Sort on Hadoop/Spark

Bobo Liang A20356451

1. Brief description of the problem

Large file can't be sorted by loaded them all into memory, so in order to sort large files, so we need use divide concur algorithm. Split the large files into small pieces file, sort these chunk files, and then merge them.

Also, to sort large datasets by one computer resource can take very long time, because the CPU speed, memory size and I/O speed of memory and disk. While use cluster, which is comprised of hundreds of computer nodes, we can divide jobs that can be processed simultaneously, by doing so we can speed up dramatically the process of large datasets.

2. Methodology

Shared Memory sort and linux sort: divide input dataset file into many chunk files which fit in the memory, sort each chunk file, write back to the file, loop read small piece of each sorted chunk file into memory buffers, merge sort into output file, if buffer is empty, load piece from file until the file is empty.

Hadoop sort: Hadoop framework divides the large file into blocks and stores in the worker nodes. A MapReduce framework (or system) is usually composed of three operations (or steps):

- Map: each worker node applies the map function to the local data, and writes the output to a temporary storage. A master node ensures that only one copy of redundant input data is processed.
- Shuffle: worker nodes redistribute data based on the output keys (produced by the map function), such that all data belonging to one key is located on the same worker node.
- Reduce: worker nodes now process each group of output data, per key, in parallel.

In the map output, the data has already been sorted automatically, so in the reduce function, we just iterate the values and write them to output.

Spark sort:

- input RDD is sampled
- input RDD is partitioned
- each partition from the second step is sorted locally

3. Performance

Table 1: Performance evaluation of sort (weak scaling – small dataset)

Experiment	Shared Memory (1VM 2GB)	Linux Sort (1VM 2GB)	Hadoop Sort (4VM 8GB)	Spark Sort (4VM 8GB)
Computation Time (sec)	182	41	444	540
Data Read (GB)	4x2=8	2x2=4	7x8=56	4x8=32
Data Write (GB)	4x2=8	2x2=4	7x8=56	4x8=32
I/O Throughput (MB/sec)	88	195	252	118
Speedup				
Efficiency				

Table 2: Performance evaluation of sort (Strong scaling – large dataset)

Experiment	Shared Memory (1VM 20GB)	Linux Sort (1VM 20GB)	Hadoop Sort (4VM 20GB)	Spark Sort (4VM 20GB)
Computation Time (sec)	1500	489	1259	1470
Data Read (GB)	6x20=120	2x20=40	7x20=140	4x20=80
Data Write (GB)	6x20=120	2x20=40	7x20=140	4x20=80
I/O Throughput (MB/sec)	160	163.5	222	109
Speedup				
Efficiency				

Table 3: Performance evaluation of sort (weak scaling – large dataset)

Experiment	Shared Memory (1VM 20GB)	Linux Sort (1VM 20GB)	Hadoop Sort (4VM 80GB)	Spark Sort (4VM 80GB)
Computation Time (sec)	1500	489	3870	3360
Data Read (GB)	6x20=120	2x20=40	7x80=560	4x80=320
Data Write (GB)	6x20=120	2x20=40	7x80=560	4x80=320
I/O Throughput (MB/sec)	160	163.5	145	95
Speedup				
Efficiency				

Performance analysis

Experiment	Shared Memory (1VM 20GB)	Linux Sort (1VM 20GB)	Hadoop Sort (4VM 80GB)	Spark Sort (4VM 80GB)
Threads Num	16 thread for first pass	-	-	-
Mappers Num	-	-	1250	-
Reducers Num	-	-	8	-
Data read times	6	2	7	4
Data write times	6	2	7	4

