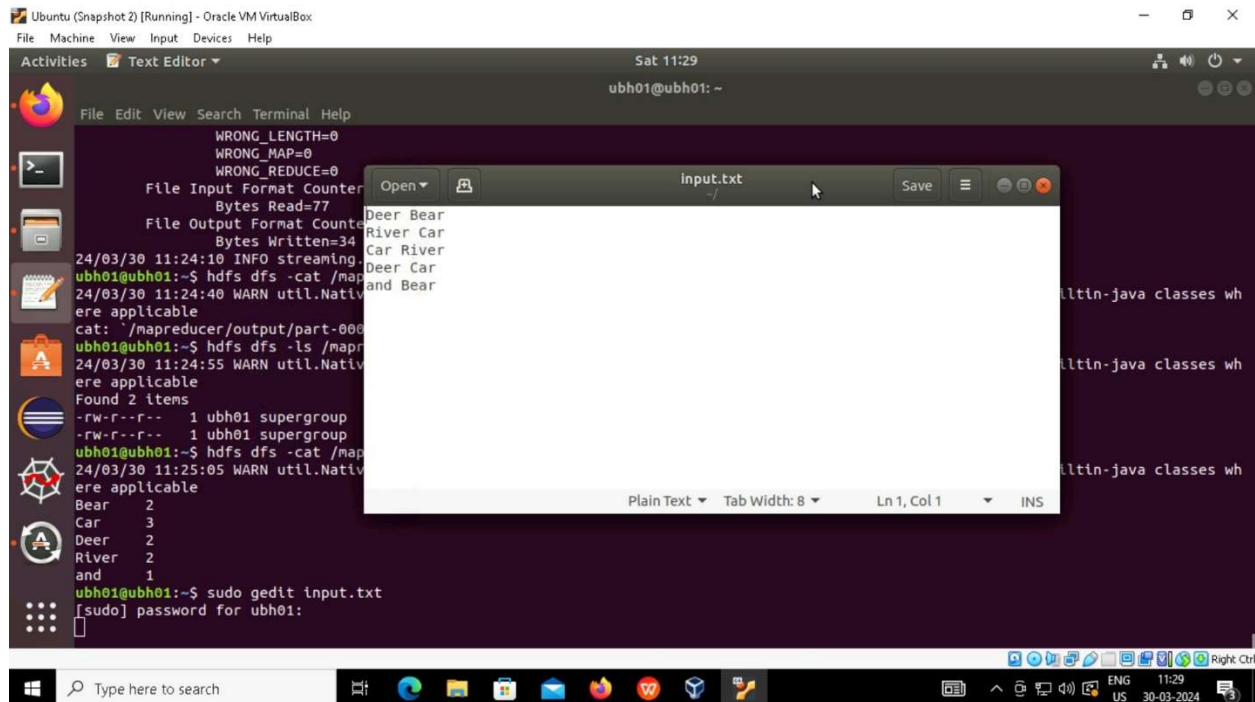


MAP REDUCER ASSIGNMENT -

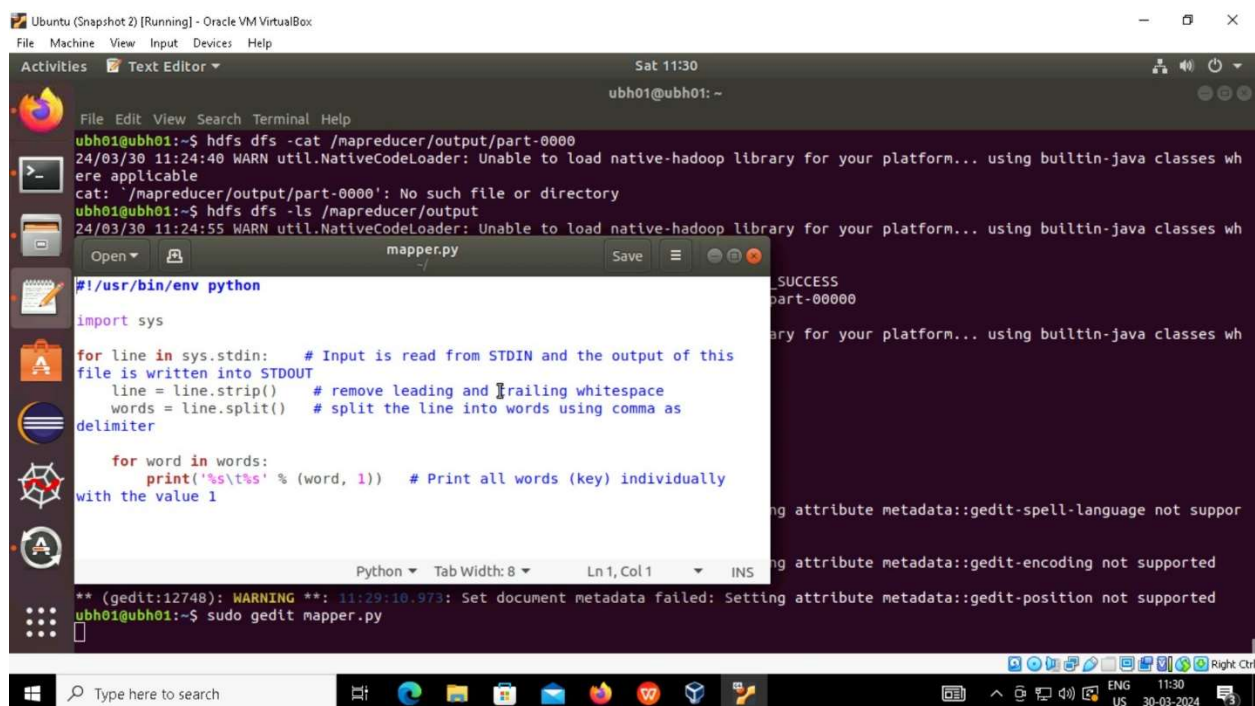
4



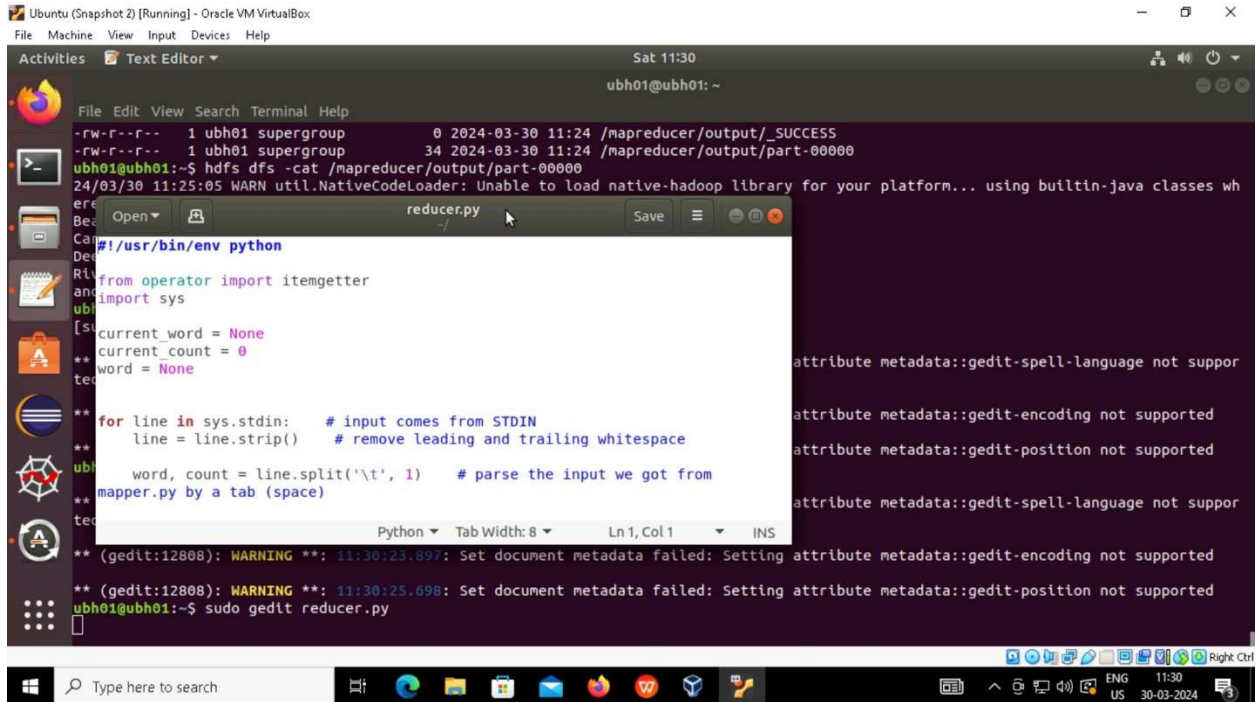
A. JOSHI SATYA VARDAN
CSDAIA24AZ003
EMP ID:2320097



- First create the file by using command **sudo gedit input.txt** file by adding some text as shown like above picture.



