



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

Big Data Analytics Lab

PMDS507P

Name: **Tufan Kundu**

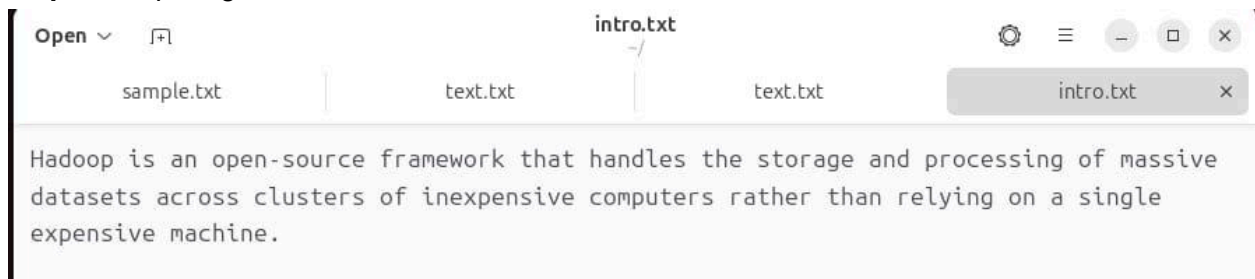
Registration number: **24MDT0184**

Slot: L29+L30

Digital Assignment 4

Explain the procedure to run the Hadoop WordCount program using the in-built mapreduce-examples.jar.

Step 1: Preparing the text file



Step 2: 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 3: Creating the directory

```
hduser@sjt217score010:~$ hdfs dfs -rm -r /user/hduser/output
2025-10-03 12:28:09,381 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
rm: '/user/hduser/output': No such file or directory
hduser@sjt217score010:~$ hdfs dfs -rm -r /user/hduser/input
2025-10-03 12:28:18,104 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
rm: '/user/hduser/input': No such file or directory
hduser@sjt217score010:~$ hdfs dfs -mkdir -p /input
2025-10-03 12:28:36,464 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hduser@sjt217score010:~$ hdfs dfs -put /home/hduser/Desktop/OperationHadoop/IntroToHadoop.txt /input/
2025-10-03 12:29:05,046 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hduser@sjt217score010:~$ hdfs dfs -ls /input/
2025-10-03 12:29:18,173 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 12:29 /input/IntroToHadoop.txt
hduser@sjt217score010:~$
```

Step 4: 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 machine.
```

Step 5: Running wordcount by inbuilt mapreduce functions

hadoop jar
/home/hduser/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.1.jar
wordcount /input /output

```
hduser@sjt217score010:~$ hadoop jar /home/hduser/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-3.3.1.jar wordcount /input /output

2025-10-03 12:51:10,296 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2025-10-03 12:51:11,558 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2025-10-03 12:51:11,781 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2025-10-03 12:51:11,781 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2025-10-03 12:51:12,322 INFO input.FileInputFormat: Total input files to process : 1
2025-10-03 12:51:12,425 INFO mapreduce.JobSubmitter: number of splits:1
2025-10-03 12:51:12,692 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1093694298_0001
2025-10-03 12:51:12,692 INFO mapreduce.JobSubmitter: Executing with tokens: []
2025-10-03 12:51:12,947 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2025-10-03 12:51:12,948 INFO mapreduce.Job: Running job: job_local1093694298_0001
2025-10-03 12:51:12,952 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2025-10-03 12:51:12,970 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-10-03 12:51:12,970 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures:
2025-10-03 12:51:12,973 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2025-10-03 12:51:13,060 INFO mapred.LocalJobRunner: Waiting for map tasks
2025-10-03 12:51:13,061 INFO mapred.LocalJobRunner: Starting task: attempt_local1093694298_0001_m_000000_0
2025-10-03 12:51:13,112 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2025-10-03 12:51:13,113 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures:
2025-10-03 12:51:13,148 INFO mapred.Task: Using ResourceCalculatorProcessTree : [ ]
2025-10-03 12:51:13,163 INFO mapred.MapTask: Processing split: hdfs://localhost:54310/input/IntroToHadoop.txt:0+187
2025-10-03 12:51:13,296 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2025-10-03 12:51:13,296 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2025-10-03 12:51:13,296 INFO mapred.MapTask: soft limit at 83886080
2025-10-03 12:51:13,296 INFO mapred.MapTask: bufstart = 0; bufvoid = 104857600
2025-10-03 12:51:13,296 INFO mapred.MapTask: kvstart = 26214396; length = 6553600
2025-10-03 12:51:13,312 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask$MapOutputBuffer
2025-10-03 12:51:13,838 INFO mapred.LocalJobRunner:
2025-10-03 12:51:13,847 INFO mapred.MapTask: Starting flush of map output
2025-10-03 12:51:13,847 INFO mapred.MapTask: Spilling map output
2025-10-03 12:51:13,847 INFO mapred.MapTask: bufstart = 0; bufend = 295; bufvoid = 104857600
2025-10-03 12:51:13,847 INFO mapred.MapTask: kvstart = 26214396(104857584); kvoid = 26214292(104857168); length = 105/6553600
2025-10-03 12:51:13,925 INFO mapred.MapTask: Finished spill 0
2025-10-03 12:51:13,959 INFO mapreduce.Job: Job job_local1093694298_0001 running in uber mode : false
2025-10-03 12:51:13,960 INFO mapreduce.Job: map 0% reduce 0%
2025-10-03 12:51:13,963 INFO mapred.Task: Task:attempt_local1093694298_0001_m_000000_0 is done. And is in the process of committing
2025-10-03 12:51:13,979 INFO mapred.LocalJobRunner: map
2025-10-03 12:51:13,980 INFO mapred.Task: Task 'attempt_local1093694298_0001_m_000000_0' done.
2025-10-03 12:51:13,996 INFO mapred.Task: Final Counters for attempt_local1093694298_0001_m_000000_0: Counters: 24
```

Step 6: Display the results

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

```
hduser@sjt217score010:~$ hdfs dfs -cat /output/part-r-00000
2025-10-03 12:57:17,083 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
```