BIG DATA

ASSIGNMENT 4

MAP REDUCE

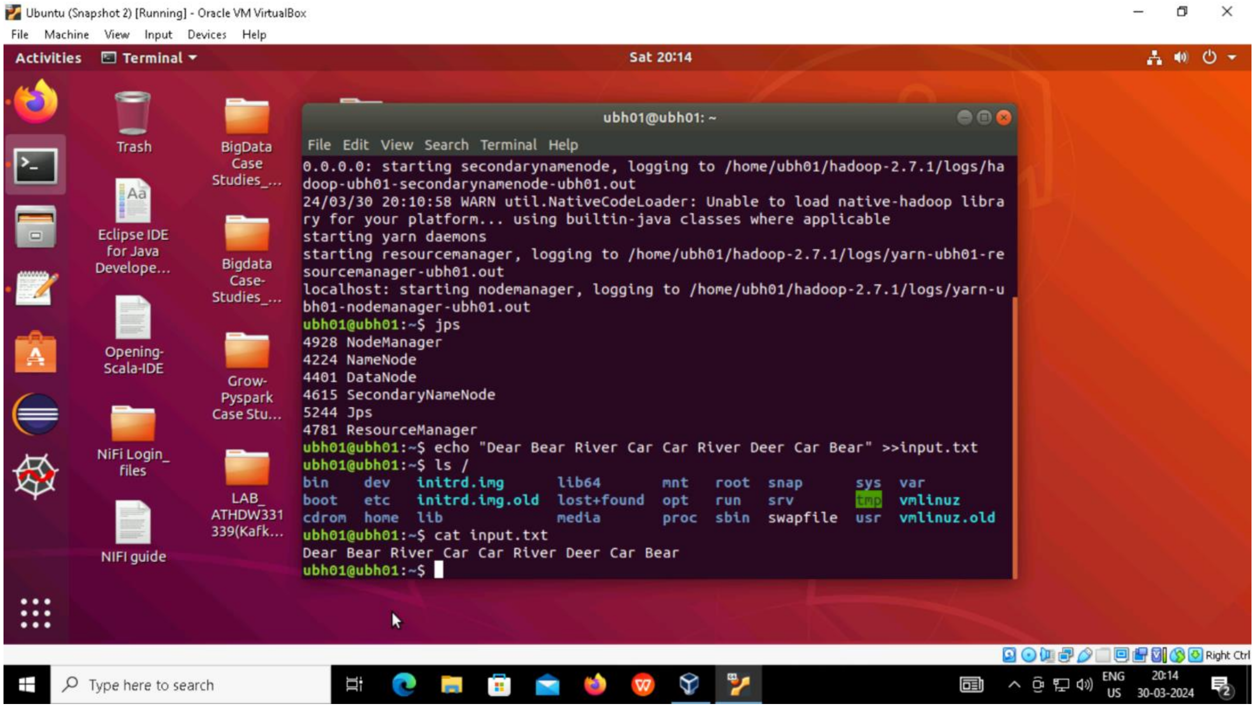
Done by:

