Exercise 03: Stop Word Elimination using Map Reduce

Previous exercise described how to count repeated words in the input file. This exercise practice the students to do MapReduce process using word counting application with elimination words.

Prerequisites

Ensure that Hadoop is installed, configured and is running. More

details: Single Node Setup for first-time users.

Cluster Setup for large, distributed clusters.

Inputs and Outputs

i. Input file should be in :/wcsw/in00/

data.txt

Copy the content text from Shakespeare.txt, Which is attached in Google classroom.

sw.txt

Add following elimination words into sw.txt file.

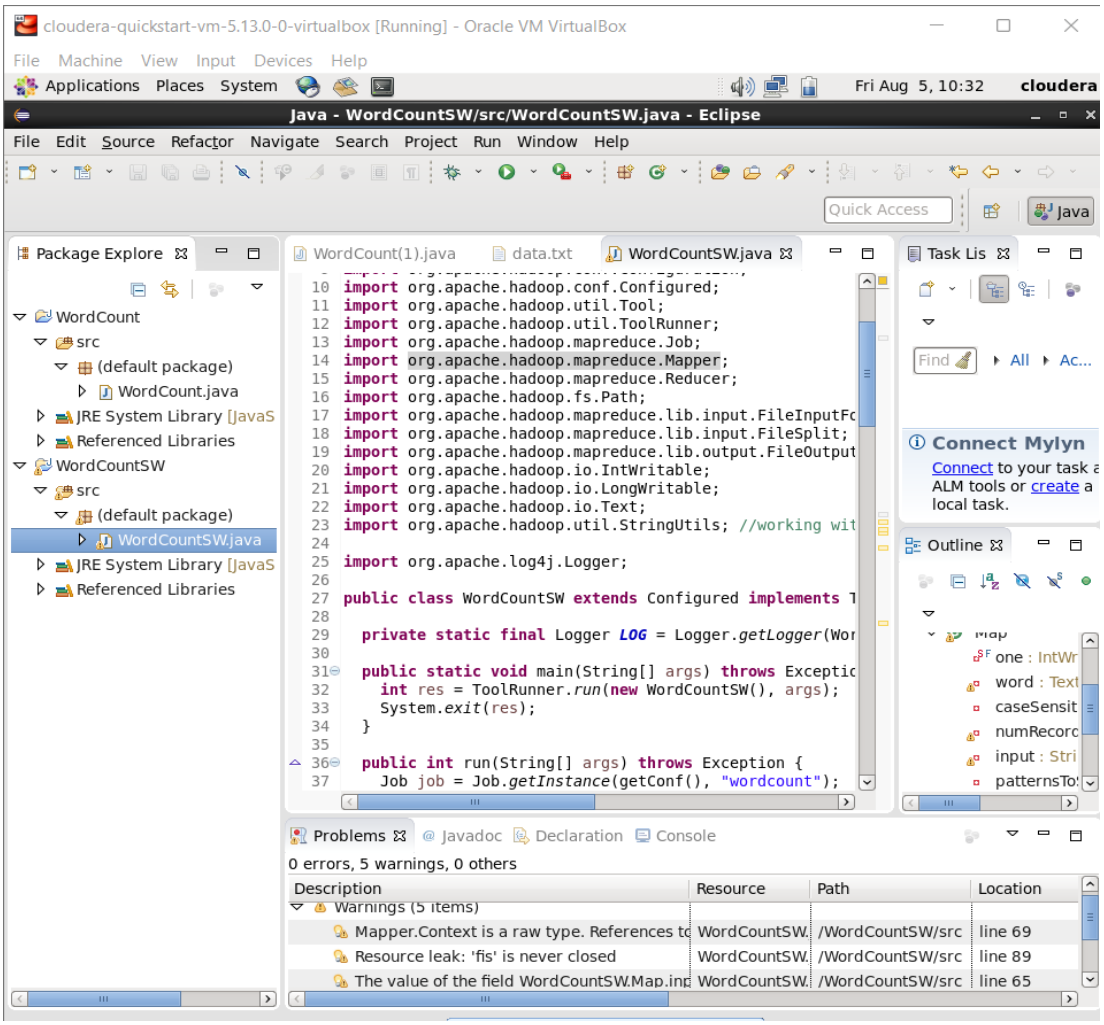
all
is
the
our
I
It

ii. Output file should be in /wcsw/out00/

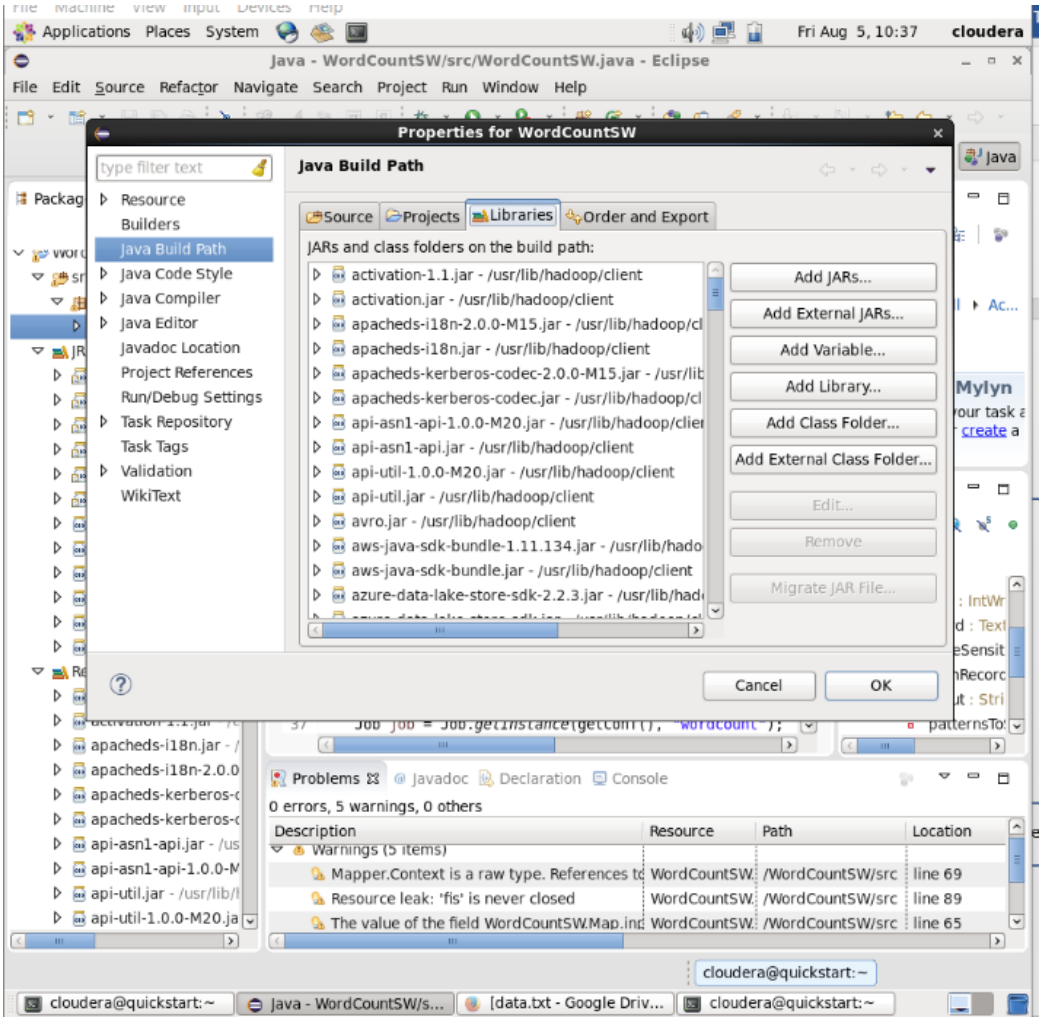
Step 1

Compile `WordCountSW.java` and create a WordCountSW.jar:

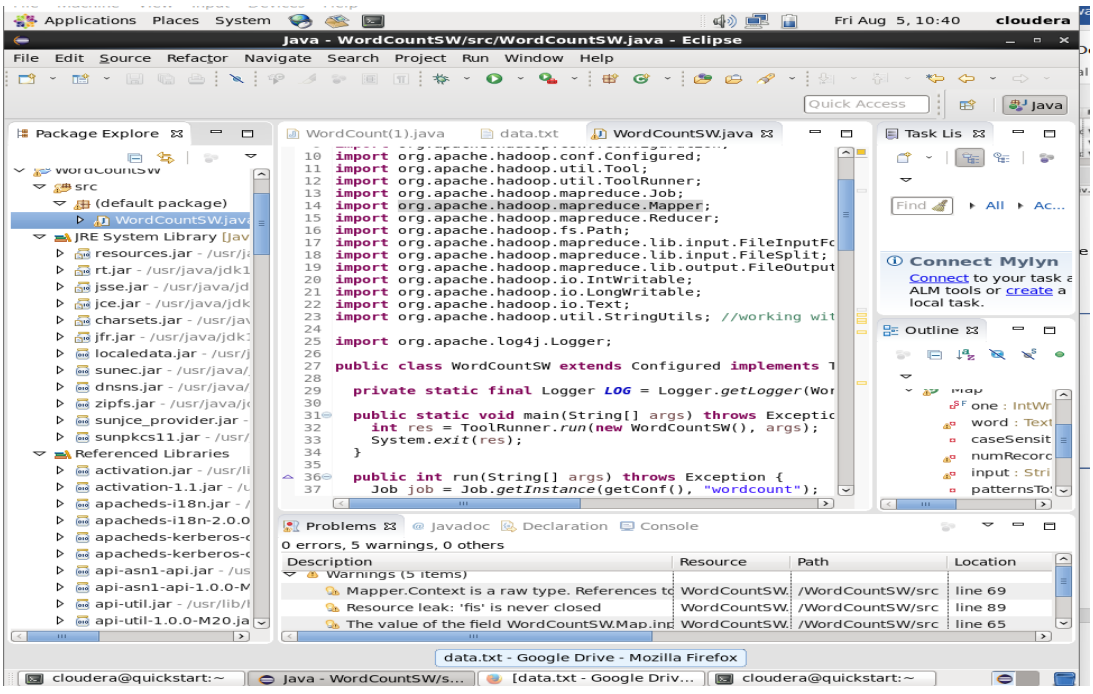
(i) Create WordCountSW.java project.


cloudera-quickstart-vm-5.13.0-0-virtualbox [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
Fri Aug 5, 10:32
cloudera
Java - WordCountSW/src/WordCountSW.java - Eclipse
File Edit Source Refactor Navigate Search Project Run Window Help
Quick Access
Java
Package Explore
WordCount
src
(default package)
WordCount.java
JRE System Library [JavaS
Referenced Libraries
WordCountSW
src
(default package)
WordCountSW.java
JRE System Library [JavaS
Referenced Libraries
WordCount(1).java
data.txt
WordCountSW.java
10 import org.apache.hadoop.conf.Configured;
11 import org.apache.hadoop.util.Tool;
12 import org.apache.hadoop.util.ToolRunner;
13 import org.apache.hadoop.mapreduce.Job;
14 import org.apache.hadoop.mapreduce.Mapper;
15 import org.apache.hadoop.mapreduce.Reducer;
16 import org.apache.hadoop.fs.Path;
17 import org.apache.hadoop.mapreduce.lib.input.FileInputFc
18 import org.apache.hadoop.mapreduce.lib.input.FileSplit;
19 import org.apache.hadoop.mapreduce.lib.output.FileOutput
20 import org.apache.hadoop.io.IntWritable;
21 import org.apache.hadoop.io.LongWritable;
22 import org.apache.hadoop.io.Text;
23 import org.apache.hadoop.util.StringUtils; //working wit
24
25 import org.apache.log4j.Logger;
26
27 public class WordCountSW extends Configured implements T
28
29 private static final Logger LOG = Logger.getLogger(Wor
30
31 public static void main(String[] args) throws Exceptic
32 int res = ToolRunner.run(new WordCountSW(), args);
33 System.exit(res);
34 }
35
36 public int run(String[] args) throws Exception {
37 Job job = Job.getInstance(getConf(), "wordcount");
Task Lis
Find
All
Ac...
Connect Mylyn
Connect to your task a
ALM tools or create a
local task.
Outline
one : IntWr
word : Text
caseSensit
numRecorc
input : Stri
patternsTo
Problems
Javadoc
Declaration
Console
0 errors, 5 warnings, 0 others
Description
Resource
Path
Location
Warnings (5 items)
Mapper.Context is a raw type. References to
WordCountSW.
/WordCountSW/src
line 69
Resource leak: 'fis' is never closed
WordCountSW.
/WordCountSW/src
line 89
The value of the field WordCountSW.Map.inp
WordCountSW.
/WordCountSW/src
line 65

