CSE 488 (Section 1)
[Summer 2022]

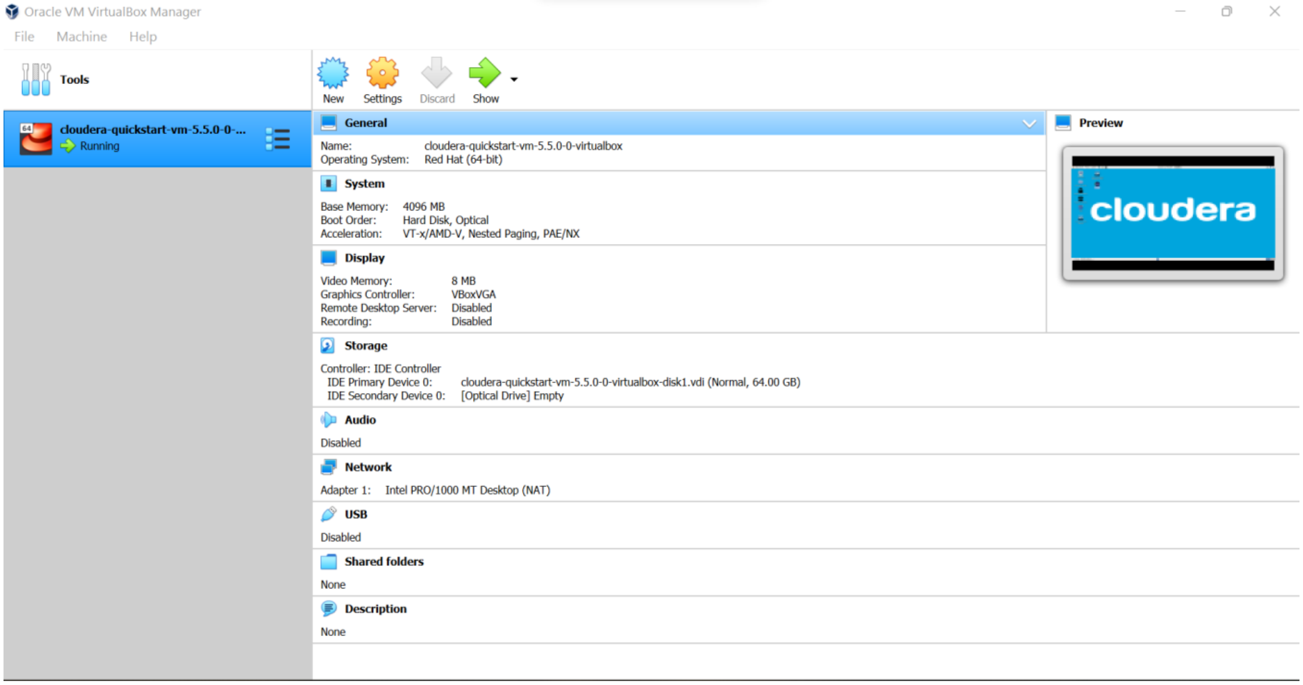
Lab Assignment Submission Report

Assignment Title: Introduction to Hadoop and MapReduce Programming

Submitted by:
Rizvee Hassan Prito
2019-3-60-041

1. Screenshots

Screenshot of the set-up of virtual machine:



General

```
Name:              cloudera-quickstart-vm-5.5.0-0-virtualbox
Operating System:  Red Hat (64-bit)
```

System

```
Base Memory:   4096 MB
Boot Order:    Hard Disk, Optical
Acceleration:  VT-x/AMD-V, Nested Paging, PAE/NX
```

Display

```
Video Memory:             8 MB
Graphics Controller:      VBoxVGA
Remote Desktop Server:    Disabled
Recording:                Disabled
```

Storage

```
Controller: IDE Controller
  IDE Primary Device 0:     cloudera-quickstart-vm-5.5.0-0-virtualbox-disk1.vdi (Normal, 64.00 GB)
  IDE Secondary Device 0:   [Optical Drive] Empty
```

