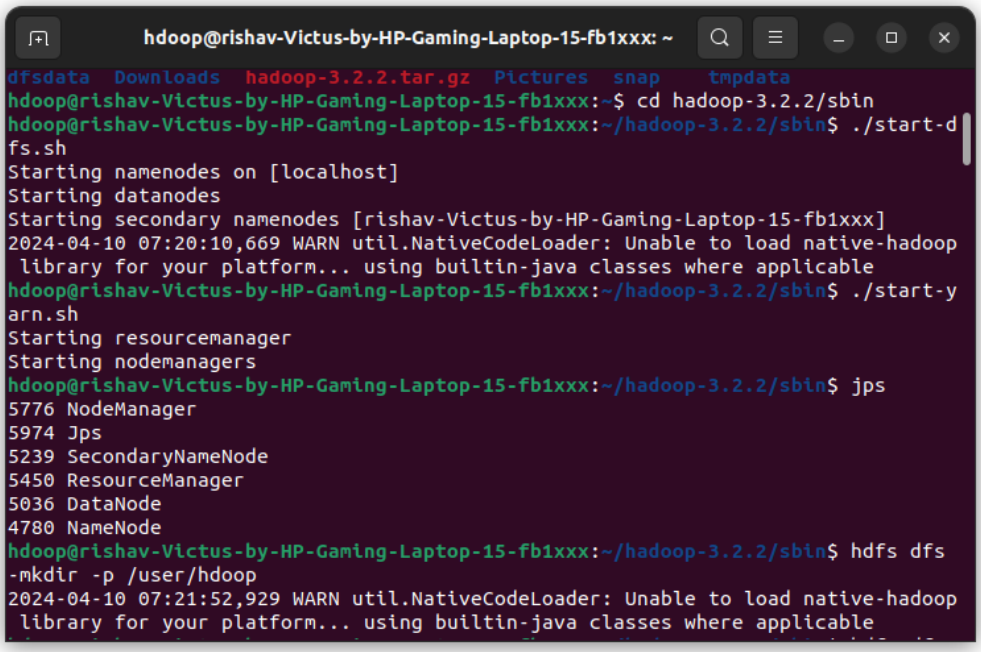
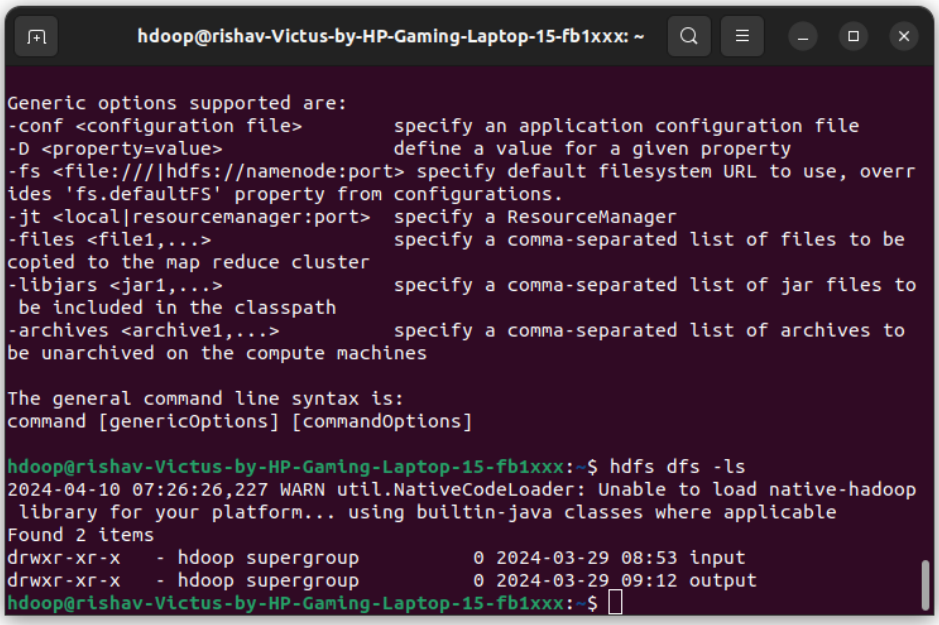
Workshop 6 – Detail Hands on Hadoop