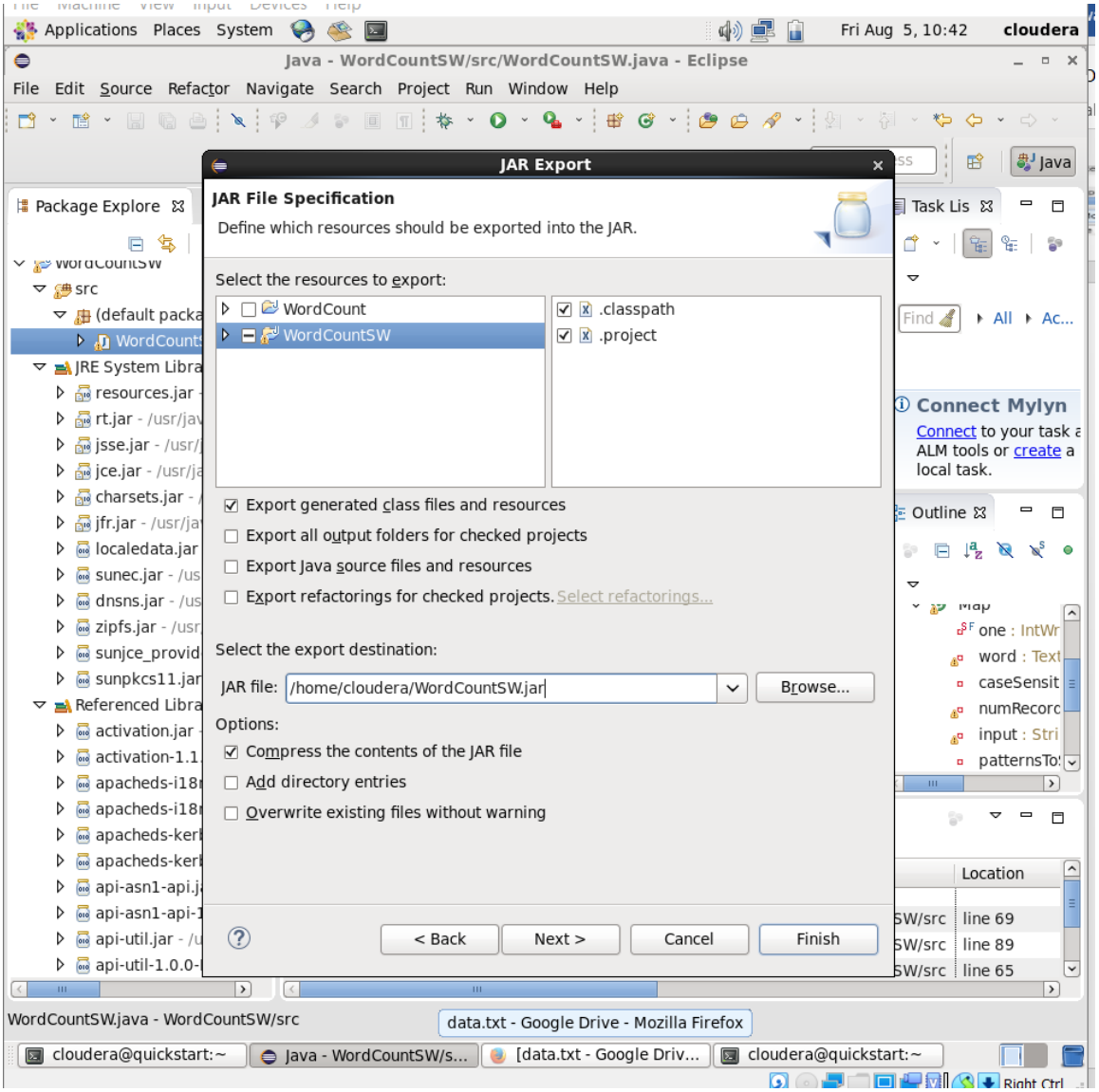
(ii) Import external .jar files



(iii) Create WordCount class file using Google classroom attached WordCount.java file.



