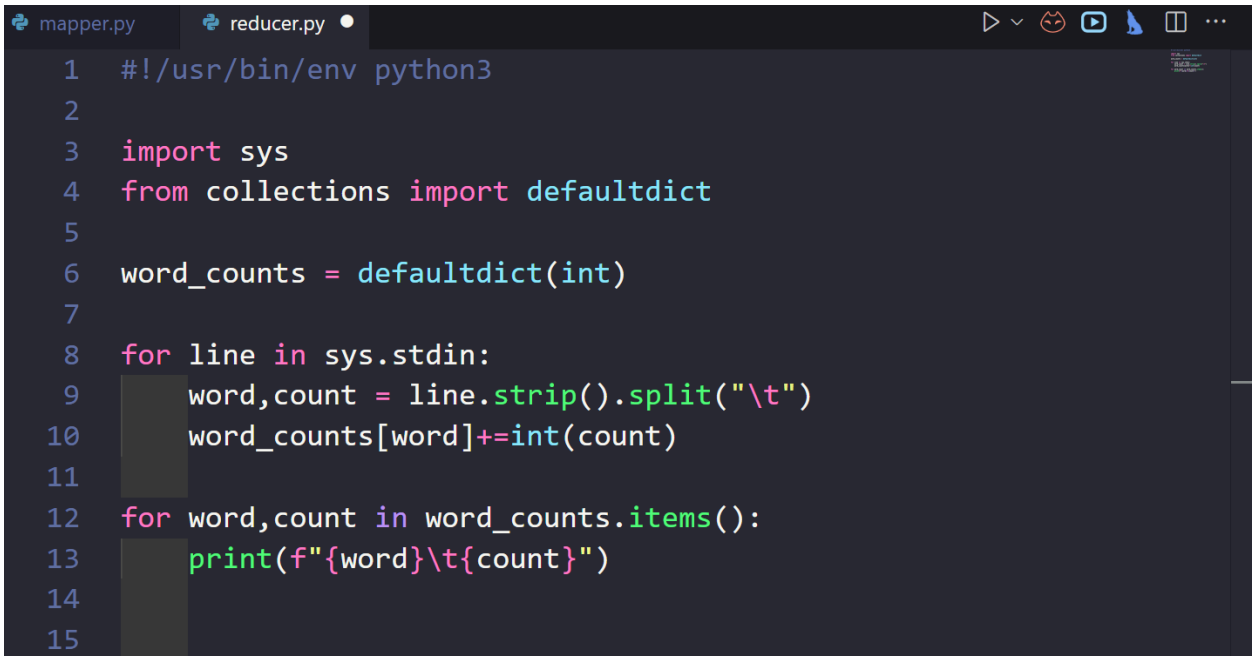
Step 2: Mapper.py

```python
#!/usr/bin/env python3

import sys

for line in sys.stdin:
    words = line.strip().split()
    for word in words:
        print(f"{word}\t1")
```

Step 3: Reducer.py



```python
#!/usr/bin/env python3

import sys
from collections import defaultdict

word_counts = defaultdict(int)

for line in sys.stdin:
    word,count = line.strip().split("\t")
    word_counts[word]+=int(count)

for word,count in word_counts.items():
    print(f"{word}\t{count}")
```

Step 4: Start HDFS and YARN and verify with jps

```
hduser@sjt217score051:~$ start-dfs.sh
Starting namenodes on [localhost]
localhost: namenode is running as process 10911.  Stop it first and ensure /tmp/hadoop-hduser-namenode.pid file is empty before retry.
Starting datanodes
localhost: datanode is running as process 11117.  Stop it first and ensure /tmp/hadoop-hduser-datanode.pid file is empty before retry.
Starting secondary namenodes [sjt217score051]
sjt217score051: secondarynamenode is running as process 11445.  Stop it first and ensure /tmp/hadoop-hduser-secondarynamenode.pid file is empty before retry.
2025-10-03 12:21:58,672 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hduser@sjt217score051:~$ start-yarn.sh
Starting resourcemanager
Starting nodemanagers
hduser@sjt217score051:~$ jps
11445 SecondaryNameNode
12695 NodeManager
13132 Jps
11117 DataNode
12542 ResourceManager
10911 NameNode
```

Step 5:

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

hdfs dfs -mkdir -p /input

```
hduser@sjt217score050:~$ hdfs dfs -mkdir -p /input
2025-10-03 12:22:14,444 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
```

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/

```
hduser@sjt217score050:~$ hdfs dfs -put /home/hduser/Desktop/introtohadoop.txt /input/
2025-10-03 12:25:02,676 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java cl
asses where applicable
```

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

```
hduser@sjt217score051:~$ hdfs dfs -ls /input
2025-10-03 12:28:04,304 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 1 items
-rw-r--r--   1 hduser supergroup        187 2025-10-03 12:26 /input/intro.txt
```

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

```
hduser@sjt217score008:~$ chmod +x /home/hduser/Desktop/operationhadoop/mapper.py
hduser@sjt217score008:~$ chmod +x /home/hduser/Desktop/operationhadoop/reducer.py
hduser@sjt217score008:~$ cd Desktop/operationhadoop
hduser@sjt217score008:~/Desktop/operationhadoop$ ls -fs
total 20
4 introtohadoop.txt  4 .  4 ..  4 reducer.py  4 mapper.py
hduser@sjt217score008:~/Desktop/operationhadoop$ ls -l
total 12
-rw-rw-r-- 1 hduser hduser 187 Sep 19 12:34 introtohadoop.txt
-rwxrwxr-x 1 hduser hduser 132 Sep 19 12:47 mapper.py
-rwxrwxr-x 1 hduser hduser 269 Sep 19 12:47 reducer.py
hduser@sjt217score008:~/Desktop/operationhadoop$
```

Step 7: Displaying the text file

hdfs dfs -cat /input/intro.txt

```
hduser@sjt217score051:~$ hdfs dfs -cat /input/intro.txt
2025-10-03 12:29:14,904 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Hadoop is an open-source framework that handles the storage and processing of massive datasets across clusters of inexpensive computers rather than relying on a single expensive machin
e.
hduser@sjt217score051:~$
```

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 -reducer /home/hduser/Desktop/operationhadoop/reducer.py

```
hduser@sjt217score008:~$ hadoop jar /home/hduser/Hadoop/share/Hadoop/tools/lib/Hadoop-streaming-3
.3.1.jar -input /user/hduser/input/introtohadoop.txt -output /user/hduser/output -mapper /home/hd
user/Desktop/operationhadoop/mapper.py -reducer /home/hduser/Desktop/operationhadoop/reducer.py
```

Step 9: Display the results

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

```
hduser@sjt217score051:~$ hdfs dfs -cat /output/part-r-00000
2025-10-03 13:00:11,043 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Hadoop  1
a       1
across  1
an      1
and     1
clusters        1
computers       1
datasets        1
expensive       1
framework       1
handles 1
inexpensive     1
is      1
machine.        1
massive 1
of      2
on      1
open-source     1
processing      1
rather  1
relying 1
single  1
storage 1
than    1
that    1
the     1
```