- Create another file by using command “**sudo gedit mapper.py**” and written some mapper function code inside that file.



The screenshot shows a Ubuntu (Snapshot 2) [Running] - Oracle VM VirtualBox window. The terminal window displays the following output:

```
-rw-r--r-- 1 ubh01 supergroup 0 2024-03-30 11:24 /mapreducer/output/_SUCCESS
-rw-r--r-- 1 ubh01 supergroup 34 2024-03-30 11:24 /mapreducer/output/part-000000
ubh01@ubh01:~$ hdfs dfs -cat /mapreducer/output/part-000000
24/03/30 11:25:05 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes wh
er
Bec
Ca
Dev
Ri
and
ubh
[st
current_word = None
** current_count = 0
word = None
**
** for line in sys.stdin: # input comes from STDIN
**     line = line.strip() # remove leading and trailing whitespace
**
**     word, count = line.split('\t', 1) # parse the input we got from
** mapper.py by a tab (space)
**
** (gedit:12808): WARNING **: 11:30:23.897: Set document metadata failed: Setting attribute metadata::gedit-encoding not supported
** (gedit:12808): WARNING **: 11:30:25.698: Set document metadata failed: Setting attribute metadata::gedit-position not supported
ubh01@ubh01:~$ sudo gedit reducer.py
```

The text editor window shows the code for `reducer.py`:

```
#!/usr/bin/env python
from operator import itemgetter
import sys

current_word = None
current_count = 0
word = None

for line in sys.stdin: # input comes from STDIN
    line = line.strip() # remove leading and trailing whitespace

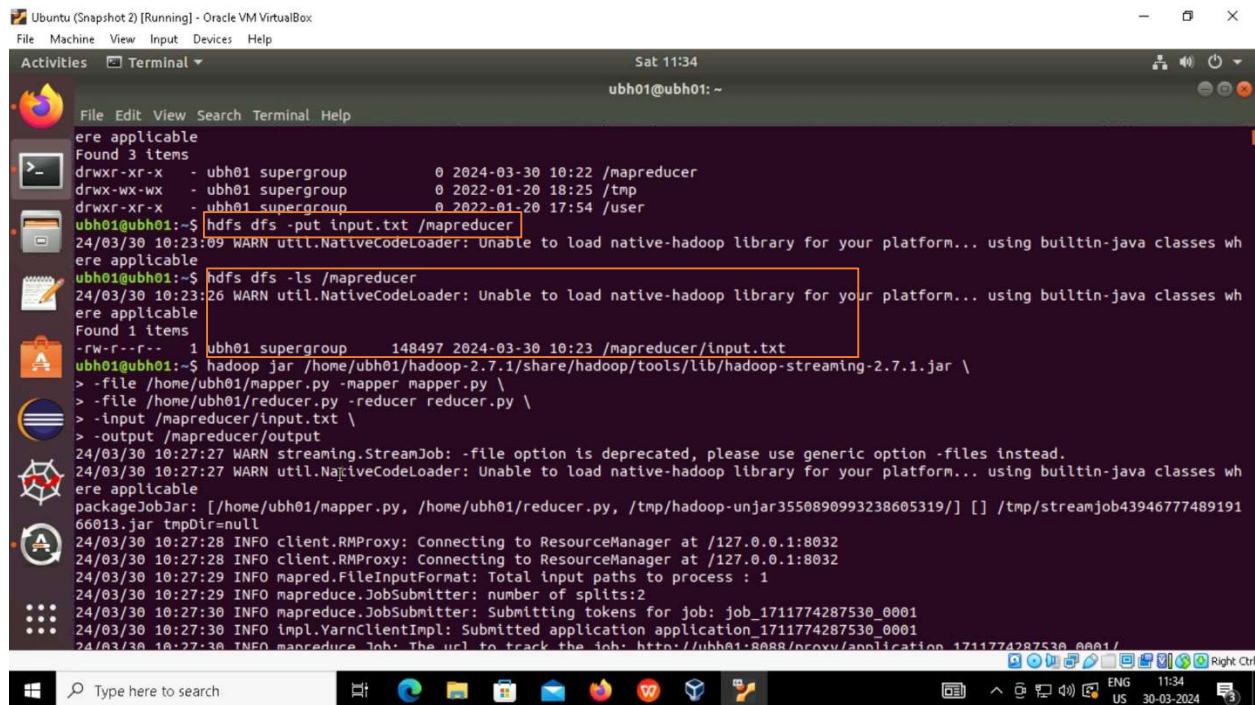
    word, count = line.split('\t', 1) # parse the input we got from
    mapper.py by a tab (space)
```