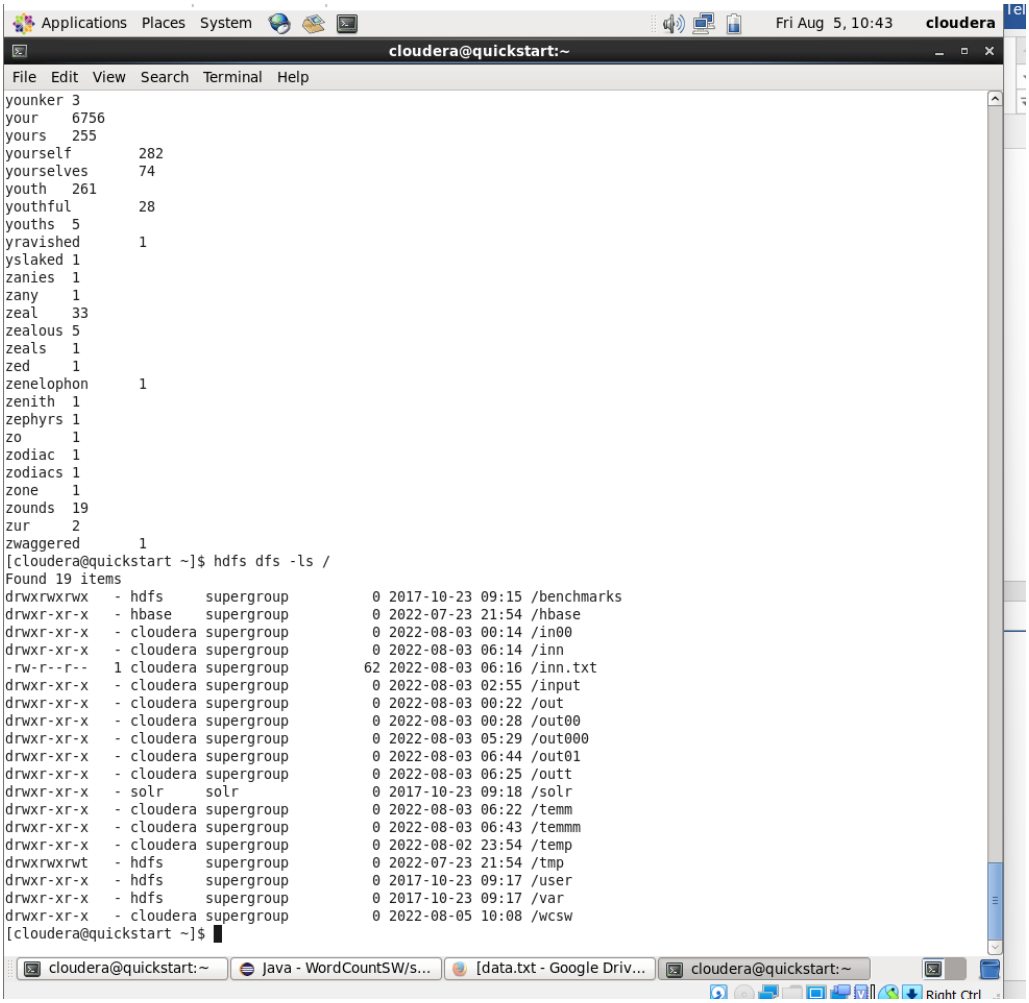
(iv) Create WordCountSW.jar file