1. Hadoop Directory Structure hdfs

dfs -mkdir -p /user/hdoop

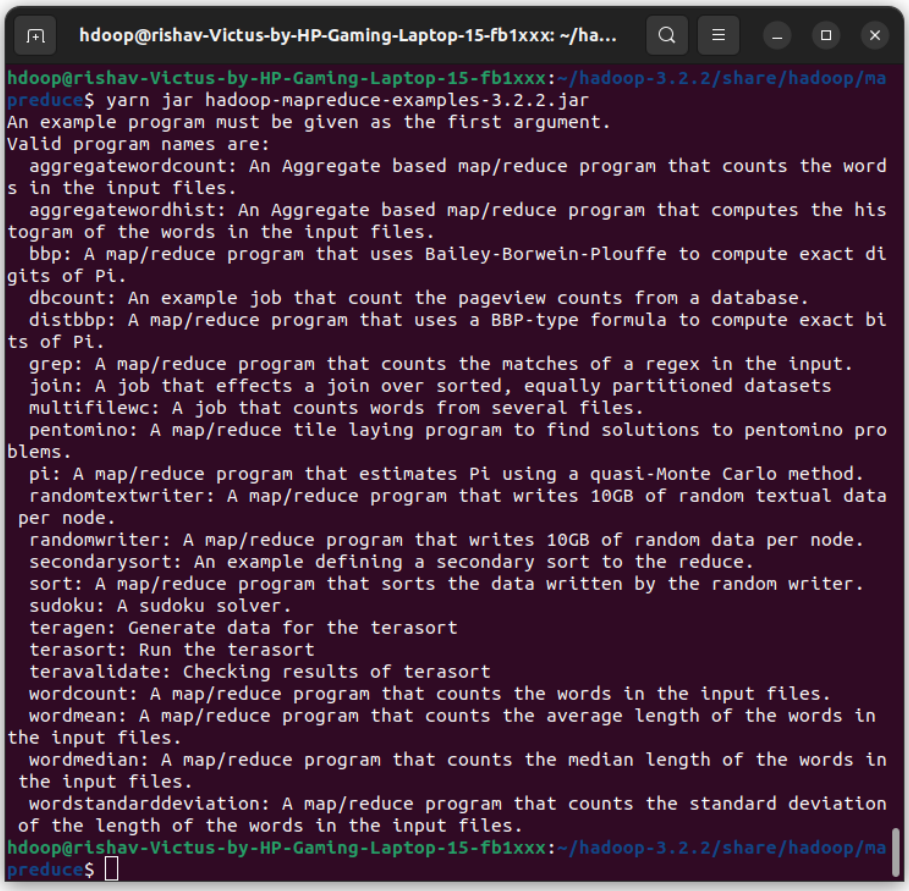


hdfs dfs –ls /user/hdoop



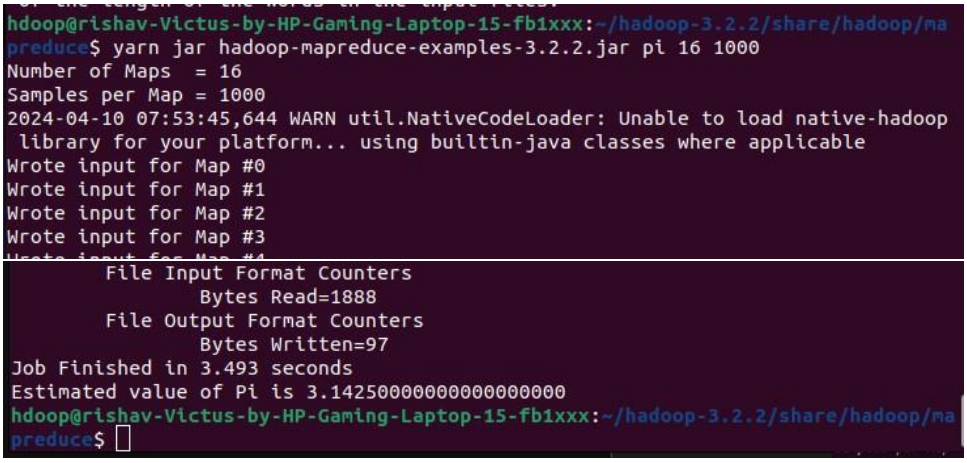
2. Example Programs

yarn jar $YARN_EXAMPLES/hadoop-mapreduce-examples.jar



2.1 Pi program

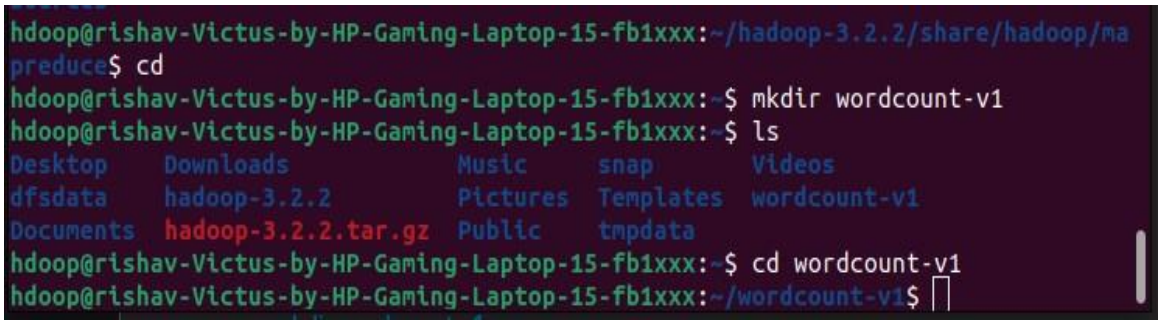
yarn jar $YARN_EXAMPLES/hadoop-mapreduce-examples.jar pi 16 1000



3. Word Count Version 1 3.1 Word

Count.java mkdir wordcount-v1 cd

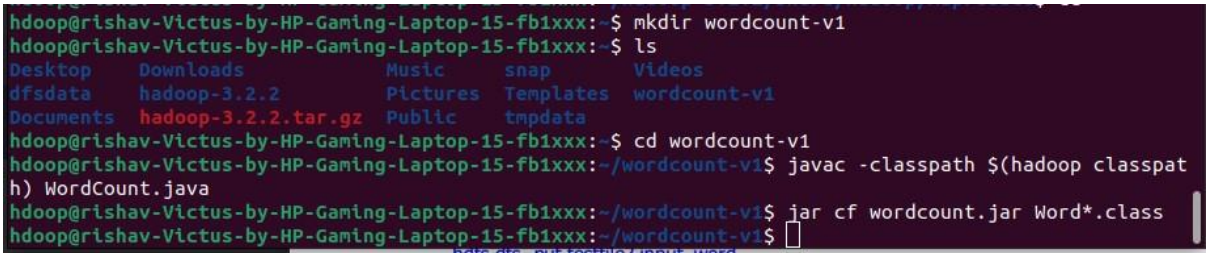
wordcount-v1



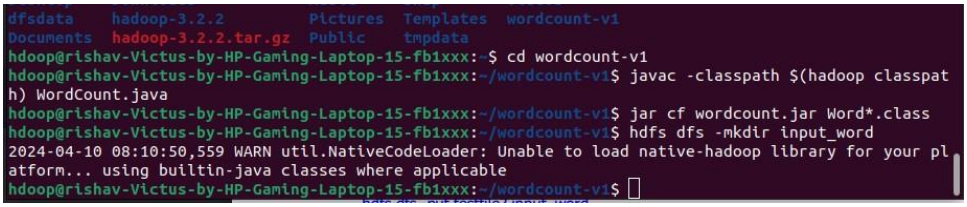
3.2 Running the Word Count Program

1. javac -classpath $(hadoop classpath) WordCount.java

2. jar cf wordcount.jar Word*.class



3. hdfs dfs -mkdir input_word



4. echo A long time ago in a galaxy far far away > testfile1

   echo Another episode of Star Wars > testfile2

```
2024-04-10 08:10:50,559 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your pl
atform... using builtin-java classes where applicable
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$ echo A long time ago in a galaxy f
ar far away > testfile1
echo Another episode of Star Wars > testfile2
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$ echo A long time ago in a galaxy f
ar far away > testfile1
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$ echo Another episode of Star Wars
> testfile2
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$
```

5. hdfs dfs -put testfile? input_word

```
ar far away > testfile1
echo Another episode of Star Wars > testfile2
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$ echo A long time ago in a galaxy f
ar far away > testfile1
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$ echo Another episode of Star Wars
> testfile2
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$ hdfs dfs -put testfile? input_word
2024-04-10 08:14:53,416 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your pl
atform... using builtin-java classes where applicable
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$
```

