

Practical 2: Map-Reduce Program for WordCount Problem

Commands:

```
[cloudera@quickstart ~]$ hdfs dfs -ls /
[cloudera@quickstart ~]$ sudo -u hdfs hadoop fs -mkdir /inputdirectory
[cloudera@quickstart ~]$ hdfs dfs -ls /
[cloudera@quickstart ~]$ cat>/home/cloudera/processfile.txt
[cloudera@quickstart ~]$ sudo -u hdfs hadoop fs -put
/home/cloudera/processfile.txt /inputdirectory
[cloudera@quickstart ~]$ hdfs dfs -ls /inputdirectory
[cloudera@quickstart ~]$ hadoop jar /home/cloudera/WordCount.jar WordCount
/inputdirectory/processfile.txt /out1
[cloudera@quickstart ~]$ hdfs dfs -ls /out1
[cloudera@quickstart ~]$ hdfs dfs -cat /out1/part-r-00000
```

OUTPUT

```
[cloudera@quickstart ~]$ hdfs dfs -ls /
sudo -u Found 6 items
drwxr-xr-x - hdfs supergroup 0 2017-10-23 10:29 /benchmarks
drwxr-xr-x - hbase supergroup 0 2023-03-20 05:39 /hbase
drwxr-xr-x - solr solr 0 2017-10-23 10:32 /solr
drwxr-xrwt - hdfs supergroup 0 2023-03-20 04:38 /tmp
drwxr-xr-x - hdfs supergroup 0 2017-10-23 10:31 /user
drwxr-xr-x - hdfs supergroup 0 2017-10-23 10:31 /var
[cloudera@quickstart ~]$ sudo -u hdfs hadoop fs -mkdir /inputdirectory
^[[A[cloudera@quickstart hdfs dfs -ls /
Found 7 items
drwxr-xr-x - hdfs supergroup 0 2017-10-23 10:29 /benchmarks
drwxr-xr-x - hbase supergroup 0 2023-03-20 05:39 /hbase
drwxr-xr-x - hdfs supergroup 0 2023-03-20 05:52 /inputdirectory
drwxr-xr-x - solr solr 0 2017-10-23 10:32 /solr
drwxr-xrwt - hdfs supergroup 0 2023-03-20 04:38 /tmp
drwxr-xr-x - hdfs supergroup 0 2017-10-23 10:31 /user
drwxr-xr-x - hdfs supergroup 0 2017-10-23 10:31 /var
[cloudera@quickstart ~]$ cat>/home/cloudera/processfile.txt
Hii How are you Hii i am fine^C
[cloudera@quickstart ~]$ sudo -u hdfs hadoop fs -put /home/cloudera/processfile.
txt /inputdirectory
^[[A[cloudera@quickstart hdfs -ls /^C
[cloudera@quickstart ~]$ hdfs dfs -ls /inputdirectory
Found 1 items
-rw-r--r-- 1 hdfs supergroup 0 2023-03-20 05:53 /inputdirectory/proce
ssfile.txt
[cloudera@quickstart ~]$ hadoop jar /home/cloudera/WordCount.jar WordCount /inpu
tdirectory/processfile.txt /out1
23/03/20 06:00:28 INFO client.RMProxy: Connecting to ResourceManager at /0.0.0.0:8032
23/03/20 06:00:28 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
23/03/20 06:00:29 INFO input.FileInputFormat: Total input paths to process : 1
23/03/20 06:00:29 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)
```

```
Applications Places System cloudera@quickstart:~
File Edit View Search Terminal Help
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)
23/03/20 06:00:29 INFO mapreduce.JobSubmitter: number of splits:1
23/03/20 06:00:29 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1679315869981_0001
23/03/20 06:00:30 INFO impl.YarnClientImpl: Submitted application application_1679315869981_0001
23/03/20 06:00:31 INFO mapreduce.Job: The url to track the job: http://quickstart.cloudera:8088/proxy/application_1679315869981_0001/
23/03/20 06:00:31 INFO mapreduce.Job: Running job: job_1679315869981_0001
23/03/20 06:00:45 INFO mapreduce.Job: Job job_1679315869981_0001 running in uber mode : false
23/03/20 06:00:45 INFO mapreduce.Job: map 0% reduce 0%
23/03/20 06:00:58 INFO mapreduce.Job: map 100% reduce 0%
hdfs dfs -ls /out123/03/20 06:01:07 INFO mapreduce.Job: map 100% reduce 100%
23/03/20 06:01:07 INFO mapreduce.Job: Job job_1679315869981_0001 completed successfully
23/03/20 06:01:07 INFO mapreduce.Job: Counters: 49
File System Counters
  FILE: Number of bytes read=6
  FILE: Number of bytes written=286727
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=127
  HDFS: Number of bytes written=0
  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
  Other local map tasks=1
  Total time spent by all maps in occupied slots (ms)=9891
  Total time spent by all reduces in occupied slots (ms)=6400
  Total time spent by all map tasks (ms)=9891
  Total time spent by all reduce tasks (ms)=6400
  Total vcore-milliseconds taken by all map tasks=9891
  Total vcore-milliseconds taken by all reduce tasks=6400
  Total megabyte-milliseconds taken by all map tasks=10128384
  Total megabyte-milliseconds taken by all reduce tasks=6553600
Map-Reduce Framework
  Map input records=0
  Map output records=0
  Map output bytes=0
  Map output materialized bytes=6
  Input split bytes=127
  Combine input records=0
Reduce input groups=0
Reduce shuffle bytes=6
Reduce input records=0
Reduce output records=0
Spilled Records=0
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=176
CPU time spent (ms)=1170
Physical memory (bytes) snapshot=322920448
Virtual memory (bytes) snapshot=3015020736
Total committed heap usage (bytes)=195301376
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=0
File Output Format Counters
  Bytes Written=0
[cloudera@quickstart ~]$ hdfs dfs -ls /out1
Found 2 items
-rw-r--r-- 1 cloudera supergroup 0 2023-03-20 06:01 /out1/ SUCCESS
-rw-r--r-- 1 cloudera supergroup 0 2023-03-20 06:01 /out1/part-r-0000
0
```

```
[cloudera@quickstart ~]$ hdfs dfs -cat /out1/part-r-00000
Hii      2
How      1
am       1
are      1
fine     1
i        1
u        1
[cloudera@quickstart ~]$
```