Step 2

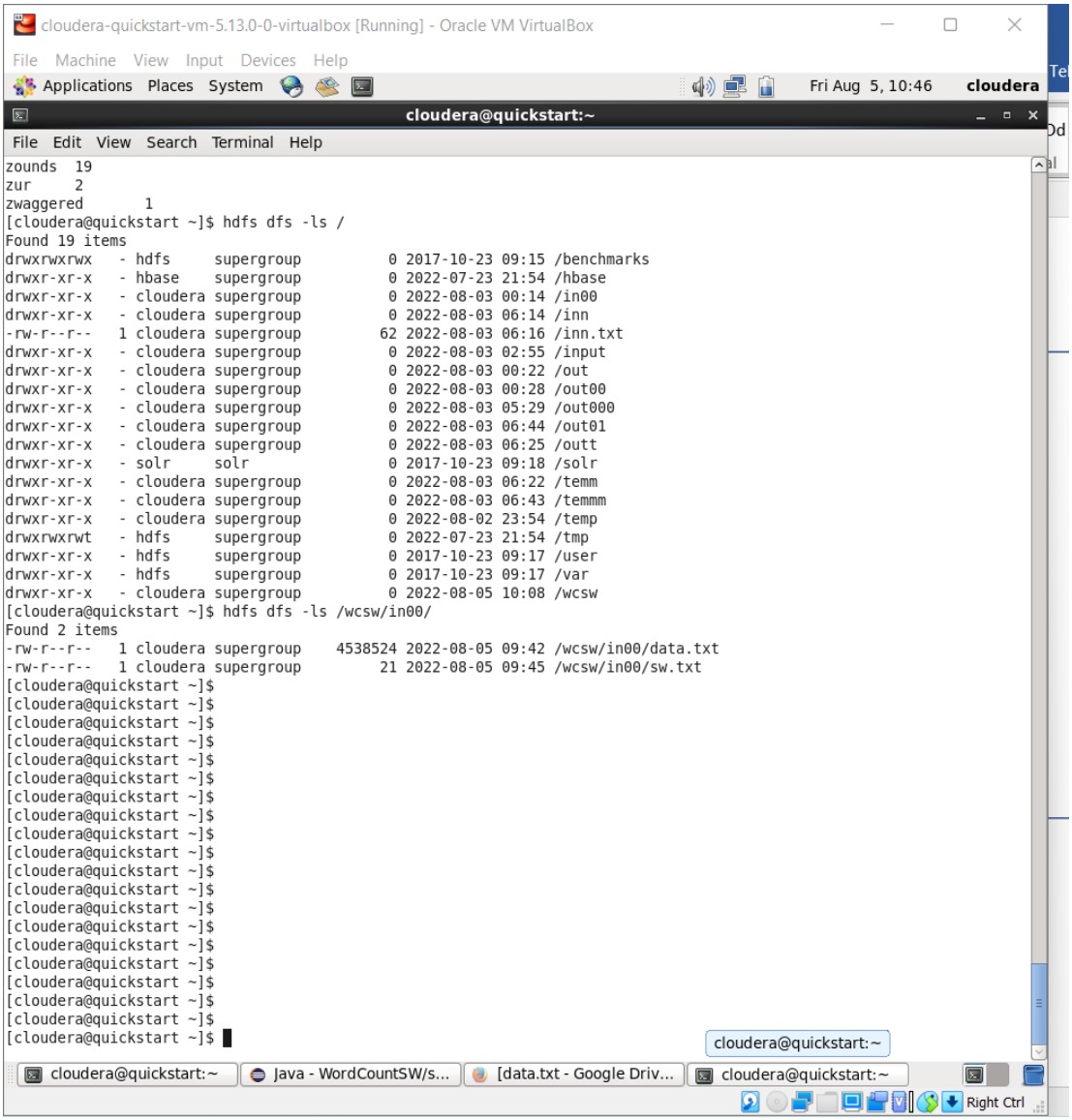
Create following folders in HDFS:

- `/wcsw/in00` - input directory in HDFS
- `/wcsw/out00` - output directory in HDFS


Applications Places System
Fri Aug 5, 10:43
cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
younker 3
your 6756
yours 255
yourself 282
yourselves 74
youth 261
youthful 28
youths 5
yravished 1
yslaked 1
zanies 1
zany 1
zeal 33
zealous 5
zeals 1
zed 1
zenelophon 1
zenith 1
zephyrs 1
zo 1
zodiac 1
zodiacs 1
zone 1
zounds 19
zur 2
zwaggered 1
[cloudera@quickstart ~]$ hdfs dfs -ls /
Found 19 items
drwxrwxrwx - hdfs supergroup 0 2017-10-23 09:15 /benchmarks
drwxr-xr-x - hbase supergroup 0 2022-07-23 21:54 /hbase
drwxr-xr-x - cloudera supergroup 0 2022-08-03 00:14 /in00
drwxr-xr-x - cloudera supergroup 0 2022-08-03 06:14 /inn
-rw-r--r-- 1 cloudera supergroup 62 2022-08-03 06:16 /inn.txt
drwxr-xr-x - cloudera supergroup 0 2022-08-03 02:55 /input
drwxr-xr-x - cloudera supergroup 0 2022-08-03 00:22 /out
drwxr-xr-x - cloudera supergroup 0 2022-08-03 00:28 /out00
drwxr-xr-x - cloudera supergroup 0 2022-08-03 05:29 /out000
drwxr-xr-x - cloudera supergroup 0 2022-08-03 06:44 /out01
drwxr-xr-x - cloudera supergroup 0 2022-08-03 06:25 /outt
drwxr-xr-x - solr solr 0 2017-10-23 09:18 /solr
drwxr-xr-x - cloudera supergroup 0 2022-08-03 06:22 /temm
drwxr-xr-x - cloudera supergroup 0 2022-08-03 06:43 /temmm
drwxr-xr-x - cloudera supergroup 0 2022-08-02 23:54 /temp
drwxrwxrwt - hdfs supergroup 0 2022-07-23 21:54 /tmp
drwxr-xr-x - hdfs supergroup 0 2017-10-23 09:17 /user
drwxr-xr-x - hdfs supergroup 0 2017-10-23 09:17 /var
drwxr-xr-x - cloudera supergroup 0 2022-08-05 10:08 /wcsw
[cloudera@quickstart ~]$
cloudera@quickstart:~
Java - WordCountSW/s...
[data.txt - Google Driv...
cloudera@quickstart:~
Right Ctrl

Step 3

Create and copy data text-files into input folder:



Step-4

Create and copy sw text-files into input folder:

```
drwxr-xr-x   - cloudera supergroup          0 2022-08-05 10:08 /wcsw
[cloudera@quickstart ~]$ hdfs dfs -ls /wcsw/in00/
Found 2 items
-rw-r--r--   1 cloudera supergroup    4538524 2022-08-05 09:42 /wcsw/in00/data.txt
-rw-r--r--   1 cloudera supergroup         21 2022-08-05 09:45 /wcsw/in00/sw.txt
[cloudera@quickstart ~]$
```