Audio

Disabled

Network

```
Adapter 1:   Intel PRO/1000 MT Desktop (NAT)
```

USB

Disabled

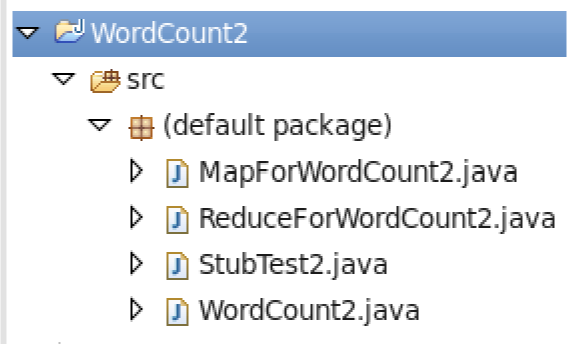
Shared folders

None

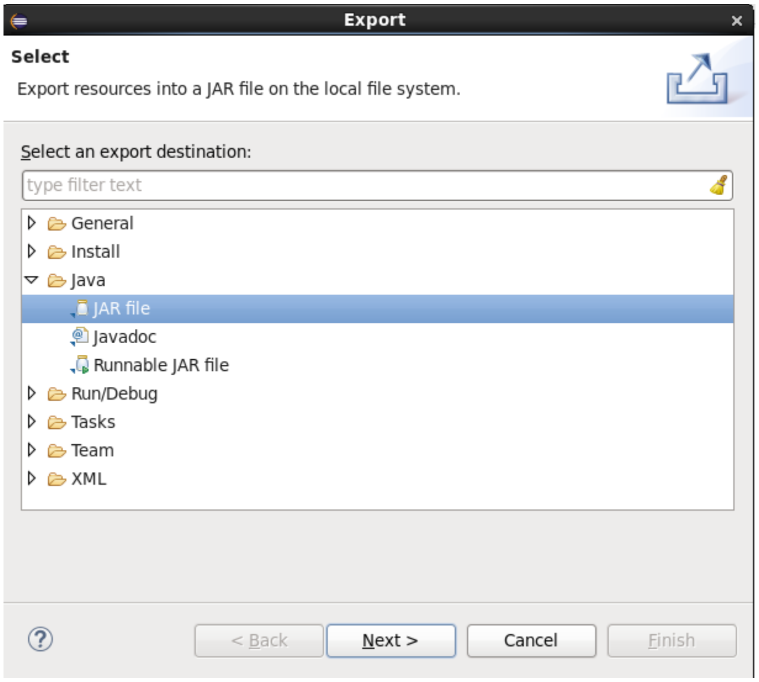
Description

None

Screenshots of different stages of the program’s execution:



```
WordCount2
  src
    (default package)
      MapForWordCount2.java
      ReduceForWordCount2.java
      StubTest2.java
      WordCount2.java
```



Export

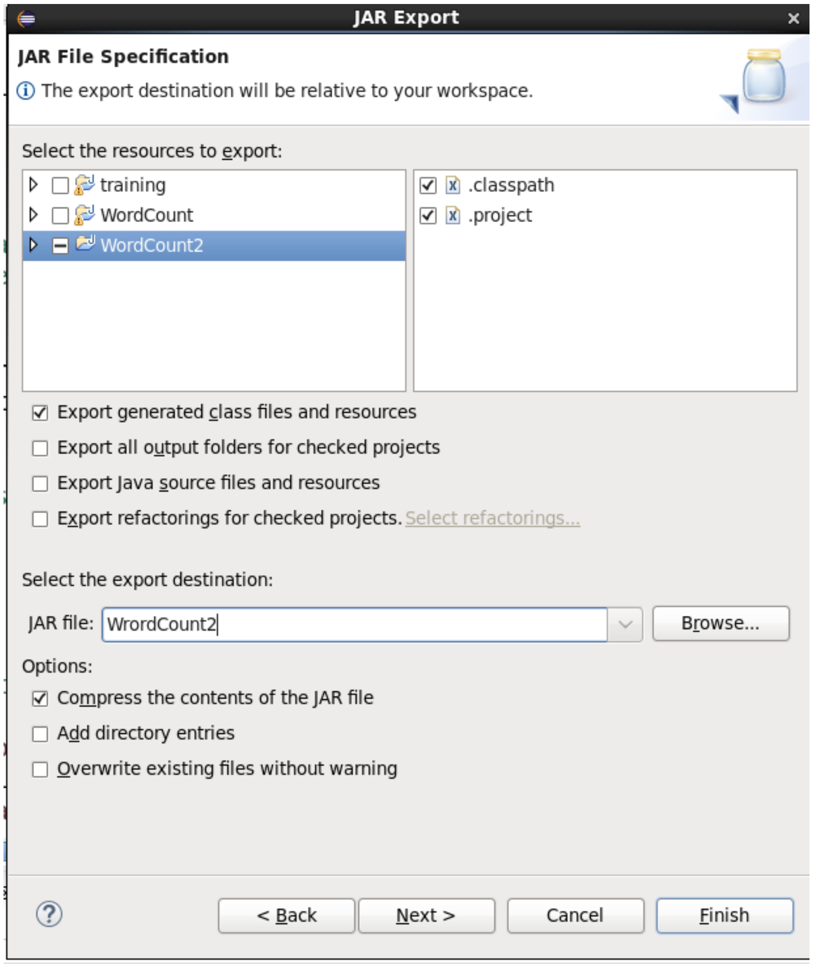
Select

Export resources into a JAR file on the local file system.