SATHIYA PRASAD. L

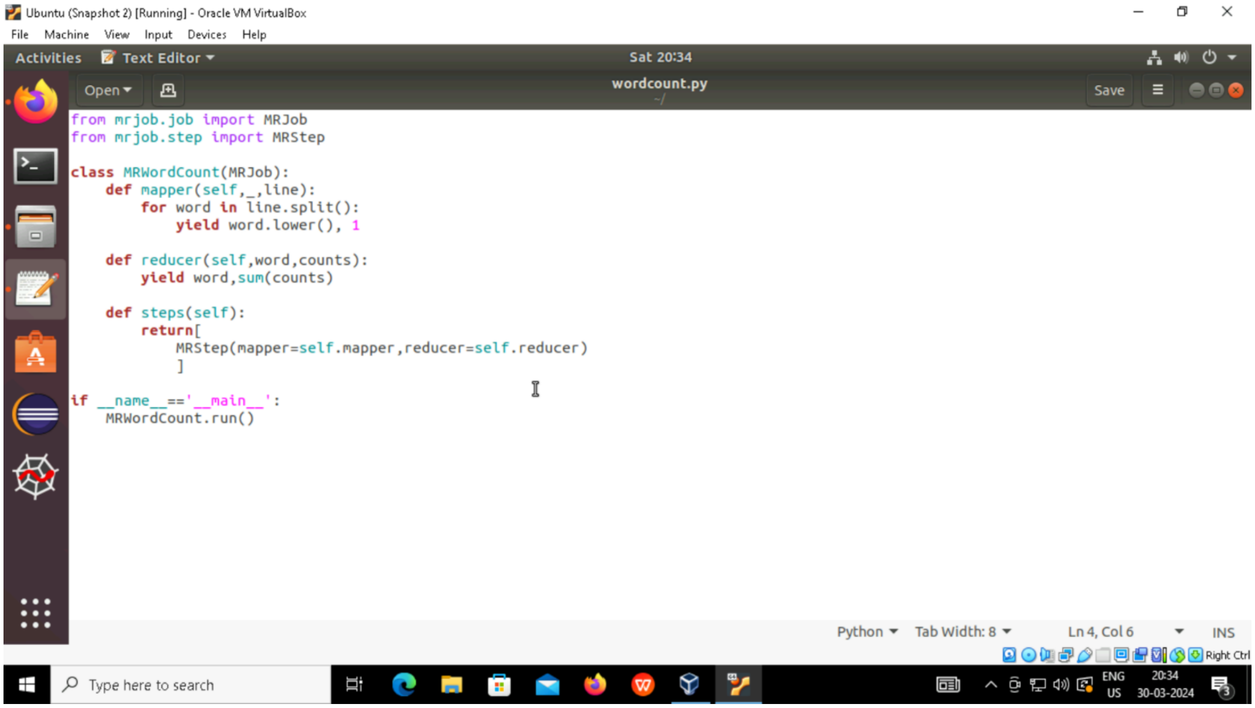
2319981

STEP 1:

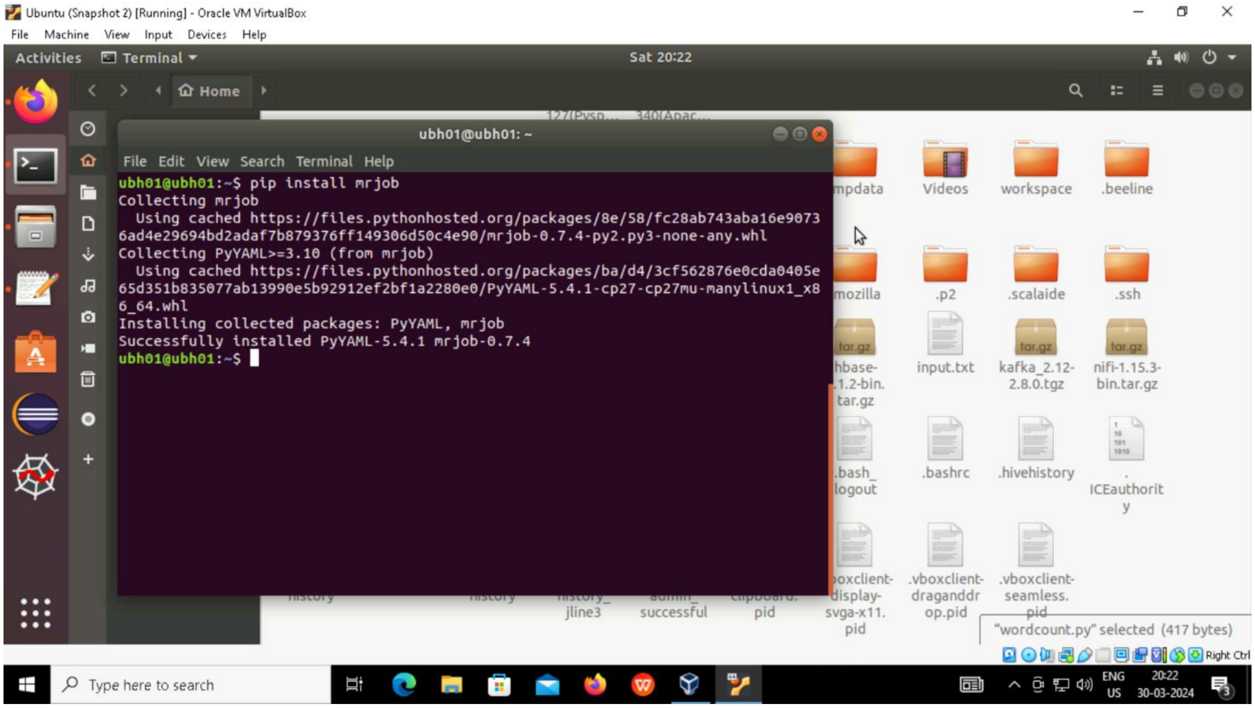
- Creating input file and wordcount python program file in local.
- Installing mrjob.
- Checking whether the function is working properly in local.



```
0.0.0.0: starting secondarynamenode, logging to /home/ubh01/hadoop-2.7.1/logs/hadoop-ubh01-secondarynamenode-ubh01.out
24/03/30 20:10:58 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
starting yarn daemons
starting resourcemanager, logging to /home/ubh01/hadoop-2.7.1/logs/yarn-ubh01-resourcemanager-ubh01.out
localhost: starting nodemanager, logging to /home/ubh01/hadoop-2.7.1/logs/yarn-ubh01-nodemanager-ubh01.out
ubh01@ubh01:~$ jps
4928 NodeManager
4224 NameNode
4401 DataNode
4615 SecondaryNameNode
5244 Jps
4781 ResourceManager
ubh01@ubh01:~$ echo "Dear Bear River Car Car River Deer Car Bear" >>input.txt
ubh01@ubh01:~$ ls /
bin    dev   initrd.img      lib64       mnt   root  snap      sys  var
boot   etc   initrd.img.old  lost+found  opt   run   srv       tmp  vmlinuz
cdrom  home  lib             media       proc  sbin  swapfile  usr  vmlinuz.old
ubh01@ubh01:~$ cat input.txt
Dear Bear River Car Car River Deer Car Bear
ubh01@ubh01:~$
```