4. Logs

Hadoopsort8GB.log

Total Time= 856.0 seconds
checksum 2626d6458319832

18/04/29 16:09:11 INFO client.RMProxy: Connecting to ResourceManager at hadoop-g/192.168.2.34:8032
18/04/29 16:09:12 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
18/04/29 16:09:12 INFO input.FileInputFormat: Total input files to process : 1
18/04/29 16:09:12 INFO mapreduce.JobSubmitter: number of splits:120
18/04/29 16:09:12 INFO Configuration.deprecation: yarn.resourcemanager.system-metrics-publisher.enabled is deprecated. Instead, use yarn.system-metrics-publisher.enabled
18/04/29 16:09:12 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1524522050925_0292
18/04/29 16:09:13 INFO impl.YarnClientImpl: Submitted application application_1524522050925_0292
18/04/29 16:09:13 INFO mapreduce.Job: The url to track the job: http://hadoop-g:8088/proxy/application_1524522050925_0292/
18/04/29 16:09:13 INFO mapreduce.Job: Running job: job_1524522050925_0292
18/04/29 16:09:19 INFO mapreduce.Job: Job job_1524522050925_0292 running in uber mode : false
18/04/29 16:09:19 INFO mapreduce.Job: map 0% reduce 0%

```
18/04/29 16:09:37 INFO mapreduce.Job:  map 1% reduce 0%
18/04/29 16:09:43 INFO mapreduce.Job:  map 9% reduce 0%
18/04/29 16:09:44 INFO mapreduce.Job:  map 14% reduce 0%
18/04/29 16:09:47 INFO mapreduce.Job:  map 15% reduce 0%
18/04/29 16:09:50 INFO mapreduce.Job:  map 18% reduce 0%
18/04/29 16:09:51 INFO mapreduce.Job:  map 21% reduce 0%
18/04/29 16:09:53 INFO mapreduce.Job:  map 22% reduce 0%
18/04/29 16:10:01 INFO mapreduce.Job:  map 23% reduce 0%
18/04/29 16:10:07 INFO mapreduce.Job:  map 24% reduce 0%
18/04/29 16:10:10 INFO mapreduce.Job:  map 26% reduce 0%
18/04/29 16:10:11 INFO mapreduce.Job:  map 29% reduce 0%
18/04/29 16:10:12 INFO mapreduce.Job:  map 31% reduce 0%
18/04/29 16:10:13 INFO mapreduce.Job:  map 33% reduce 8%
18/04/29 16:10:14 INFO mapreduce.Job:  map 36% reduce 8%
18/04/29 16:10:15 INFO mapreduce.Job:  map 37% reduce 8%
18/04/29 16:10:16 INFO mapreduce.Job:  map 38% reduce 8%
18/04/29 16:10:17 INFO mapreduce.Job:  map 39% reduce 8%
18/04/29 16:10:18 INFO mapreduce.Job:  map 40% reduce 8%
18/04/29 16:10:19 INFO mapreduce.Job:  map 41% reduce 9%
18/04/29 16:10:20 INFO mapreduce.Job:  map 42% reduce 9%
18/04/29 16:10:21 INFO mapreduce.Job:  map 43% reduce 9%
18/04/29 16:10:23 INFO mapreduce.Job:  map 44% reduce 9%
18/04/29 16:10:26 INFO mapreduce.Job:  map 44% reduce 13%
18/04/29 16:10:28 INFO mapreduce.Job:  map 45% reduce 13%
18/04/29 16:10:31 INFO mapreduce.Job:  map 46% reduce 13%
18/04/29 16:10:32 INFO mapreduce.Job:  map 46% reduce 14%
18/04/29 16:10:38 INFO mapreduce.Job:  map 50% reduce 15%
18/04/29 16:10:40 INFO mapreduce.Job:  map 53% reduce 15%
18/04/29 16:10:41 INFO mapreduce.Job:  map 54% reduce 15%
18/04/29 16:10:42 INFO mapreduce.Job:  map 55% reduce 15%
18/04/29 16:10:43 INFO mapreduce.Job:  map 57% reduce 15%
