

Roll No: J016 (Avneesh Dubey) & J054(Aayush Talekar)

Aim: Word count using MapReduce Java

Objectives:

1. To run Java command.
2. Copy Data file from Local to HDFS.
3. Generate a Word count query.
4. Display Word count of the file.

Code & Output:

1. Create a Text file in Local

2. Transfer to Hdfs using:

```
hdfs dfs -put random.txt random4.txt
```

3. Create a Jar file(Code supplied by Sir) in eclipse with 3 classes:

- a. WCDriver
- b. WCMapper
- c. WCReducer

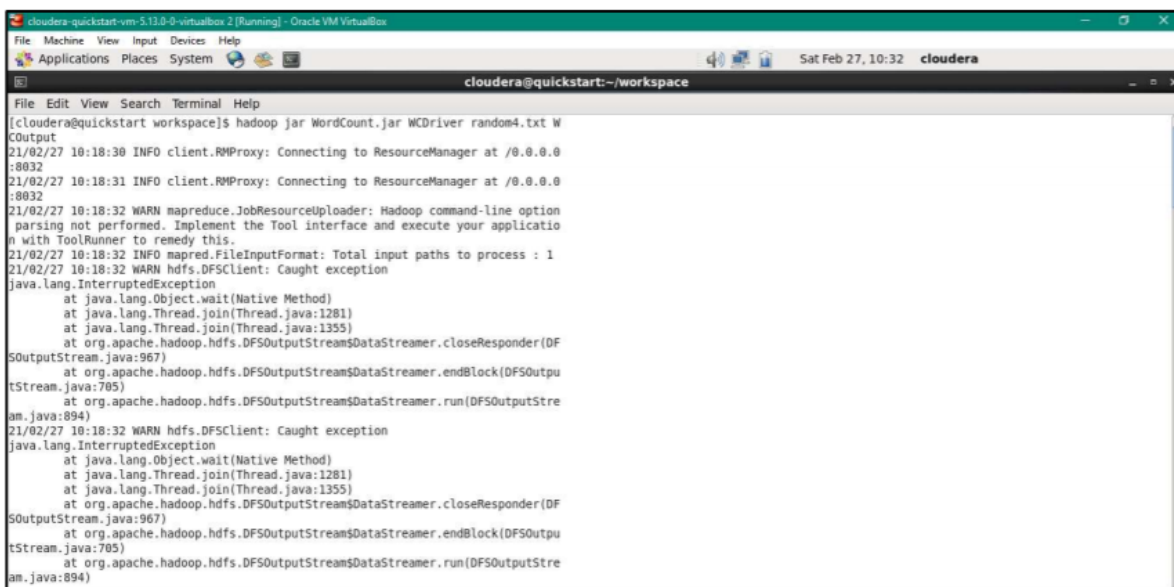
4. In Building add 2 external JARs

- a. /usr/lib/hadoop-0.20-mapreduce/hadoop-core-2.6.0-mr1-cdh5.13.0.jar
- b. /usr/lib/hadoop/hadoop-common-2.6.0-cdh5.13.0.jar