- Create another file by using command “**sudo gedit reducer.py**” and written some reducer function code inside that file.

The screenshot shows a terminal window titled "Ubuntu (Snapshot 2) [Running] - Oracle VM VirtualBox". The terminal is running a series of commands to test Hadoop MapReduce. The first command is `cat input.txt | python mapper.py`, which outputs a list of animals and their counts: Deer 1, Bear 1, River 1, Car 1, Car 1, River 1, Deer 1, Car 1, and Bear 1. The second command is `cat input.txt | python mapper.py | sort | python reducer.py`, which outputs the sorted results: and 1, Bear 2, Car 3, Deer 2, and River 2. The third command is `hdfs dfs -rmrdir /mapreducer/output`. The fourth command is `hadoop jar /home/ubh01/hadoop-2.7.1/share/hadoop/tools/lib/hadoop-streaming-2.7.1.jar -file /home/ubh01/mapper.py -map per mapper.py -file /home/ubh01/reducer.py -reducer reducer.py -input /mapreducer/input.txt -output /mapreducer/output`. The output of the Hadoop command shows various warnings and information, including the Hadoop version, the input and output paths, and the status of the job. The terminal window also shows the Ubuntu desktop environment with various icons and a taskbar at the bottom.

```
ubh01@ubh01:~$ cat input.txt | python mapper.py
Deer 1
Bear 1
River 1
Car 1
Car 1
River 1
Deer 1
Car 1
Bear 1
ubh01@ubh01:~$ cat input.txt | python mapper.py | sort | python reducer.py
and 1
Bear 2
Car 3
Deer 2
River 2
ubh01@ubh01:~$ hdfs dfs -rmrdir /mapreducer/output
24/03/30 11:10:39 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
^[[ubh01@ubh01:~$ hadoop jar /home/ubh01/hadoop-2.7.1/share/hadoop/tools/lib/hadoop-streaming-2.7.1.jar -file /home/ubh01/mapper.py -map
per mapper.py -file /home/ubh01/reducer.py -reducer reducer.py -input /mapreducer/input.txt -output /mapreducer/output
24/03/30 11:10:46 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
24/03/30 11:10:46 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes wh
ere applicable
packageJobJar: [/home/ubh01/mapper.py, /home/ubh01/reducer.py, /tmp/hadoop-unjar7112785140824316704/] [] /tmp/streamjob69534560369003
5066.jar tmpDir=null
24/03/30 11:10:47 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
24/03/30 11:10:47 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
24/03/30 11:10:48 INFO mapred.FileInputFormat: Total input paths to process : 1
```

- By using command “**cat input.txt | python mapper.py**” and check whether mapper function is working on the given text file in local
- By using command “**cat input.txt | python mapper.py | sort | python reducer.py**” and checking whether the mapper and reducer functions are working on the given text file in local.



