Map Reduce Tutorial 3

Venkatesh Vinayakarao
venkateshv@cmi.ac.in
http://vvtesh.co.in

Chennai Mathematical Institute

Map Reduce

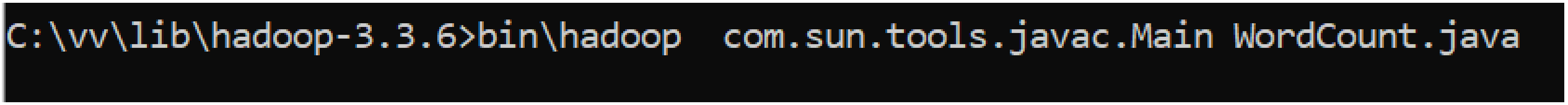
- Our Hadoop installation is quite old! It works with the following configuration:
    - hadoop-2.6.0
    - java version "1.7.0_67“
- Start Cloudera Hadoop
    - Start Docker Desktop

    -
      ```
      docker run --hostname=quickstart.cloudera --privileged=true -t -i --publish-all=true -p 8888:8888 -p 8080:80 -p 50070:50070 -p 8088:8088  -p 50075:50075 -p 8032:8032 -p 8042:8042 -p 19888:19888 cloudera/quickstart /usr/bin/docker-quickstart
      ```

Find detailed documentation here

Compile

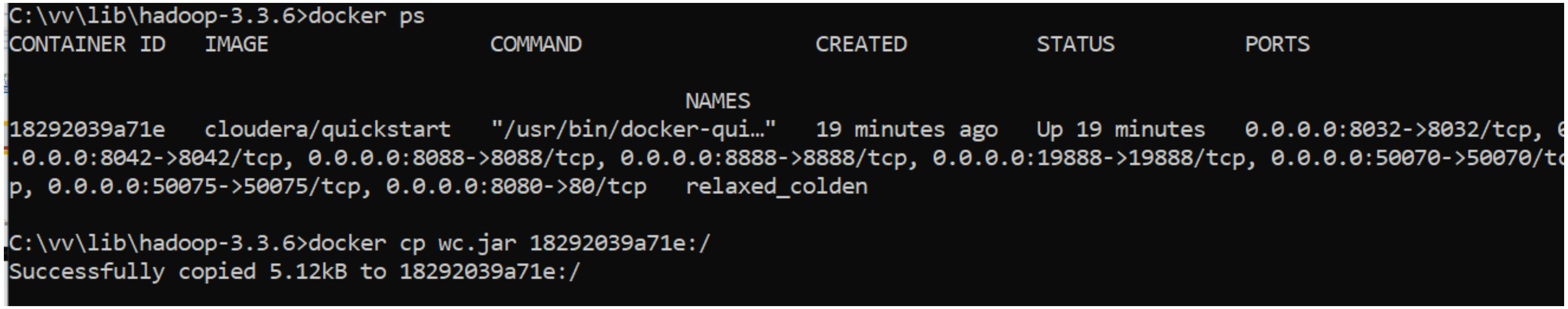
- Compile WordCount.Java
  - It is already provided to you on the course page. You may also get it from https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html
  - bin/hadoop com.sun.tools.javac.Main WordCount.java
  - jar cf wc.jar WordCount*.class



```
C:\vv\lib\hadoop-3.3.6>bin\hadoop  com.sun.tools.javac.Main WordCount.java
```