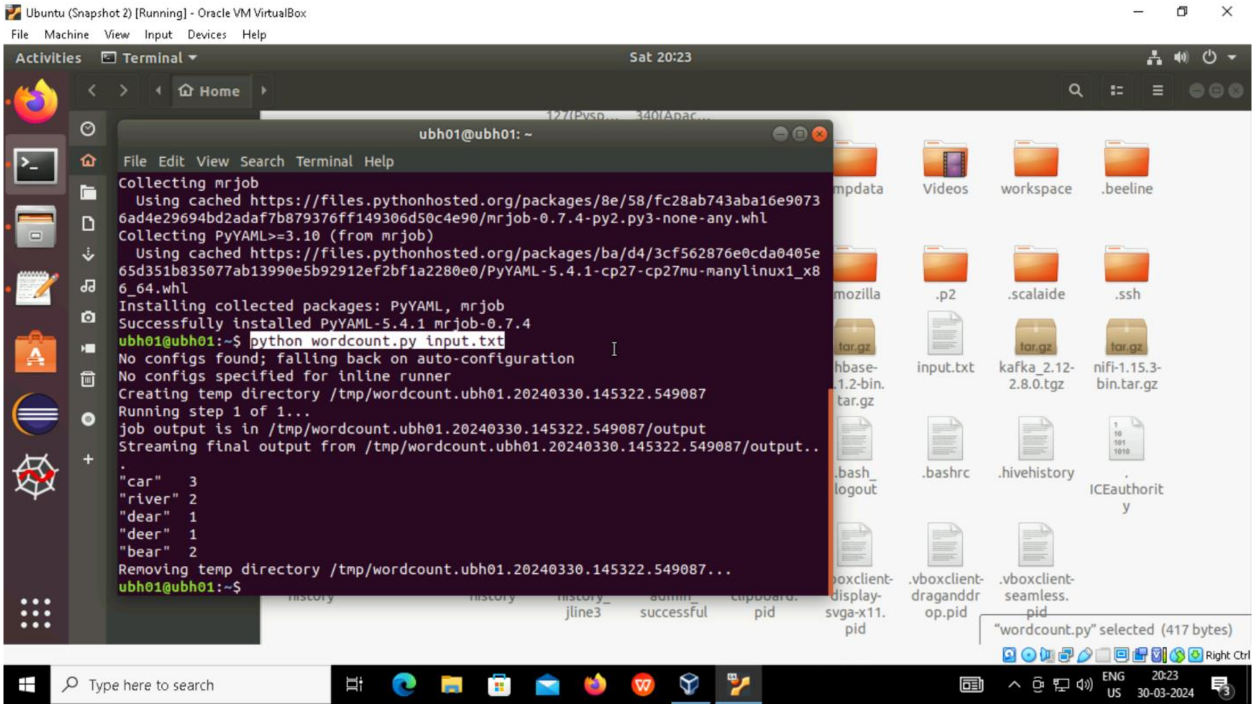


```python
from mrjob.job import MRJob
from mrjob.step import MRStep

class MRWordCount(MRJob):
    def mapper(self,_,line):
        for word in line.split():
            yield word.lower(), 1

    def reducer(self,word,counts):
        yield word,sum(counts)

    def steps(self):
        return[
            MRStep(mapper=self.mapper,reducer=self.reducer)
            ]

if __name__=='__main__':
    MRWordCount.run()
```



```
ubh01@ubh01:~$ pip install mrjob
Collecting mrjob
  Using cached https://files.pythonhosted.org/packages/8e/58/fc28ab743aba16e9073
6ad4e29694bd2adaf7b879376ff149306d50c4e90/mrjob-0.7.4-py2.py3-none-any.whl
Collecting PyYAML>=3.10 (from mrjob)
  Using cached https://files.pythonhosted.org/packages/ba/d4/3cf562876e0cda0405e
65d351b835077ab13990e5b92912ef2bf1a2280e0/PyYAML-5.4.1-cp27-cp27mu-manylinux1_x8
6_64.whl
Installing collected packages: PyYAML, mrjob
Successfully installed PyYAML-5.4.1 mrjob-0.7.4
ubh01@ubh01:~$ 
```

