

215229140

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

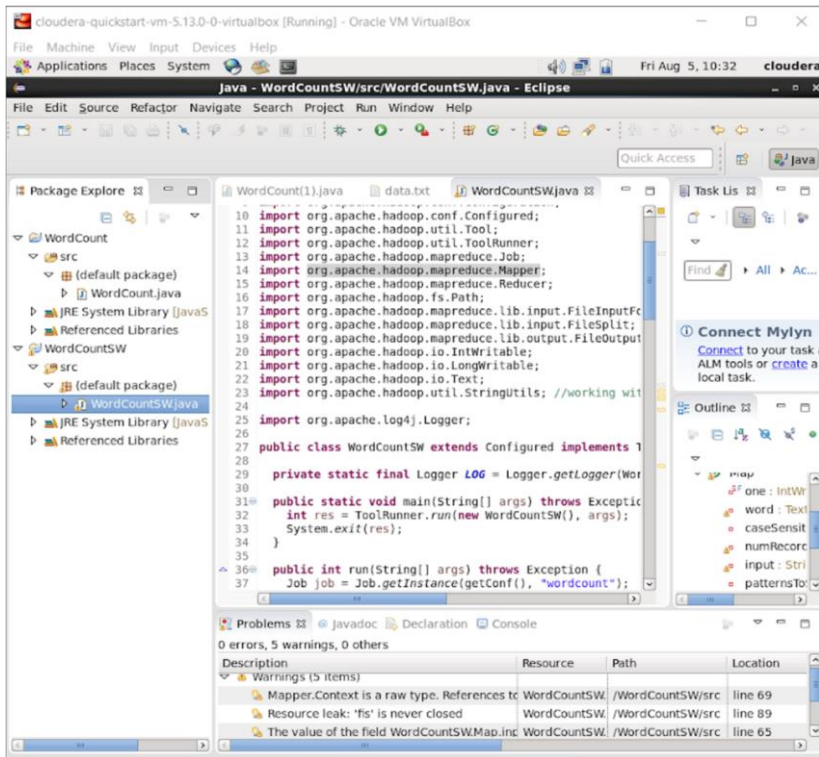
It

ii. **Output file should be in /wcsW/out00/**

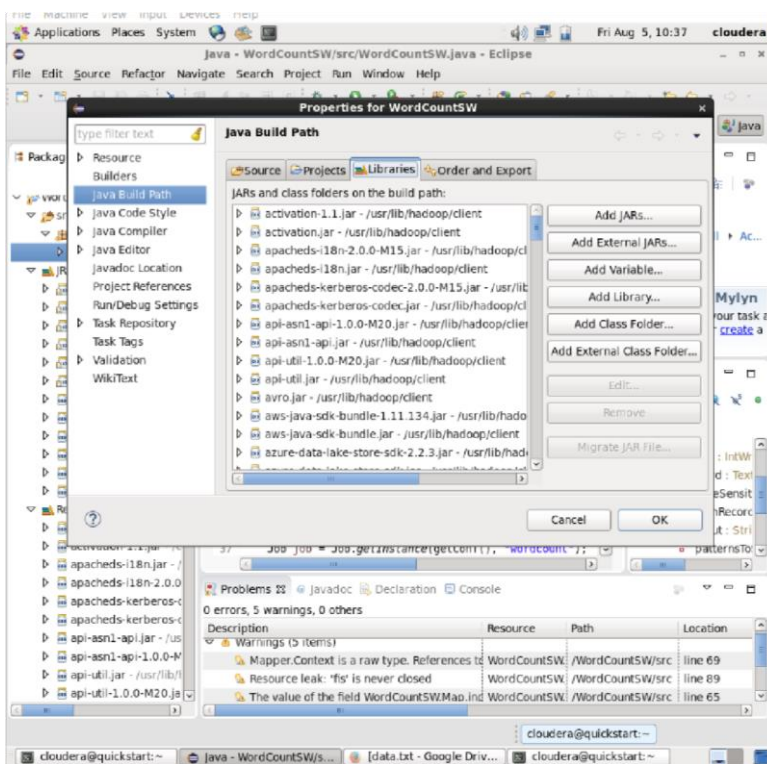
Step 1

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

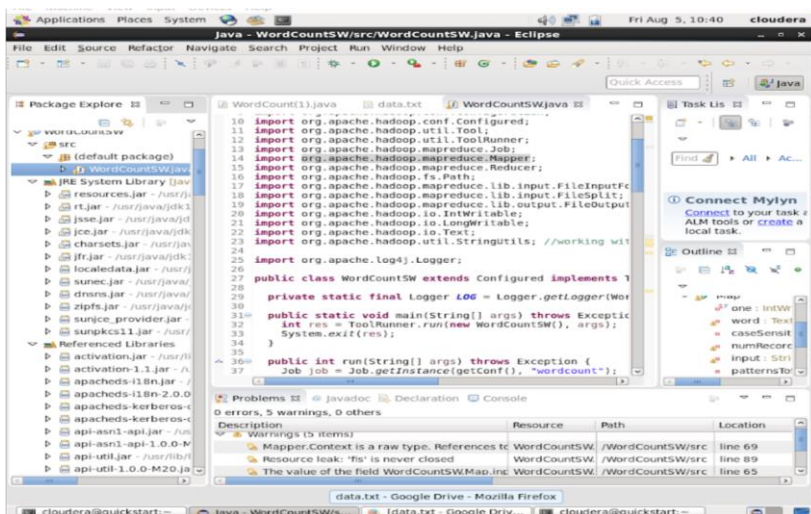
(i) Create `WordCountSW.java` project.



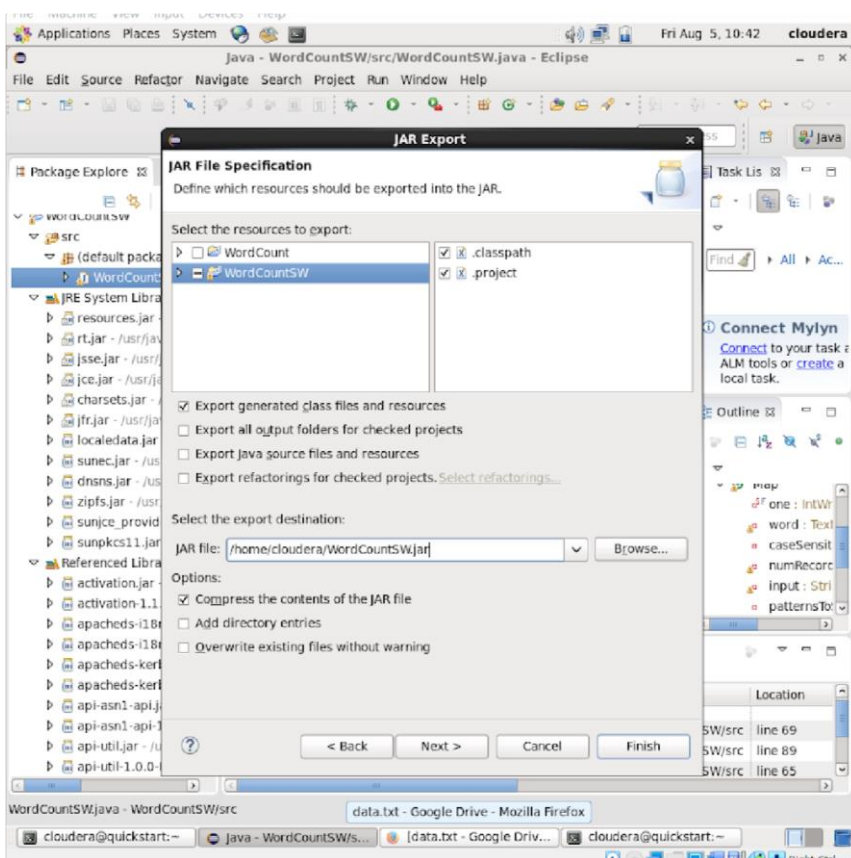
(ii) Import external .jar files



(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 cloudera@quickstart:~
File Edit View Search Terminal Help
younger 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
ze 1
zodiac 1
zodiacs 1
zone 1
zounds 19
zur 2
zwaggered 1
[cloudera@quickstart ~]$ hdfs dfs -ls /
Found 19 items
drwxr-xr-x - 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 /temm
drwxr-xr-x - cloudera supergroup 0 2022-08-02 23:54 /temp
drwxr-xr-x - hdfs supergroup 0 2022-07-23 21:54 /tap
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 /wcsv
[cloudera@quickstart ~]$
```

Step 3

Create and copy data text-files into input folder:

[illegible]

Step-4

Step 4
Create and copy sw text-files into input folder:

```

drwxr-xr-x - cloudera supergroup          0 2022-08-05 10:00 /wcs
[cloudera@quickstart ~]$ hdfs dfs -ls /wcs/in00/
Found 2 items
-rw-r--r-- 1 cloudera supergroup      4538524 2022-08-05 09:42 /wcs/in00/data.txt
-rw-r--r-- 1 cloudera supergroup       21 2022-08-05 09:45 /wcs/in00/sw.txt
[cloudera@quickstart ~]$

```

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

Found 2 items

```
- rw-r--r-- 1 cloudera supergroup 3309 2021-08-24 07:00 /wcsww/in00/data.txt
- rw-r--r-- 1 cloudera supergroup 15 2021-08-24 07:02 /wcsww/in00/sw.txt
```

Step 5

Run the MapReduce application with skip option:

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

Show MapReduce Framework

```
File Machine View Input Devices Help
Applications Places System Fri Aug 5, 10:09 cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
at org.apache.hadoop.mapreduce.Job$10.run(Job.java:1367)
at org.apache.hadoop.mapreduce.Job$10.run(Job.java:1364)
at java.security.AccessController.doPrivileged(Native Method)
at javax.security.auth.Subject.doAs(Subject.java:415)
at org.apache.hadoop.security.UserGroupInformation.doAs(UserGroupInformation.java:1917)
at org.apache.hadoop.mapreduce.Job.submit(Job.java:1364)
at org.apache.hadoop.mapreduce.Job.waitForCompletion(Job.java:1325)
at WordCountSW.run(WordCountSW.java:57)
at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:79)
at org.apache.hadoop.util.ToolRunner.run(ToolRunner.java:84)
at WordCountSW.main(WordCountSW.java:32)
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:606)
at org.apache.hadoop.util.RunJar.run(RunJar.java:221)
at org.apache.hadoop.util.RunJar.main(RunJar.java:136)
[cloudera@quickstart ~]$ hadoop jar /home/cloudera/WordCountSW.jar WordCountSW /wscw/in00/data.txt /wscw/out22 -skip
22/08/05 10:08:49 INFO WordCountSW: Added file to the distributed cache: /wscw/in00/sw.txt
22/08/05 10:08:50 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
22/08/05 10:08:50 INFO input.FileInputFormat: Total input paths to process : 1
22/08/05 10:08:50 INFO mapreduce.JobSubmitter: number of splits:1
22/08/05 10:08:51 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1658638425302_0011
22/08/05 10:08:51 INFO impl.YarnClientImpl: Submitted application application_1658638425302_0011
22/08/05 10:08:51 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_165
22/08/05 10:08:51 INFO mapreduce.Job: Running job: job_1658638425302_0011
22/08/05 10:08:58 INFO mapreduce.Job: Job job_1658638425302_0011 running in uber mode : false
22/08/05 10:08:58 INFO mapreduce.Job: map 0% reduce 0%
22/08/05 10:09:00 INFO mapreduce.Job: map 100% reduce 0%
22/08/05 10:09:13 INFO mapreduce.Job: map 100% reduce 100%
22/08/05 10:09:14 INFO mapreduce.Job: Job job_1658638425302_0011 completed successfully
22/08/05 10:09:14 INFO mapreduce.Job: Counters: 49
File System Counters
  FILE: Number of bytes read=324588
  FILE: Number of bytes written=938413
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=4538639
  HDFS: Number of bytes written=238434
  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
  Java - WordCountSW/src/WordCountSW.java - Eclipse (ms)=5653
```