Move the jar

- Copy this wc.jar to Hadoop

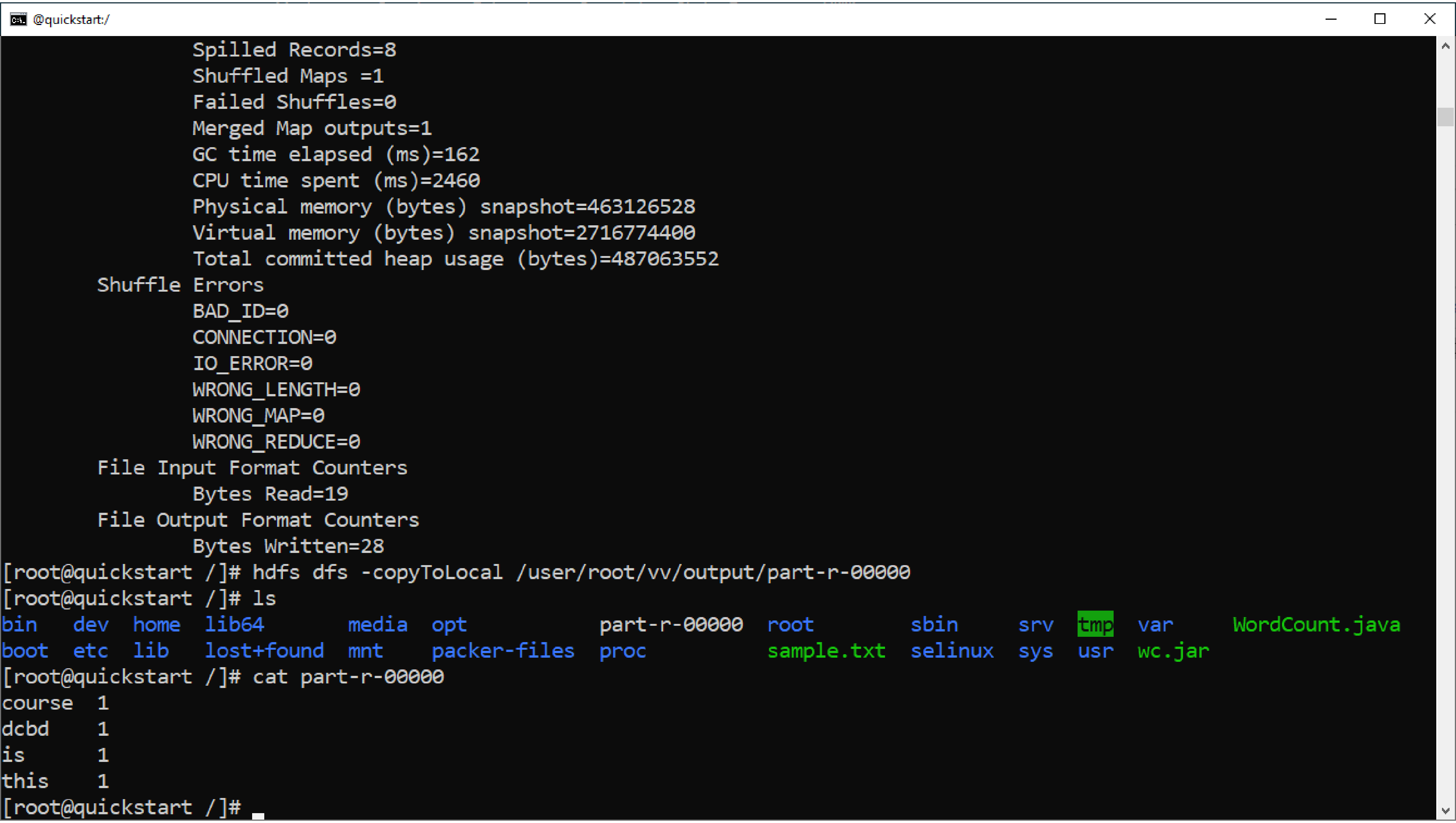


```
C:\vv\lib\hadoop-3.3.6>docker ps
CONTAINER ID   IMAGE                 COMMAND                  CREATED          STATUS          PORTS

                                                    NAMES
18292039a71e   cloudera/quickstart   "/usr/bin/docker-qui…"   19 minutes ago   Up 19 minutes   0.0.0.0:8032->8032/tcp, 0
.0.0.0:8042->8042/tcp, 0.0.0.0:8088->8088/tcp, 0.0.0.0:8888->8888/tcp, 0.0.0.0:19888->19888/tcp, 0.0.0.0:50070->50070/tc
p, 0.0.0.0:50075->50075/tcp, 0.0.0.0:8080->80/tcp   relaxed_colden

C:\vv\lib\hadoop-3.3.6>docker cp wc.jar 18292039a71e:/
Successfully copied 5.12kB to 18292039a71e:/
```

Run the Program

- Switch to [root@quickstart /]#
- Setup the input
  - hdfs dfs -mkdir /user/root/vv/
  - hdfs dfs -mkdir /user/root/vv/input/
  - docker cp sample.txt 18292039a71e:/
  - hdfs dfs -put sample.txt /user/root/vv/input/
  - hdfs dfs -rm -r /user/root/vv/output
- Now, we are ready to run the MR job
  - hadoop jar wc.jar WordCount /user/root/vv/input /user/root/vv/output
- Output is in /user/root/vv/output folder

Done!



```
                Spilled Records=8
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=162
                CPU time spent (ms)=2460
                Physical memory (bytes) snapshot=463126528
                Virtual memory (bytes) snapshot=2716774400
                Total committed heap usage (bytes)=487063552
        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=19
        File Output Format Counters
                Bytes Written=28
[root@quickstart /]# hdfs dfs -copyToLocal /user/root/vv/output/part-r-00000
[root@quickstart /]# ls
bin   dev  home  lib64       media  opt           part-r-00000  root        sbin     srv  tmp  var     WordCount.java
boot  etc  lib   lost+found  mnt    packer-files  proc          sample.txt  selinux  sys  usr  wc.jar
[root@quickstart /]# cat part-r-00000
course  1
dcbd    1
is      1
this    1
[root@quickstart /]# _
```

Thank You