Ubuntu (Snapshot 2) [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Activities
Terminal
Sat 20:22
Home
ubh01@ubh01: ~
File Edit View Search Terminal Help
"wordcount.py" selected (417 bytes)
Type here to search
ENG US
20:22
30-03-2024

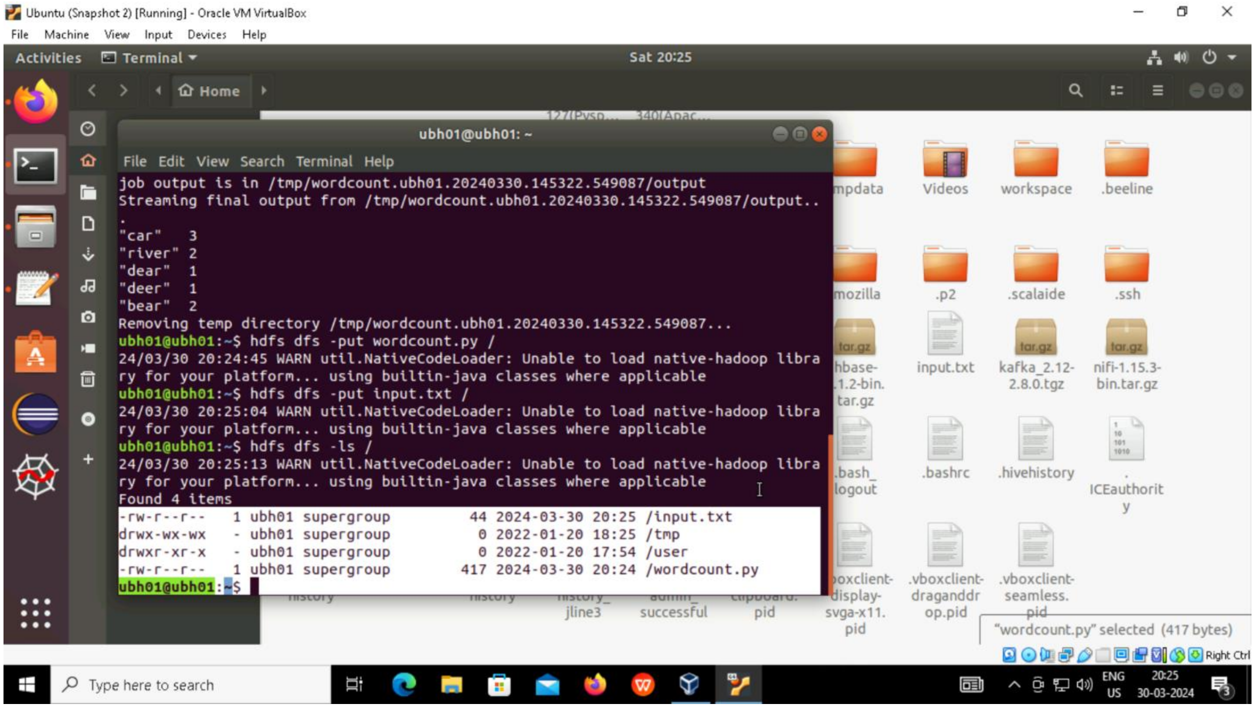


```
Collecting mrjob
  Using cached https://files.pythonhosted.org/packages/8e/58/fc28ab743aba16e9073
6ad4e29694bd2adaf7b879376ff149306d50c4e90/mrjob-0.7.4-py2.py3-none-any.whl
Collecting PyYAML>=3.10 (from mrjob)
  Using cached https://files.pythonhosted.org/packages/ba/d4/3cf562876e0cda0405e
65d351b835077ab13990e5b92912ef2bf1a2280e0/PyYAML-5.4.1-cp27-cp27mu-manylinux1_x8
6_64.whl
Installing collected packages: PyYAML, mrjob
Successfully installed PyYAML-5.4.1 mrjob-0.7.4
ubh01@ubh01:~$ python wordcount.py input.txt
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory /tmp/wordcount.ubh01.20240330.145322.549087
Running step 1 of 1...
job output is in /tmp/wordcount.ubh01.20240330.145322.549087/output
Streaming final output from /tmp/wordcount.ubh01.20240330.145322.549087/output..
.
"car"   3
"river" 2
"dear"  1
"deer"  1
"bear"  2
Removing temp directory /tmp/wordcount.ubh01.20240330.145322.549087...
ubh01@ubh01:~$ 
```