5. Now build the Jar file.

6. Now change directory to workspace and execute below command

a. `hadoop jar WordCount.jar WCDriver WCFFile.txt WCOOutput`



```
cloudera@quickstart:~/workspace$ hadoop jar WordCount.jar WCDriver random4.txt WCOOutput
21/02/27 10:18:30 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
21/02/27 10:18:31 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
21/02/27 10:18:32 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
21/02/27 10:18:32 INFO mapred.FileInputFormat: Total input paths to process : 1
21/02/27 10:18:32 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
21/02/27 10:18:32 WARN hdfs.DFSClient: Caught exception
java.lang.InterruptedException
    at java.lang.Object.wait(Native Method)
    at java.lang.Thread.join(Thread.java:1281)
    at java.lang.Thread.join(Thread.java:1355)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.closeResponder(DFSOutputStream.java:967)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.endBlock(DFSOutputStream.java:705)
    at org.apache.hadoop.hdfs.DFSOutputStream$DataStreamer.run(DFSOutputStream.java:894)
```

```
cloudera@quickstart:~/workspace$ hcat -i /workspace/quickstart-vm-5.11.0-0-vmtoolsd-2 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~/workspace
File Edit View Search Terminal Help
21/02/27 10:18:32 INFO mapreduce.JobSubmitter: number of splits=2
21/02/27 10:18:33 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_16
14417601119_0007
21/02/27 10:18:33 INFO impl.YarnClientImpl: Submitted application application_16
14417601119_0007
21/02/27 10:18:33 INFO mapreduce.Job: The url to track the job: http://quickstar
t.cloudera:8088/proxy/application_1614417601119_0007/
21/02/27 10:18:33 INFO mapreduce.Job: Running job: job_1614417601119_0007
21/02/27 10:18:49 INFO mapreduce.Job: Job job_1614417601119_0007 running in uber
mode : false
21/02/27 10:18:49 INFO mapreduce.Job: map 0% reduce 0%
21/02/27 10:19:11 INFO mapreduce.Job: map 50% reduce 0%
21/02/27 10:19:12 INFO mapreduce.Job: map 100% reduce 0%
21/02/27 10:19:25 INFO mapreduce.Job: map 100% reduce 100%
21/02/27 10:19:26 INFO mapreduce.Job: Job job_1614417601119_0007 completed succe
ssfully
21/02/27 10:19:26 INFO mapreduce.Job: Counters: 50
File System Counters
  FILE: Number of bytes read=1237
  FILE: Number of bytes written=433093
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=1884
  HDFS: Number of bytes written=390
  HDFS: Number of read operations=9
  HDFS: Number of large read operations=8
  HDFS: Number of write operations=2
Job Counters
  Killed map tasks=1
  Launched map tasks=2
  Launched reduce tasks=1
```

```
cloudera@quickstart:~/workspace$ hcat -i /workspace/quickstart-vm-5.11.0-0-vmtoolsd-2 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System
cloudera@quickstart:~/workspace
File Edit View Search Terminal Help
Job Counters
  Killed map tasks=1
  Launched map tasks=2
  Launched reduce tasks=1
  Data-local map tasks=2
  Total time spent by all maps in occupied slots (ms)=36274
  Total time spent by all reduces in occupied slots (ms)=11461
  Total time spent by all map tasks (ms)=36274
  Total time spent by all reduce tasks (ms)=11461
  Total vcore-milliseconds taken by all map tasks=36274
  Total vcore-milliseconds taken by all reduce tasks=11461
  Total megabyte-milliseconds taken by all map tasks=37144576
  Total megabyte-milliseconds taken by all reduce tasks=11736064
Map-Reduce Framework
  Map input records=2
  Map output records=109
  Map output bytes=1013
  Map output materialized bytes=1243
  Input split bytes=218
  Combine input records=0
  Combine output records=0
  Reduce input groups=72
  Reduce shuffle bytes=1243
  Reduce input records=109
  Reduce output records=72
  Spilled Records=218
  Shuffled Maps =2
  Failed Shuffles=0
  Merged Map outputs=2
  GC time elapsed (ms)=642
  CPU time spent (ms)=3140
  Physical memory (bytes) snapshot=686768320
Total committed heap usage (bytes)=391979008
Shuffle Errors
  BAD_ID=0
  CONNECTION=0
  ID_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=966
File Output Format Counters
```

b. `hadoop fs -cat WCOOutput/part-00000`

```
cloudera-quickstart-vm-5.11.0-0-virtualbox 2 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System Sat Feb 27, 10:31 cloudera
cloudera@quickstart:~/workspace
File Edit View Search Terminal Help
$
[cloudera@quickstart workspace]$ hadoop fs -cat MCOutput/part-000009
.      13
2021   1
27th   1
BDT    1
Bye    1
Career 1
Data   4
Databases, 1
Excel  1
February 1
Forms  1
Google 1
It      1
It's   1
J078   1
My      1
People 1
Science 1
Scientist 1
Shah   2
Skill  1
Spreadsheet, 1
Thank  1
This   2
To     1
Ujwal  2
a      4
algorithms 1
an     2
and    4
```

```
cloudera-quickstart-vm-5.11.0-0-virtualbox 2 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
Applications Places System Sat Feb 27, 10:32 cloudera
cloudera@quickstart:~/workspace
File Edit View Search Terminal Help
data 3
etc 1
extract 1
field 1
from 2
gain 1
has 1
hello 1
hidden 1
in 1
information 2
insights 1
inter-disciplinary 1
is 6
knowledge 1
lab 1
methods, 1
name 1
no: 1
of 3
processes, 1
random 1
roll 1
science 2
scientific 1
scientist 1
study 1
studying 1
systems 1
that 1
the 1
to 1
```