The screenshot shows a terminal window titled "Ubuntu (Snapshot 2) [Running] - Oracle VM VirtualBox". The terminal is running a series of commands to set up a Hadoop environment. The first command is `hdfs dfs -ls /mapreducer`, which shows three files: `drwxr-xr-x - ubh01 supergroup 0 2024-03-30 10:22 /mapreducer`, `drwx-wx-wx - ubh01 supergroup 0 2022-01-20 18:25 /tmp`, and `drwxr-xr-x - ubh01 supergroup 0 2022-01-20 17:54 /user`. The second command is `hdfs dfs -put input.txt /mapreducer`, which uploads a file named `input.txt` to the `/mapreducer` directory. The third command is `hdfs dfs -ls /mapreducer`, which shows the file `-rw-r--r-- 1 ubh01 supergroup 148497 2024-03-30 10:23 /mapreducer/input.txt`. The fourth command is `hadoop jar /home/ubh01/hadoop-2.7.1/share/hadoop/tools/lib/hadoop-streaming-2.7.1.jar -file /home/ubh01/mapper.py -mapper mapper.py -file /home/ubh01/reducer.py -reducer reducer.py -input /mapreducer/input.txt -output /mapreducer/output`, which runs a MapReduce job. The output shows the job is submitted and the URL to track the job is `http://ubh01:8088/proxy/application_1711774287530_0001/`.

```
File Edit View Search Terminal Help
Sat 11:34
ubh01@ubh01: ~

ere applicable
Found 3 items
drwxr-xr-x - ubh01 supergroup 0 2024-03-30 10:22 /mapreducer
drwx-wx-wx - ubh01 supergroup 0 2022-01-20 18:25 /tmp
drwxr-xr-x - ubh01 supergroup 0 2022-01-20 17:54 /user
ubh01@ubh01:~$ hdfs dfs -put input.txt /mapreducer
24/03/30 10:23:09 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes wh
ere applicable
ubh01@ubh01:~$ hdfs dfs -ls /mapreducer
24/03/30 10:23:26 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes wh
ere applicable
Found 1 items
-rw-r--r-- 1 ubh01 supergroup 148497 2024-03-30 10:23 /mapreducer/input.txt
ubh01@ubh01:~$ hadoop jar /home/ubh01/hadoop-2.7.1/share/hadoop/tools/lib/hadoop-streaming-2.7.1.jar \
> -file /home/ubh01/mapper.py -mapper mapper.py \
> -file /home/ubh01/reducer.py -reducer reducer.py \
> -input /mapreducer/input.txt \
> -output /mapreducer/output
24/03/30 10:27:27 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
24/03/30 10:27:27 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes wh
ere applicable
packageJobJar: [/home/ubh01/mapper.py, /home/ubh01/reducer.py, /tmp/hadoop-unjar3550890993238605319/] [] /tmp/streamjob43946777489191
66013.jar tmpDir=null
24/03/30 10:27:28 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
24/03/30 10:27:28 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
24/03/30 10:27:29 INFO mapred.FileInputFormat: Total input paths to process : 1
24/03/30 10:27:29 INFO mapreduce.JobSubmitter: number of splits:2
24/03/30 10:27:30 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1711774287530_0001
24/03/30 10:27:30 INFO impl.YarnClientImpl: Submitted application application_1711774287530_0001
24/03/30 10:27:30 INFO mapreduce.Job: The url to track the job: http://ubh01:8088/proxy/application_1711774287530_0001/
```

- By using this command “**hdfs dfs -mkdir /mapreducer**” create a directory in hdfs
- By using this command “**hdfs dsf -put input.txt /mapreducer**” moving the file input.txt from local to mapreducer directory in Hadoop.


```
24/03/30 11:23:44 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
ubh01@ubh01:~$ hadoop jar /home/ubh01/hadoop-2.7.1/share/hadoop/tools/lib/hadoop-streaming-2.7.1.jar -file /home/ubh01/mapper.py -mapper mapper.py -reducer reducer.py -input /mapreducer/input.txt -output /mapreducer/output
24/03/30 11:23:53 WARN streaming.StreamJob: -file option is deprecated, please use generic option -files instead.
24/03/30 11:23:53 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
packageJobJar: [/home/ubh01/mapper.py, /home/ubh01/reducer.py, /tmp/hadoop-unjar9179552531365288731/] [] /tmp/streamjob8148666092968664984.jar tmpDir=null
24/03/30 11:23:54 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
24/03/30 11:23:54 INFO client.RMProxy: Connecting to ResourceManager at /127.0.0.1:8032
24/03/30 11:23:54 INFO mapred.FileInputFormat: Total input paths to process : 1
24/03/30 11:23:55 INFO mapreduce.JobSubmitter: number of splits:2
24/03/30 11:23:55 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1711774287530_0007
24/03/30 11:23:55 INFO impl.YarnClientImpl: Submitted application application_1711774287530_0007
24/03/30 11:23:55 INFO mapreduce.Job: The url to track the job: http://ubh01:8088/proxy/application_1711774287530_0007/
24/03/30 11:23:55 INFO mapreduce.Job: Running job: job_1711774287530_0007
24/03/30 11:24:00 INFO mapreduce.Job: Job job_1711774287530_0007 running in uber mode : false
24/03/30 11:24:00 INFO mapreduce.Job: map 0% reduce 0%
24/03/30 11:24:05 INFO mapreduce.Job: map 100% reduce 0%
24/03/30 11:24:09 INFO mapreduce.Job: map 100% reduce 100%
24/03/30 11:24:10 INFO mapreduce.Job: Job job_1711774287530_0007 completed successfully
24/03/30 11:24:10 INFO mapreduce.Job: Counters: 49
File System Counters
  FILE: Number of bytes read=94
  FILE: Number of bytes written=356587
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=265
```