6. hdfs dfs -rm -R output_word

```
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$ echo Another episode of Star Wars
> testfile2
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$ hdfs dfs -put testfile? input_word
2024-04-10 08:14:53,416 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your pl
atform... using builtin-java classes where applicable
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$ hdfs dfs -rm -R output_word
2024-04-10 08:16:27,007 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your pl
atform... using builtin-java classes where applicable
rm: `output_word': No such file or directory
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$
```

7. hadoop jar wordcount.jar WordCount input_word output_word

```
        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=70
        File Output Format Counters
                Bytes Written=94
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$
```

```
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$ hadoop jar wordcount.jar WordCount
 input_word output_word
2024-04-10 08:21:48,865 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your pl
atform... using builtin-java classes where applicable
2024-04-10 08:21:49,268 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2024-04-10 08:21:49,329 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s
).
2024-04-10 08:21:49,329 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2024-04-10 08:21:49,432 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not p
erformed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2024-04-10 08:21:49,504 INFO input.FileInputFormat: Total input files to process : 2
2024-04-10 08:21:49,538 INFO mapreduce.JobSubmitter: number of splits:2
2024-04-10 08:21:49,628 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local1421412149_
0001
2024-04-10 08:21:49,628 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-04-10 08:21:49,716 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2024-04-10 08:21:49,716 INFO mapreduce.Job: Running job: job_local1421412149_0001
2024-04-10 08:21:49,717 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2024-04-10 08:21:49,722 INFO output.FileOutputCommitter: File Output Committer Algorithm version is
2
2024-04-10 08:21:49,722 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary
 folders under output directory:false, ignore cleanup failures: false
2024-04-10 08:21:49,722 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.l
ib.output.FileOutputCommitter
2024-04-10 08:21:49,747 INFO mapred.LocalJobRunner: Waiting for map tasks
2024-04-10 08:21:49,748 INFO mapred.LocalJobRunner: Starting task: attempt_local1421412149_0001_m_00
0000_0
2024-04-10 08:21:49,763 INFO output.FileOutputCommitter: File Output Committer Algorithm version is
2
2024-04-10 08:21:49,763 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary
 folders under output directory:false, ignore cleanup failures: false
```

8. hdfs dfs -ls output_word

```
                Bytes Written=94
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$ hdfs dfs -ls output_word
2024-04-10 08:24:24,942 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your pl
atform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 hdoop supergroup          0 2024-04-10 08:21 output_word/_SUCCESS
-rw-r--r--   1 hdoop supergroup         94 2024-04-10 08:21 output_word/part-r-00000
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$
```

9. hdfs dfs -cat output_word/part-r-00000

```
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$ hdfs dfs -cat output_word/part-r-0
0000
2024-04-10 08:26:30,633 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your pl
atform... using builtin-java classes where applicable
A       1
Another 1
Star    1
Wars    1
a       1
ago     1
away    1
episode 1
far     2
galaxy  1
in      1
long    1
of      1
time    1
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$
```

10. hdfs dfs -get output_word/part-r-00000 word-results.txt

```
time    1
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$ hdfs dfs -get output_word/part-r-0
0000 word-results.txt
2024-04-10 08:31:09,315 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your pl
atform... using builtin-java classes where applicable
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$
```

3.3 Rerunning the Word Count

hdfs dfs -rm -R output_word hadoop jar wordcount.jar

WordCount input_word output_word

```
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$ hdfs dfs -rm -R output_word
2024-04-10 08:54:46,441 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your pl
atform... using builtin-java classes where applicable
Deleted output_word
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$ hadoop jar wordcount.jar WordCount
 input_word output_word
2024-04-10 10:41:39,325 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your pl
atform... using builtin-java classes where applicable
2024-04-10 10:41:39,688 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2024-04-10 10:41:39,733 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s
).
2024-04-10 10:41:39,733 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2024-04-10 10:41:39,827 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not p
```

hdfs dfs -cat output_word/part-r-00000

```
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$ hdfs dfs -cat output_word/part-r-0
0000
2024-04-10 10:46:09,985 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your pl
atform... using builtin-java classes where applicable
cat: `output_word/part-r-00000 ': No such file or directory
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$
```

3.4 Using a larger dataset

1. Hadoop will work on any file in the input directory, so remove the previous testfiles:

   hdfs dfs -rm input_word/testfile?

```
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~$ hdfs dfs -rm input_word/testfile?
2024-04-10 10:55:33,548 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your pl
atform... using builtin-java classes where applicable
Deleted input_word/testfile1
Deleted input_word/testfile2
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~$
```

2. Save the new file to the input directory:

   hdfs dfs -put shakespeare.txt input_word

```
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~$ hdfs dfs -put t8.shakespeare.txt input_word
2024-04-10 11:01:58,776 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your pl
atform... using builtin-java classes where applicable
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~$
```

3. By now you will have created an output directory in the previous run of the program, so you need to delete this:

   hdfs dfs -rm -R output_word

```
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~$ hdfs dfs -rm -R output_word
2024-04-10 11:04:05,702 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your pl
atform... using builtin-java classes where applicable
Deleted output_word
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~$
```

4. Run the Map Reduce program:

   hadoop jar wordcount.jar WordCount input_word output_word

```
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$ hadoop jar wordcount.jar WordCount
 input_word output_word
2024-04-10 11:20:50,309 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your pl
atform... using builtin-java classes where applicable
2024-04-10 11:20:50,731 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2024-04-10 11:20:50,777 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s
).
2024-04-10 11:20:50,777 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2024-04-10 11:20:50,879 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not p
erformed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
```

```
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=5458199
        File Output Format Counters
                Bytes Written=717768
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$
```

5. Check what files are in the output directory:

    hdfs dfs -ls output_word

```
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$ hdfs dfs -ls output_word
2024-04-10 11:23:45,778 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your pl
atform... using builtin-java classes where applicable
Found 2 items
-rw-r--r--   1 hdoop supergroup          0 2024-04-10 11:20 output_word/_SUCCESS
-rw-r--r--   1 hdoop supergroup     717768 2024-04-10 11:20 output_word/part-r-00000
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$
```

6. See what is in the output file:

    hdfs dfs -cat output_word/part-r-00000

```
zip     1
zir,    1
zir.    1
zo      1
zodiac  1
zodiacs 1
zone,   1
zounds! 1
zounds, 1
zwagger'd       1
}       2
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$
```

7. This time there will be a bigger results set, so to view it properly, copy the file locally:

    hdfs dfs -get output_word/part-r-00000 shakespeare-results.txt

```
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$ hdfs dfs -get output_word/part-r-0
0000 shakespeare-results.txt
2024-04-10 11:27:11,954 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your pl
atform... using builtin-java classes where applicable
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$
```