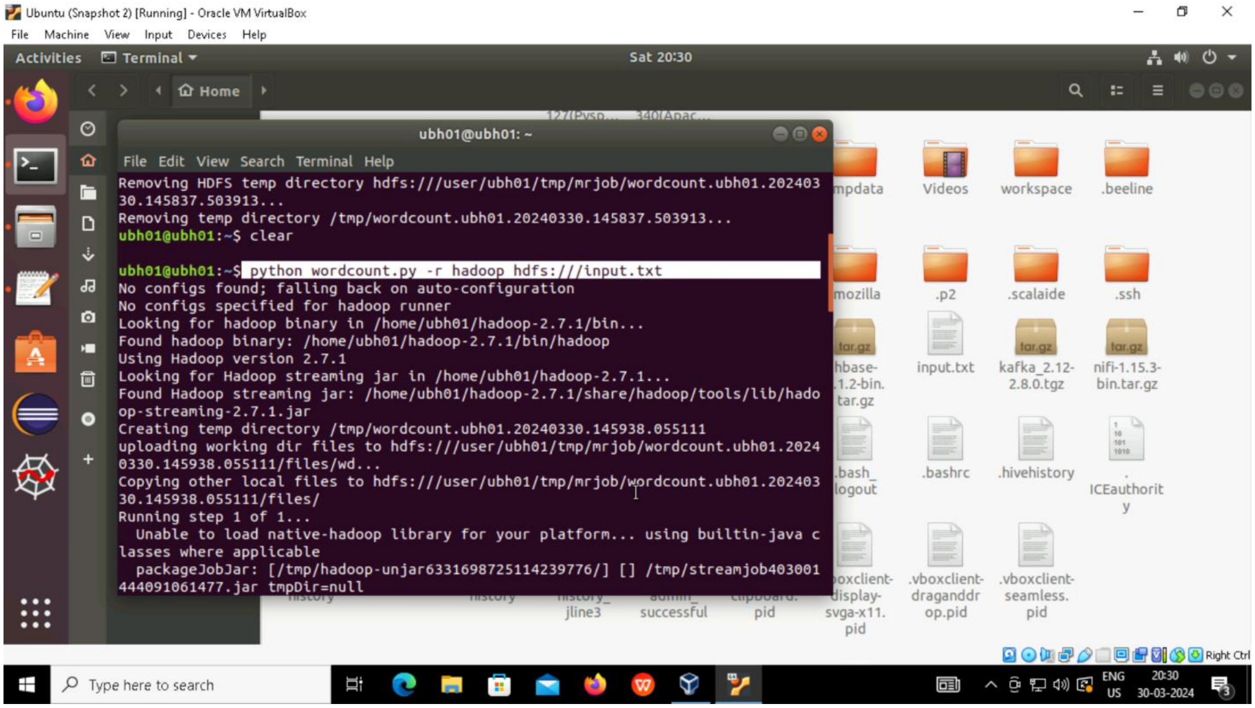
Ubuntu (Snapshot 2) [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Activities
Terminal
Sat 20:23
Home
ubh01@ubh01: ~
File Edit View Search Terminal Help
"wordcount.py" selected (417 bytes)
Type here to search
ENG US
20:23
30-03-2024

STEP 2:

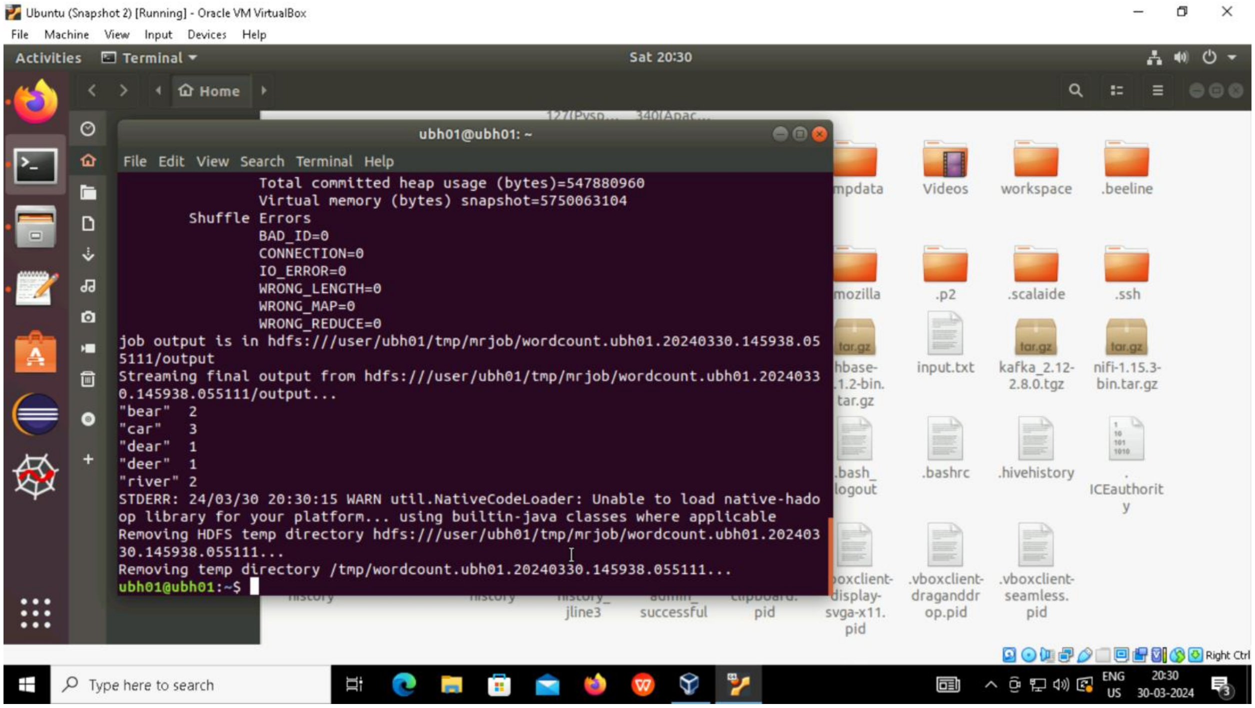
- Moving both the files to Hdfs.
- Executing the command to get the wordcount output.



```
job output is in /tmp/wordcount.ubh01.20240330.145322.549087/output
Streaming final output from /tmp/wordcount.ubh01.20240330.145322.549087/output..
.
"car"   3
"river" 2
"dear"  1
"deer"  1
"bear"  2
Removing temp directory /tmp/wordcount.ubh01.20240330.145322.549087...
ubh01@ubh01:~$ hdfs dfs -put wordcount.py /
24/03/30 20:24:45 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
ubh01@ubh01:~$ hdfs dfs -put input.txt /
24/03/30 20:25:04 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
ubh01@ubh01:~$ hdfs dfs -ls /
24/03/30 20:25:13 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
Found 4 items
-rw-r--r--   1 ubh01 supergroup         44 2024-03-30 20:25 /input.txt
drwx-wx-wx   - ubh01 supergroup          0 2022-01-20 18:25 /tmp
drwxr-xr-x   - ubh01 supergroup          0 2022-01-20 17:54 /user
-rw-r--r--   1 ubh01 supergroup        417 2024-03-30 20:24 /wordcount.py
ubh01@ubh01:~$
```



```
Removing HDFS temp directory hdfs:///user/ubh01/tmp/mrjob/wordcount.ubh01.202403
30.145837.503913...
Removing temp directory /tmp/wordcount.ubh01.20240330.145837.503913...
ubh01@ubh01:~$ clear

ubh01@ubh01:~$ python wordcount.py -r hadoop hdfs:///input.txt
No configs found; falling back on auto-configuration
No configs specified for hadoop runner
Looking for hadoop binary in /home/ubh01/hadoop-2.7.1/bin...
Found hadoop binary: /home/ubh01/hadoop-2.7.1/bin/hadoop
Using Hadoop version 2.7.1
Looking for Hadoop streaming jar in /home/ubh01/hadoop-2.7.1...
Found Hadoop streaming jar: /home/ubh01/hadoop-2.7.1/share/hadoop/tools/lib/hado
op-streaming-2.7.1.jar
Creating temp directory /tmp/wordcount.ubh01.20240330.145938.055111
uploading working dir files to hdfs:///user/ubh01/tmp/mrjob/wordcount.ubh01.2024
0330.145938.055111/files/wd...
Copying other local files to hdfs:///user/ubh01/tmp/mrjob/wordcount.ubh01.202403
30.145938.055111/files/
Running step 1 of 1...
  Unable to load native-hadoop library for your platform... using builtin-java c
lasses where applicable
  packageJobJar: [/tmp/hadoop-unjar6331698725114239776/] [] /tmp/streamjob403001
444091061477.jar tmpDir=null
```



```
                Total committed heap usage (bytes)=547880960
                Virtual memory (bytes) snapshot=5750063104
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
job output is in hdfs:///user/ubh01/tmp/mrjob/wordcount.ubh01.20240330.145938.05
5111/output
Streaming final output from hdfs:///user/ubh01/tmp/mrjob/wordcount.ubh01.2024033
0.145938.055111/output...
"bear"  2
"car"   3
"dear"  1
"deer"  1
"river" 2
STDERR: 24/03/30 20:30:15 WARN util.NativeCodeLoader: Unable to load native-hado
op library for your platform... using builtin-java classes where applicable
Removing HDFS temp directory hdfs:///user/ubh01/tmp/mrjob/wordcount.ubh01.202403
30.145938.055111...
Removing temp directory /tmp/wordcount.ubh01.20240330.145938.055111...
ubh01@ubh01:~$
```