Select an export destination:

type filter text

- General
- Install
- Java
  - JAR file
  - Javadoc
  - Runnable JAR file
- Run/Debug
- Tasks
- Team
- XML



JAR Export

JAR File Specification

ⓘ The export destination will be relative to your workspace.

Select the resources to export:

- □ training
- □ WordCount
- ⊟ WordCount2

☑ .classpath
☑ .project

☑ Export generated class files and resources

□ Export all output folders for checked projects

□ Export Java source files and resources

□ Export refactorings for checked projects. Select refactorings...

Select the export destination:

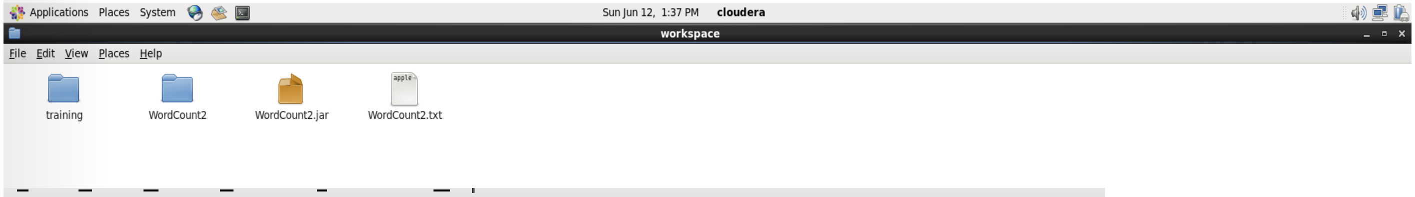
JAR file: WordCount2

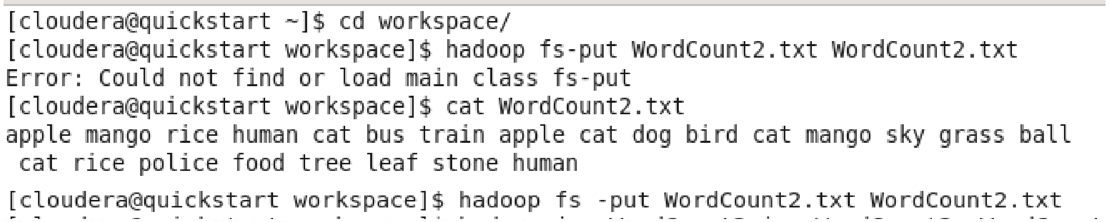
Options:

☑ Compress the contents of the JAR file

□ Add directory entries

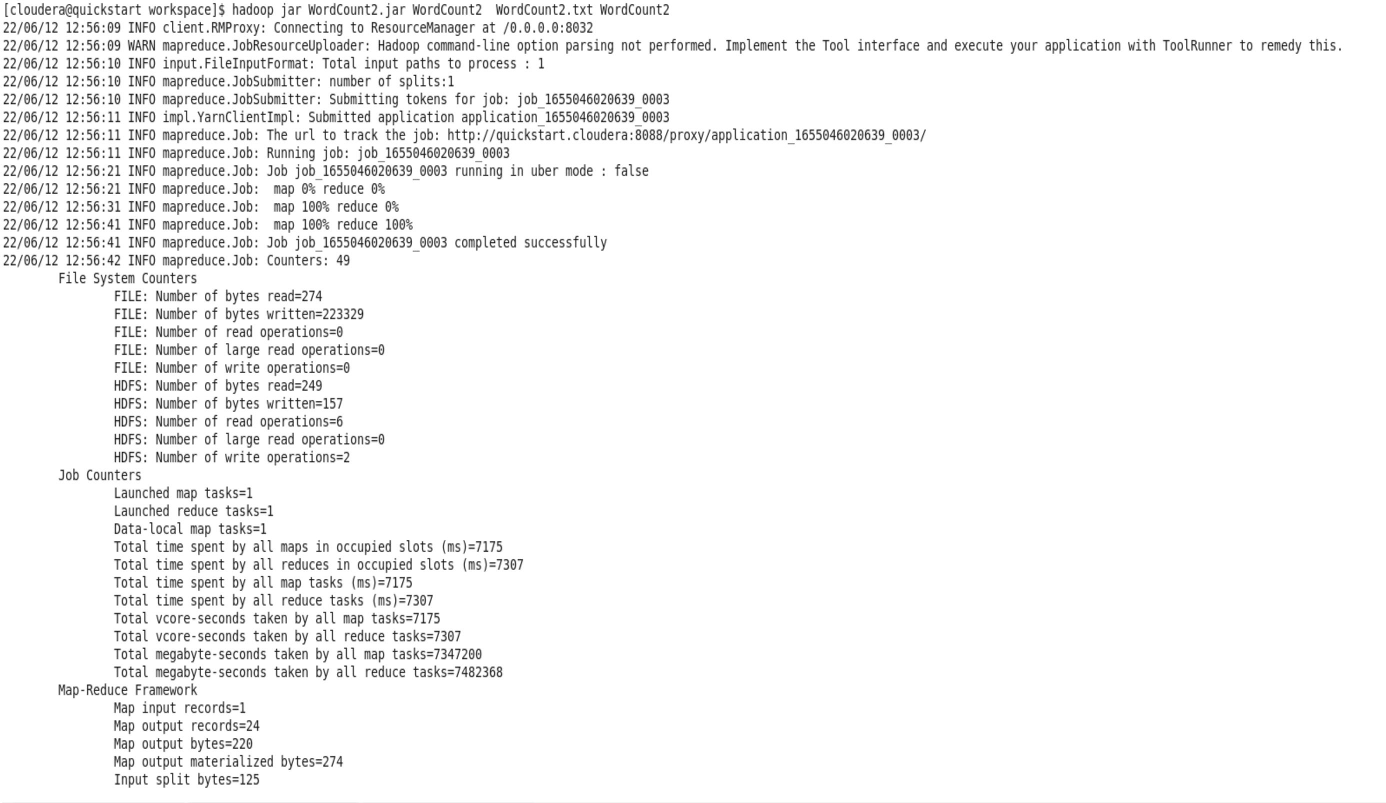
□ Overwrite existing files without warning