[cloudera@quickstart ~]$ hdfs dfs -ls /wcsw/in00/

Found 2 items

-rw-r--r-- 1 cloudera supergroup 3309 2021-08-24 07:00 /wcsw/in00/data.txt

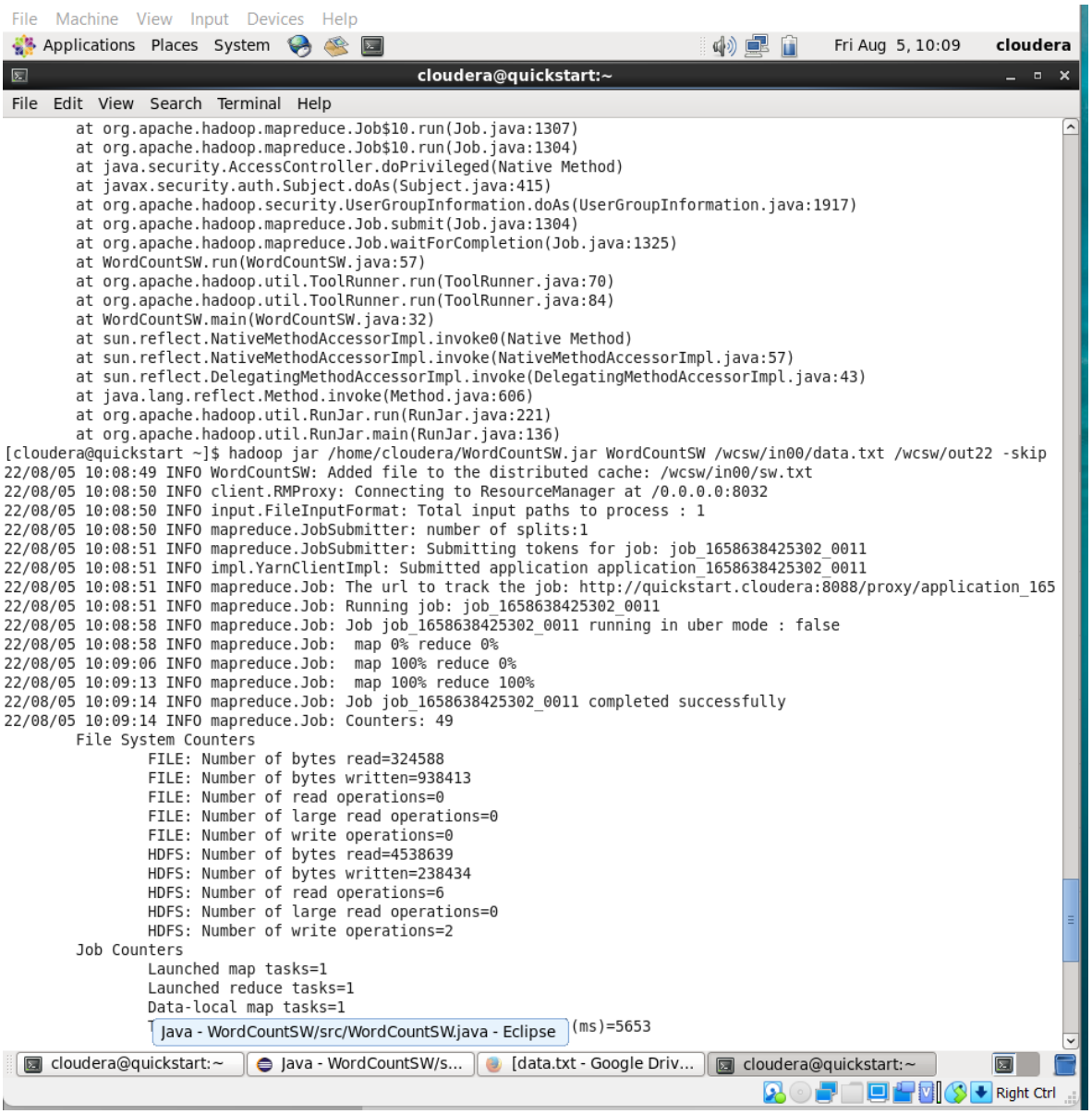
-rw-r--r-- 1 cloudera supergroup 15 2021-08-24 07:02 /wcsw/in00/sw.txt

Step 5

Run the MapReduce application with skip option:

[cloudera@quickstart ~]$ hadoop jar /home/cloudera/WordCountSW.jar /wcsw/in00/data.txt /wcsw/out00/ -skip /wcsw/in00/sw.txt

Show MapReduce Framework



Step 6

Output:

[cloudera@quickstart ~]$ hdfs dfs -ls /wcsw/out00/

Found 2 items

-rw-r--r-- 1 cloudera supergroup 0 2021-08-24 07:05 /wcsw/out00/_SUCCESS

-rw-r--r-- 1 cloudera supergroup 2384 2021-08-24 07:05 /wcsw/out00/part-r-

00000 [cloudera@quickstart ~]$ hdfs dfs -cat /wcsw/out00/part-r-00000