- By using this command

“ `hadoop jar /home/ubh01/hadoop-2.7.1/share/hadoop/tools/lib/hadoop-streaming-2.7.1.jar -file /home/ubh01/mapper.py -mapper mapper.py -file /home/ubh01/reducer.py -reducer reducer.py -input /mapreducer/input.txt -output /mapreducer/output` “

we are running mapreducer job by using hadoop streaming-2.7.1.jar and checking whether mapper and reducer are working 100% on the given input.txt file.

```
Ubuntu (Snapshot 2) [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Activities Terminal Sat 11:26 ubh01@ubh01: ~
File Edit View Search Terminal Help
Failed Shuffles=0
Merged Map outputs=2
GC time elapsed (ms)=339
CPU time spent (ms)=1780
Physical memory (bytes) snapshot=713625600
Virtual memory (bytes) snapshot=5755002880
Total committed heap usage (bytes)=548405248
Shuffle Errors
BAD_ID=0
CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
Bytes Read=77
File Output Format Counters
Bytes Written=34
24/03/30 11:24:10 INFO streaming.StreamJob: Output directory: /mapreducer/output
ubh01@ubh01:~$ hdfs dfs -cat /mapreducer/output/part-0000
24/03/30 11:24:40 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes wh
ere applicable
cat: '/mapreducer/output/part-0000': No such file or directory
ubh01@ubh01:~$ hdfs dfs -ls /mapreducer/output
24/03/30 11:24:55 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes wh
ere applicable
Found 2 items
-rw-r--r-- 1 ubh01 supergroup 0 2024-03-30 11:24 /mapreducer/output/_SUCCESS
-rw-r--r-- 1 ubh01 supergroup 34 2024-03-30 11:24 /mapreducer/output/part-000000
ubh01@ubh01:~$ hdfs dfs -cat /mapreducer/output/part-000000
```

- After running map reducer job output get succeed and output file given by map reducer is stored in hadoop mapreducer/output directory.

```
Ubuntu (Snapshot 2) [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Activities Terminal
Sat 11:26
ubh01@ubh01: ~

CONNECTION=0
IO_ERROR=0
WRONG_LENGTH=0
WRONG_MAP=0
WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=77
File Output Format Counters
  Bytes Written=34
24/03/30 11:24:10 INFO streaming.StreamJob: Output directory: /mapreducer/output
ubh01@ubh01:~$ hdfs dfs -ls /mapreducer/output
24/03/30 11:24:40 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
cat: /mapreducer/output/part-0000: No such file or directory
ubh01@ubh01:~$ hdfs dfs -ls /mapreducer/output
24/03/30 11:24:55 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 2 items
-rw-r--r-- 1 ubh01 supergroup          0 2024-03-30 11:24 /mapreducer/output/_SUCCESS
-rw-r--r-- 1 ubh01 supergroup        34 2024-03-30 11:24 /mapreducer/output/part-00000
ubh01@ubh01:~$ hdfs dfs -cat /mapreducer/output/part-00000
24/03/30 11:25:05 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Bear 2
Car 3
Deer 2
River 2
and 1
ubh01@ubh01:~$
```

- By using this command “**hdfs dfs -ls /mapreducer/output**” we are checking whether the output files given by mapreducer are stored in output directory or not in Hadoop.
- By using this command “**hdfs -cat /mapreducer/output/part-0000**” we are checking the output which is given by the mapreducer is correct comparing with the local output.