8. Use the Linux command more to view the results file in the Operating System:

    more shakespeare-results.txt

```
zounds, 1
zwagger'd       1
}       2
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$
```

4. Word Count Version 2

The results file is quite large, so you can use the Linux command grep to search for a particular word:

grep anon shakespeare-results.txt

```
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$ grep anon shakespeare-results.txt
Canonized,      1
abroad-anon     1
anon    30
anon!   2
anon,   12
anon-   2
anon.   41
anon.-  1
anon;   5
anon?   1
anonymous       1
canon   2
canon,  1
canon.  2
canon;  1
canoniz'd       1
canoniz'd,      1
canonize        1
canonized       1
canons, 1
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$
```

grep Anon shakespeare-results.txt

```
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$ grep Anon shakespeare-results.txt
'Anon   1
'Anon!' 1
'Anon,  1
Anon    9
Anon!   1
Anon,   18
Anon.   2
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$
```

4.1 Making Word Count Case Insensitive

Some extra code is needed to make the word count not case sensitive.

Return to your home directory and create a new directory to keep the versions separate, for example:

cd

mkdir wordcount-v2

Change to this new directory:

cd wordcount-v2

Then copy the new version here.

Compile the file and create a Jar file as seen in steps 1 and 2 of Section 4.2

```
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v1$ cd
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~$ mkdir wordcount-v2
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~$ cd wordcount-v2
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v2$ javac -classpath $(hadoop classpat
h) WordCount.java
javac: file not found: WordCount.java
Usage: javac <options> <source files>
use -help for a list of possible options
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v2$ javac -classpath $(hadoop classpat
h) WordCount.java
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v2$
```

Remove the output directory: hdfs

dfs -rm -R output_word

```
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v2$ hdfs dfs -rm -R output_word
2024-04-10 11:52:47,540 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your pl
atform... using builtin-java classes where applicable
Deleted output_word
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v2$
```

Assuming that the shakespeare.txt file is still stored in the hdfs run the program:

hadoop jar wordcount.jar WordCount input_word output_word

```
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v2$ hadoop jar wordcount.jar WordCount
 input_word output_word
2024-04-10 11:56:19,317 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your pl
atform... using builtin-java classes where applicable
2024-04-10 11:56:19,697 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2024-04-10 11:56:19,742 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s
```

If everything has worked correctly, retrieve the results file from the hdfs:

hdfs dfs -get output_word/part-r-00000 shakespeare-results.txt

```
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v2$ hdfs dfs -get output_word/part-r-0
0000 shakespeare-results.txt
2024-04-10 12:01:19,981 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your pl
atform... using builtin-java classes where applicable
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v2$
```

Try using grep to search for anon again:

grep anon shakespeare-results.txt

```
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v2$ grep anon shakespeare-results.txt
Canonized,      1
abroad-anon     1
anon    30
anon!   2
anon,   12
anon-   2
anon.   41
anon.-  1
anon;   5
anon?   1
anonymous       1
canon   2
canon,  1
canon.  2
canon;  1
canoniz'd       1
canoniz'd,      1
canonize        1
canonized       1
canons, 1
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v2$
```

grep Anon shakespeare-results.txt

```
canonized,      1
canons, 1
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v2$ grep Anon shakespeare-results.txt
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v2$
```

5. Word Count Version 3

5.1 Removing Punctuation from Words

This final version of Word Count will use the Pattern class from java.util.regex

Return to your home directory and create another new directory to keep this version separate:

cd mkdir wordcount-

v3

```
canons, 1
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v2$ grep Anon shakespeare-results.txt
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v2$ cd
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~$ mkdir wordcount-v3
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~$
```

Change to this new directory: cd

wordcount-v3

```
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~$ cd wordcount-v3
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v3$
```

Then copy the new version here.

Compile the file and create a Jar file as seen in steps 1 and 2 earlier.

```
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v3$ javac -classpath $(hadoop classpat
h) WordCount.java
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v3$ jar cf wordcount.jar Word*.class
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v3$
```

Don't forget to remove the output directory:

hdfs dfs -rm -R output_word

```
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v3$ hdfs dfs -rm -R output_word
2024-04-10 12:38:00,440 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your pl
atform... using builtin-java classes where applicable
Deleted output_word
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v3$
```

Assuming that the shakespeare.txt file is still stored in the hdfs run the program:

hadoop jar wordcount.jar WordCount input_word output_word

```
        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=5458199
        File Output Format Counters
                Bytes Written=245910
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v3$
```

If everything has worked correctly, retrieve the results file from the hdfs:

hdfs dfs -get output_word/part-r-00000 shakespeare-results.csv

```
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v3$ hdfs dfs -get output_word/part-r-0
0000 shakespeare-results.csv
2024-04-10 13:06:00,324 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your pl
atform... using builtin-java classes where applicable
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v3$
```

Try using grep to search for anon again: grep

anon shakespeare-results.csv

```
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v3$ grep anon shakespeare-results.csv
anon,128
anonymous,1
canon,6
canoniz,2
canonize,1
canonized,2
canons,1
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v3$
```

This time anon will only appear once:

anon,128

anonymous,1 canon,6

canoniz,2 canonize,1

canonized,2 canons,1

The results are still case insensitive too, this should produce no results: grep

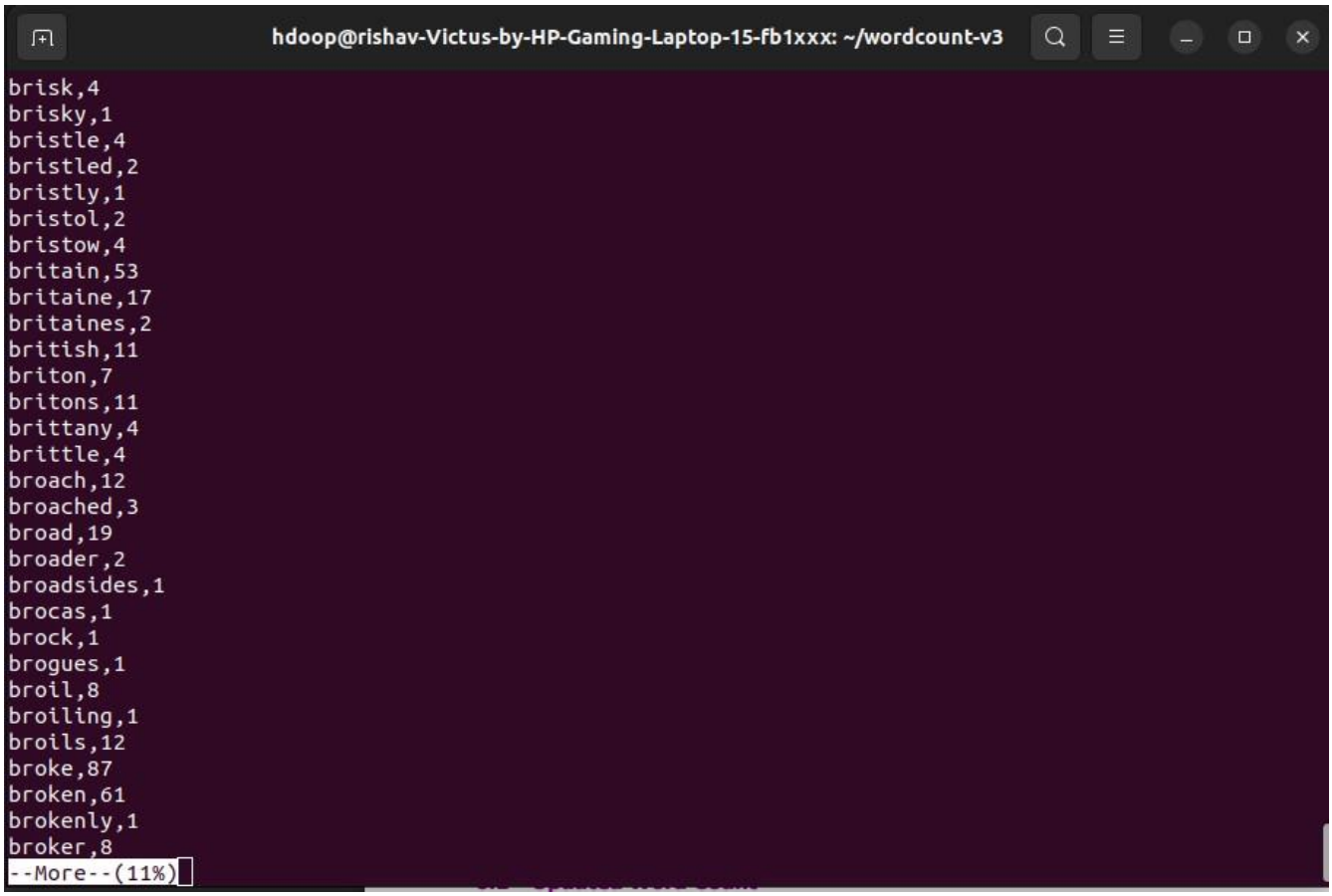
Anon shakespeare-results.csv

```
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v3$ grep Anon shakespeare-results.csv
grep: shakespeare-results.csv : No such file or directory
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v3$
```

6. Exercises to do

The results are still not perfect. If you use more to list the whole file: more

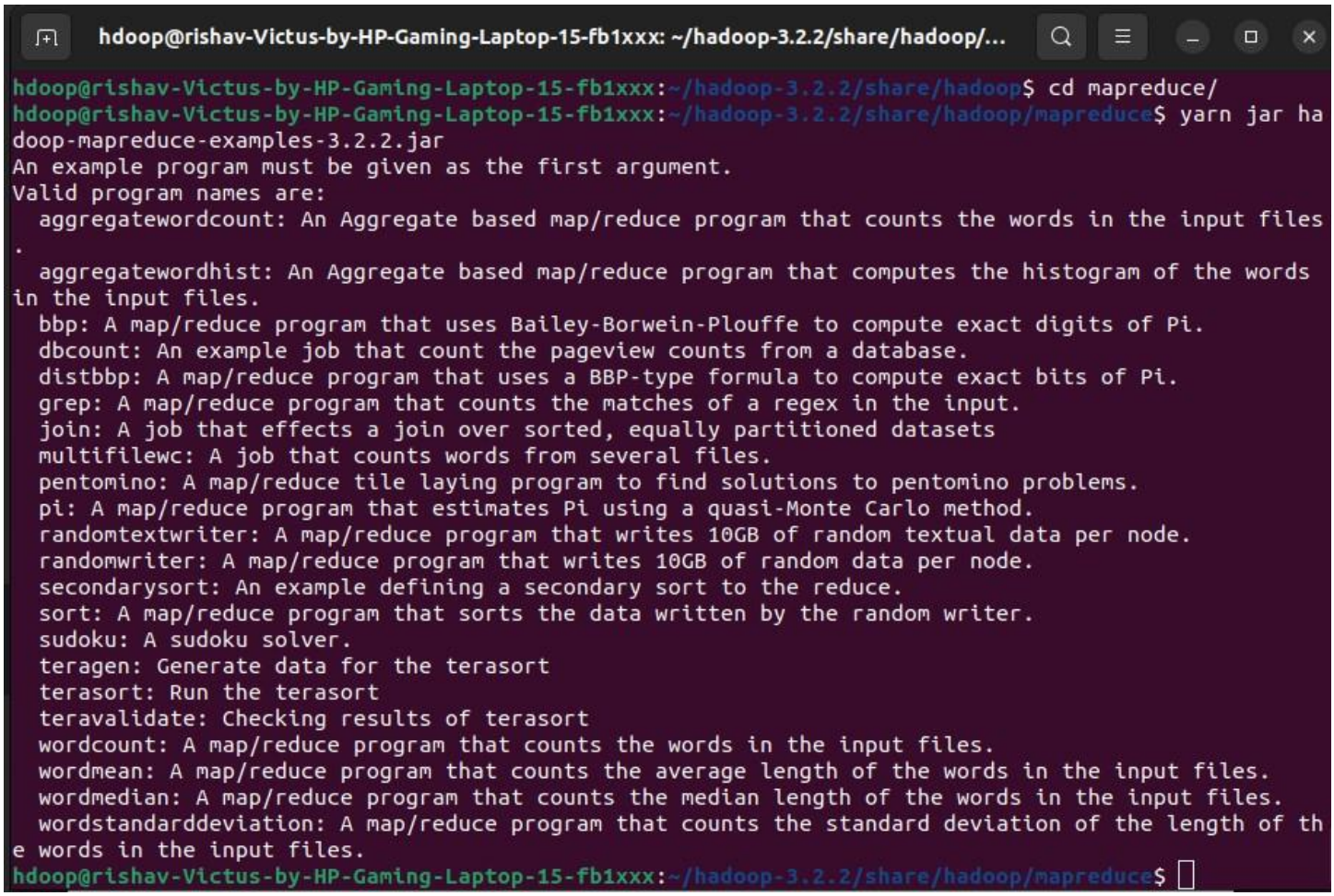
shakespeare-results.csv



6.2 Sample Programs

Section 3 Introduced the sample programs that come with Hadoop. As a reminder to see the full list type:

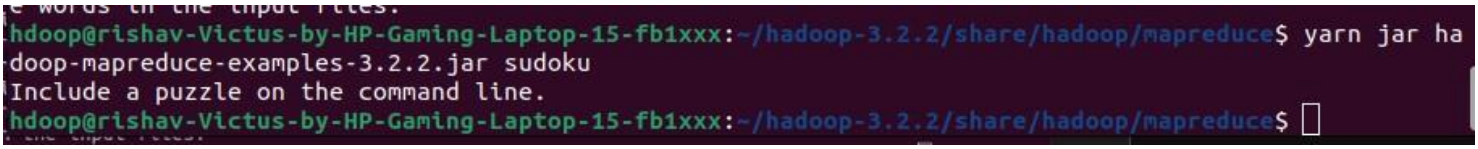
yarn jar $YARN_EXAMPLES/hadoop-mapreduce-examples.jar



To find out what options are required, add one of the above to the command, for example, if you are a Sudoku fan:

yarn jar $YARN_EXAMPLES/hadoop-mapreduce-examples.jar sudoku Which tells

you to: Include a puzzle on the command line.



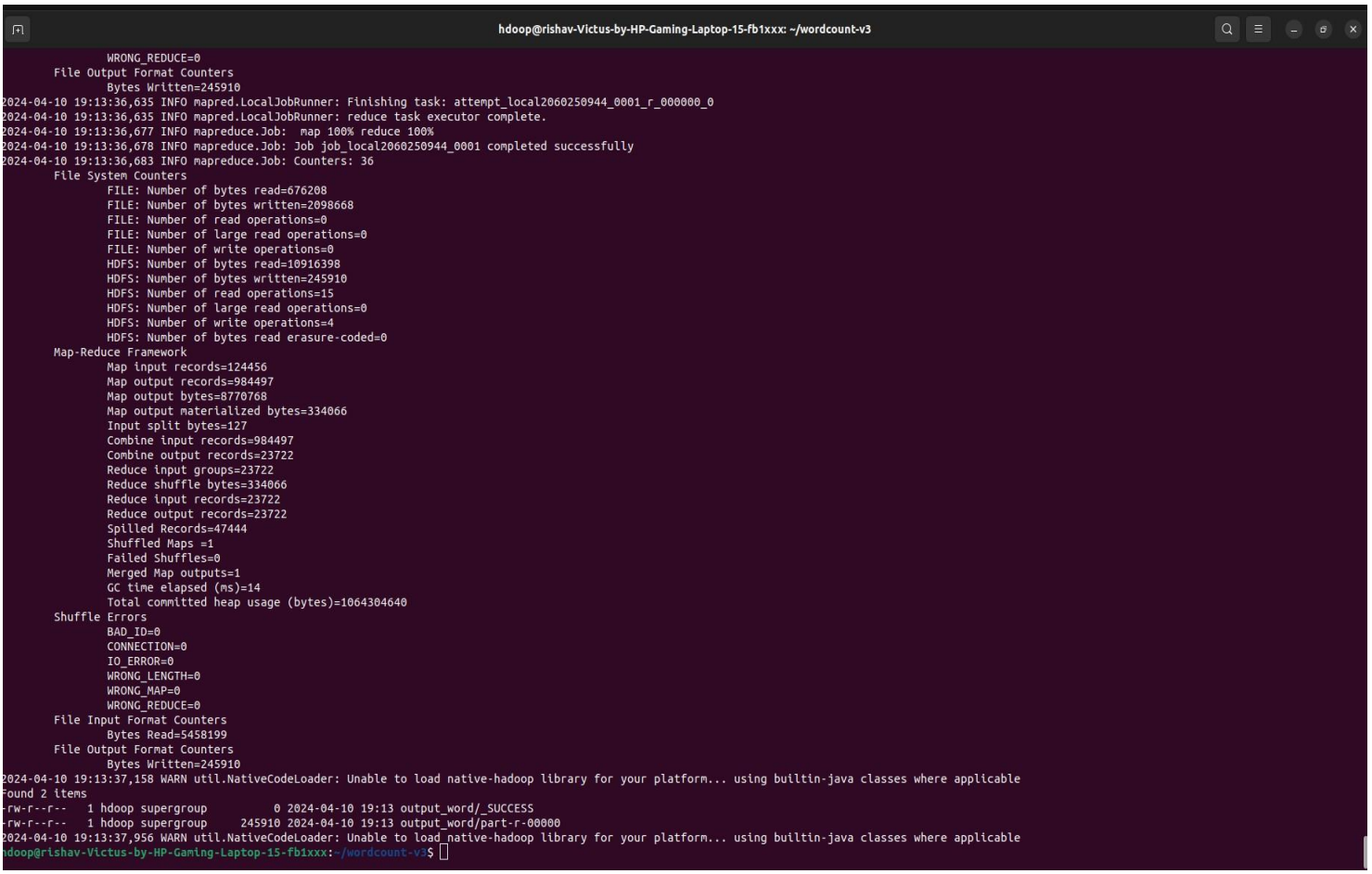
To run Sudoku with this puzzle:

yarn jar $YARN_EXAMPLES/hadoop-mapreduce-examples.jar sudoku puzzle.txt

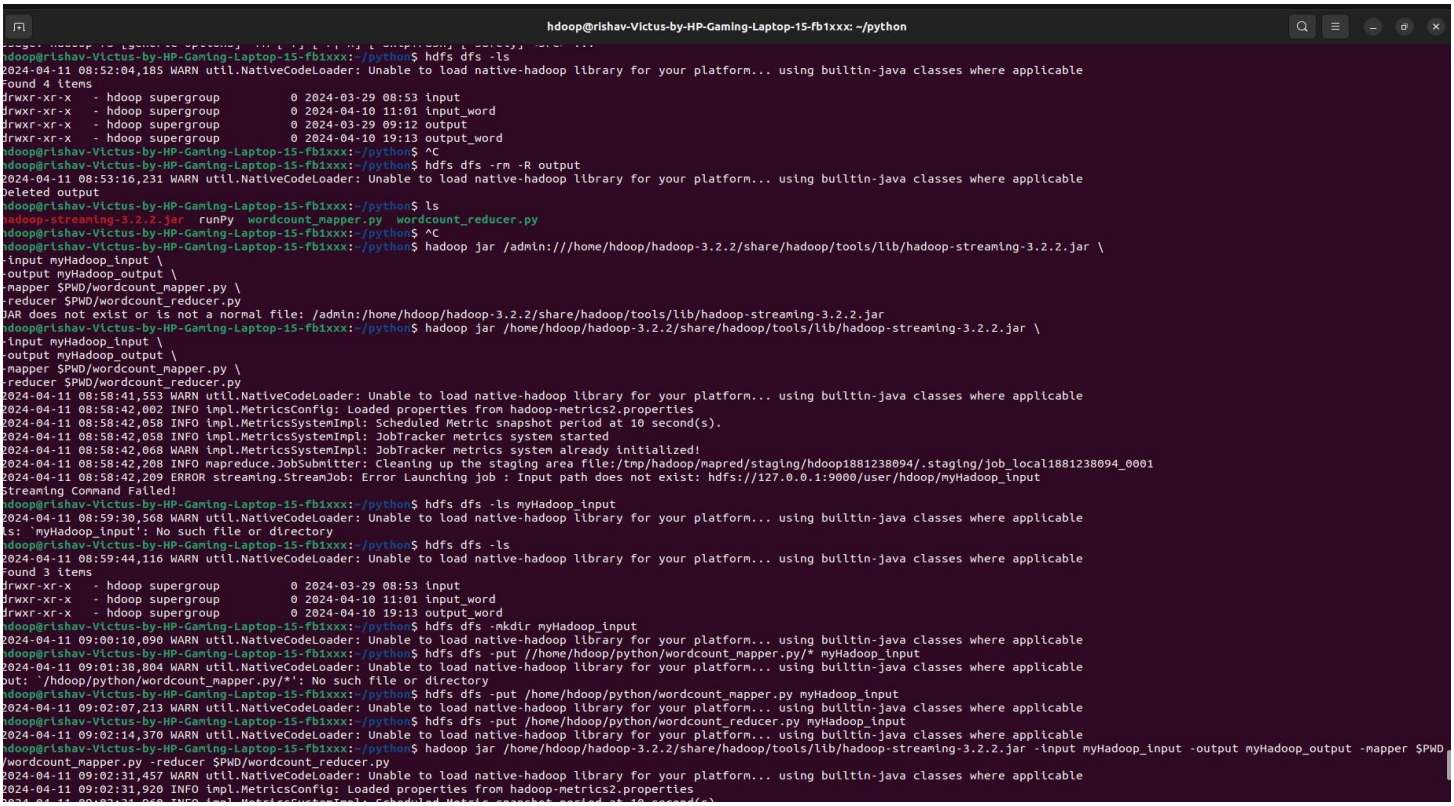
```
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/hadoop-3.2.2/share/hadoop/mapreduce$ yarn jar ha
doop-mapreduce-examples-3.2.2.jar sudoku
Include a puzzle on the command line.
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/hadoop-3.2.2/share/hadoop/mapreduce$ yarn jar ha
doop-mapreduce-examples-3.2.2.jar sudoku puzzle.txt
Solving puzzle.txt
8 5 1 3 9 2 6 4 7
4 3 2 6 7 8 1 9 5
7 9 6 5 1 4 3 8 2
6 1 4 8 2 3 7 5 9
5 7 8 9 6 1 4 2 3
3 2 9 4 5 7 8 1 6
9 4 7 2 8 6 5 3 1
1 8 5 7 3 9 2 6 4
2 6 3 1 4 5 9 7 8

Found 1 solutions
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/hadoop-3.2.2/share/hadoop/mapreduce$
```