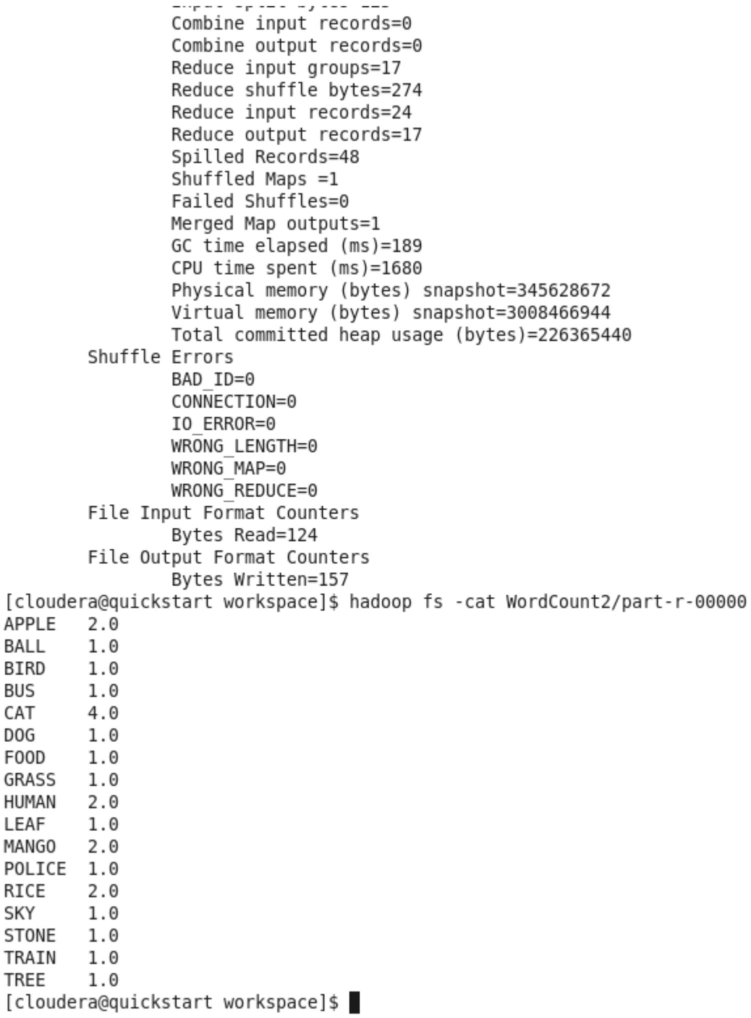


```
[cloudera@quickstart ~]$ cd workspace/
[cloudera@quickstart workspace]$ hadoop fs-put WordCount2.txt WordCount2.txt
Error: Could not find or load main class fs-put
[cloudera@quickstart workspace]$ cat WordCount2.txt
apple mango rice human cat bus train apple cat dog bird cat mango sky grass ball
 cat rice police food tree leaf stone human

[cloudera@quickstart workspace]$ hadoop fs -put WordCount2.txt WordCount2.txt
```



```
[cloudera@quickstart workspace]$ hadoop jar WordCount2.jar WordCount2  WordCount2.txt WordCount2
22/06/12 12:56:09 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
22/06/12 12:56:09 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
22/06/12 12:56:10 INFO input.FileInputFormat: Total input paths to process : 1
22/06/12 12:56:10 INFO mapreduce.JobSubmitter: number of splits:1
22/06/12 12:56:10 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1655046020639_0003
22/06/12 12:56:11 INFO impl.YarnClientImpl: Submitted application application_1655046020639_0003
22/06/12 12:56:11 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1655046020639_0003/
22/06/12 12:56:11 INFO mapreduce.Job: Running job: job_1655046020639_0003
22/06/12 12:56:21 INFO mapreduce.Job: Job job_1655046020639_0003 running in uber mode : false
22/06/12 12:56:21 INFO mapreduce.Job:  map 0% reduce 0%
22/06/12 12:56:31 INFO mapreduce.Job:  map 100% reduce 0%
22/06/12 12:56:41 INFO mapreduce.Job:  map 100% reduce 100%
22/06/12 12:56:41 INFO mapreduce.Job: Job job_1655046020639_0003 completed successfully
22/06/12 12:56:42 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=274
                FILE: Number of bytes written=223329
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=249
                HDFS: Number of bytes written=157
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=7175
                Total time spent by all reduces in occupied slots (ms)=7307
                Total time spent by all map tasks (ms)=7175
                Total time spent by all reduce tasks (ms)=7307
                Total vcore-seconds taken by all map tasks=7175
                Total vcore-seconds taken by all reduce tasks=7307
                Total megabyte-seconds taken by all map tasks=7347200
                Total megabyte-seconds taken by all reduce tasks=7482368
        Map-Reduce Framework
                Map input records=1
                Map output records=24
                Map output bytes=220
                Map output materialized bytes=274
                Input split bytes=125
```



```
                Input split bytes=[illegible]
                Combine input records=0
                Combine output records=0
                Reduce input groups=17
                Reduce shuffle bytes=274
                Reduce input records=24
                Reduce output records=17
                Spilled Records=48
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=189
                CPU time spent (ms)=1680
                Physical memory (bytes) snapshot=345628672
                Virtual memory (bytes) snapshot=3008466944
                Total committed heap usage (bytes)=226365440
        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=124
        File Output Format Counters
                Bytes Written=157
[cloudera@quickstart workspace]$ hadoop fs -cat WordCount2/part-r-00000
APPLE   2.0
BALL    1.0
BIRD    1.0
BUS     1.0
CAT     4.0
DOG     1.0
FOOD    1.0
GRASS   1.0
HUMAN   2.0
LEAF    1.0
MANGO   2.0
POLICE  1.0
RICE    2.0
SKY     1.0
STONE   1.0
TRAIN   1.0
TREE    1.0
[cloudera@quickstart workspace]$ █
```

2. Learning Outcomes

From this I have learned, how to set up virtual machine and use Hadoop framework in Cloudera platform which is used for data analytics, data warehousing etc. Hadoop framework is used for the distributed processing of large data sets across the clusters of computers using simple programming models. From this lab, I have learned how to write a MapReduce program in Java language in Eclipse IDE and execute the program in Hadoop framework through Cloudera command line. By doing this lab I have known that how Hadoop framework run the MapReduce program by dividing the input into chunks, transforming the values of the chunks in key value pairs then the values are grouped by the keys and applying the reducing operation on the values of those keys.