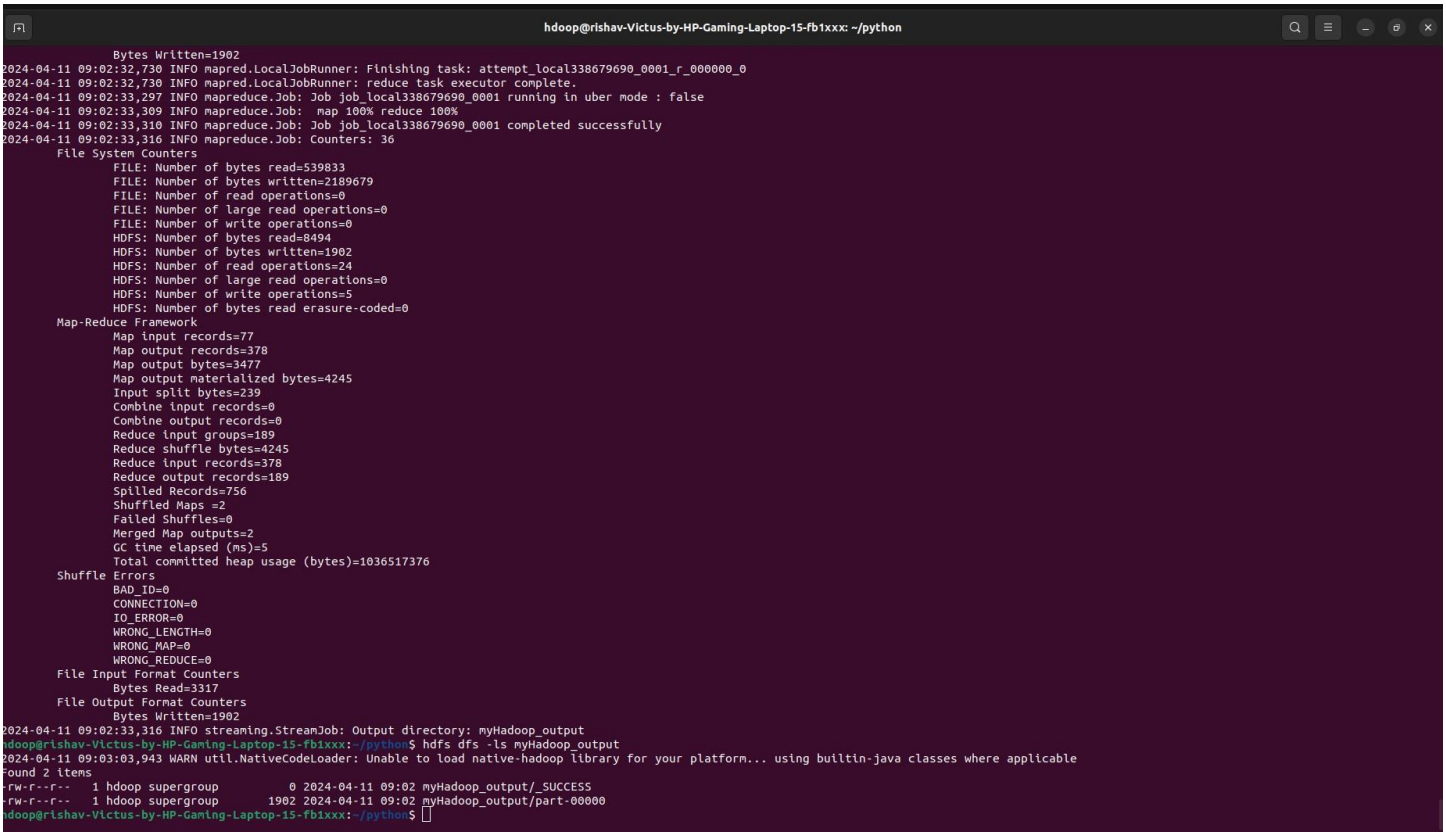
6.3 Shell Script

```
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx: ~/wordcount-v3
zir,2
zo,1
zodiac,1
zodiacs,1
zone,1
zounds,24
zwagger,1
|,33
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v3$ if [ -f word-results.txt ];
then rm word-results.txt;
fi
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v3$ hdfs dfs -get output_word/part-r-00000 word-results.txt
2024-04-10 19:06:59,217 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v3$ chmod u+x runProg
chmod: cannot access 'runProg': No such file or directory
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v3$ chmod u+x runProg
chmod: changing permissions of 'runProg': Operation not permitted
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v3$ ^C
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v3$ ls -l runProg
-rw-r--r-- 1 root root 521 मार्च    8  2019 runProg
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v3$ sudo chmod u+x runProg
[sudo] password for hdoop:
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v3$ ./runProg
bash: ./runProg: Permission denied
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v3$ ^C
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v3$ chmod +x runProg
chmod: changing permissions of 'runProg': Operation not permitted
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v3$ sudo chmod +x runProg
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/wordcount-v3$ ./runProg
2024-04-10 19:13:32,261 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2024-04-10 19:13:33,048 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Deleted output_word
2024-04-10 19:13:33,905 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
2024-04-10 19:13:34,259 INFO impl.MetricsConfig: Loaded properties from hadoop-metrics2.properties
2024-04-10 19:13:34,311 INFO impl.MetricsSystemImpl: Scheduled Metric snapshot period at 10 second(s).
2024-04-10 19:13:34,311 INFO impl.MetricsSystemImpl: JobTracker metrics system started
2024-04-10 19:13:34,405 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
2024-04-10 19:13:34,475 INFO input.FileInputFormat: Total input files to process : 1
2024-04-10 19:13:34,506 INFO mapreduce.JobSubmitter: number of splits:1
2024-04-10 19:13:34,592 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_local2060250944_0001
2024-04-10 19:13:34,592 INFO mapreduce.JobSubmitter: Executing with tokens: []
2024-04-10 19:13:34,660 INFO mapreduce.Job: The url to track the job: http://localhost:8080/
2024-04-10 19:13:34,660 INFO mapreduce.Job: Running job: job_local2060250944_0001
2024-04-10 19:13:34,661 INFO mapred.LocalJobRunner: OutputCommitter set in config null
2024-04-10 19:13:34,666 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-04-10 19:13:34,666 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2024-04-10 19:13:34,666 INFO mapred.LocalJobRunner: OutputCommitter is org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter
2024-04-10 19:13:34,688 INFO mapred.LocalJobRunner: Waiting for map tasks
2024-04-10 19:13:34,689 INFO mapred.LocalJobRunner: Starting task: attempt_local2060250944_0001_m_000000_0
2024-04-10 19:13:34,703 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 2
2024-04-10 19:13:34,703 INFO output.FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
2024-04-10 19:13:34,716 INFO mapred.Task:  Using ResourceCalculatorProcessTree : [ ]
2024-04-10 19:13:34,718 INFO mapred.MapTask: Processing split: hdfs://127.0.0.1:9000/user/hdoop/input_word/t8.shakespeare.txt:0+5458199
2024-04-10 19:13:34,765 INFO mapred.MapTask: (EQUATOR) 0 kvi 26214396(104857584)
2024-04-10 19:13:34,765 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100
2024-04-10 19:13:34,765 INFO mapred.MapTask: soft limit at 83886080
```



7. Python Code




hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx: ~/python
Bytes Written=1902
2024-04-11 09:02:32,730 INFO mapred.LocalJobRunner: Finishing task: attempt_local338679690_0001_r_000000_0
2024-04-11 09:02:32,730 INFO mapred.LocalJobRunner: reduce task executor complete.
2024-04-11 09:02:33,297 INFO mapreduce.Job: Job job_local338679690_0001 running in uber mode : false
2024-04-11 09:02:33,309 INFO mapreduce.Job: map 100% reduce 100%
2024-04-11 09:02:33,310 INFO mapreduce.Job: Job job_local338679690_0001 completed successfully
2024-04-11 09:02:33,316 INFO mapreduce.Job: Counters: 36
File System Counters
FILE: Number of bytes read=539833
FILE: Number of bytes written=2189679
FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=8494
HDFS: Number of bytes written=1902
HDFS: Number of read operations=24
HDFS: Number of large read operations=0
HDFS: Number of write operations=5
HDFS: Number of bytes read erasure-coded=0
Map-Reduce Framework
Map input records=77
Map output records=378
Map output bytes=3477
Map output materialized bytes=4245
Input split bytes=239
Combine input records=0
Combine output records=0
Reduce input groups=189
Reduce shuffle bytes=4245
Reduce input records=378
Reduce output records=189
Spilled Records=756
Shuffled Maps =2
Failed Shuffles=0
Merged Map outputs=2
GC time elapsed (ms)=5
Total committed heap usage (bytes)=1036517376
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=3317
File Output Format Counters
Bytes Written=1902
2024-04-11 09:02:33,316 INFO streaming.StreamJob: Output directory: myHadoop_output
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/python$ hdfs dfs -ls myHadoop_output
2024-04-11 09:03:03,943 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 hdoop supergroup 0 2024-04-11 09:02 myHadoop_output/_SUCCESS
-rw-r--r-- 1 hdoop supergroup 1902 2024-04-11 09:02 myHadoop_output/part-00000
hdoop@rishav-Victus-by-HP-Gaming-Laptop-15-fb1xxx:~/python$