18/04/29 16:10:44 INFO mapreduce.Job:  map 58% reduce 17%
18/04/29 16:10:45 INFO mapreduce.Job:  map 59% reduce 17%
18/04/29 16:10:46 INFO mapreduce.Job:  map 62% reduce 17%
18/04/29 16:10:47 INFO mapreduce.Job:  map 64% reduce 17%
18/04/29 16:10:50 INFO mapreduce.Job:  map 64% reduce 20%
18/04/29 16:10:51 INFO mapreduce.Job:  map 65% reduce 20%
18/04/29 16:10:52 INFO mapreduce.Job:  map 66% reduce 20%
18/04/29 16:10:53 INFO mapreduce.Job:  map 67% reduce 20%
```

```
18/04/29 16:10:56 INFO mapreduce.Job:  map 68% reduce 22%
18/04/29 16:10:58 INFO mapreduce.Job:  map 69% reduce 22%
18/04/29 16:11:01 INFO mapreduce.Job:  map 70% reduce 22%
18/04/29 16:11:02 INFO mapreduce.Job:  map 71% reduce 23%
18/04/29 16:11:03 INFO mapreduce.Job:  map 72% reduce 23%
18/04/29 16:11:04 INFO mapreduce.Job:  map 74% reduce 23%
18/04/29 16:11:05 INFO mapreduce.Job:  map 75% reduce 23%
18/04/29 16:11:06 INFO mapreduce.Job:  map 76% reduce 23%
18/04/29 16:11:07 INFO mapreduce.Job:  map 78% reduce 23%
18/04/29 16:11:08 INFO mapreduce.Job:  map 79% reduce 25%
18/04/29 16:11:10 INFO mapreduce.Job:  map 80% reduce 25%
18/04/29 16:11:13 INFO mapreduce.Job:  map 81% reduce 25%
18/04/29 16:11:14 INFO mapreduce.Job:  map 82% reduce 26%
18/04/29 16:11:16 INFO mapreduce.Job:  map 83% reduce 26%
18/04/29 16:11:20 INFO mapreduce.Job:  map 84% reduce 27%
18/04/29 16:11:23 INFO mapreduce.Job:  map 85% reduce 27%
18/04/29 16:11:24 INFO mapreduce.Job:  map 86% reduce 27%
18/04/29 16:11:25 INFO mapreduce.Job:  map 87% reduce 27%
18/04/29 16:11:26 INFO mapreduce.Job:  map 88% reduce 28%
18/04/29 16:11:29 INFO mapreduce.Job:  map 90% reduce 28%
18/04/29 16:11:30 INFO mapreduce.Job:  map 91% reduce 28%
18/04/29 16:11:37 INFO mapreduce.Job:  map 98% reduce 30%
18/04/29 16:11:38 INFO mapreduce.Job:  map 98% reduce 33%
18/04/29 16:11:40 INFO mapreduce.Job:  map 99% reduce 33%
18/04/29 16:11:41 INFO mapreduce.Job:  map 100% reduce 33%
18/04/29 16:18:58 INFO mapreduce.Job:  map 100% reduce 35%
18/04/29 16:19:04 INFO mapreduce.Job:  map 100% reduce 37%
18/04/29 16:19:10 INFO mapreduce.Job:  map 100% reduce 39%
18/04/29 16:19:16 INFO mapreduce.Job:  map 100% reduce 41%
18/04/29 16:19:22 INFO mapreduce.Job:  map 100% reduce 43%
18/04/29 16:19:28 INFO mapreduce.Job:  map 100% reduce 45%
18/04/29 16:19:34 INFO mapreduce.Job:  map 100% reduce 47%
18/04/29 16:19:40 INFO mapreduce.Job:  map 100% reduce 48%
18/04/29 16:19:46 INFO mapreduce.Job:  map 100% reduce 50%
18/04/29 16:19:52 INFO mapreduce.Job:  map 100% reduce 52%
18/04/29 16:19:58 INFO mapreduce.Job:  map 100% reduce 54%
18/04/29 16:20:04 INFO mapreduce.Job:  map 100% reduce 56%
18/04/29 16:20:10 INFO mapreduce.Job:  map 100% reduce 58%
18/04/29 16:20:17 INFO mapreduce.Job:  map 100% reduce 59%
```

```
18/04/29 16:20:23 INFO mapreduce.Job:  map 100% reduce 61%
18/04/29 16:20:29 INFO mapreduce.Job:  map 100% reduce 63%
18/04/29 16:20:35 INFO mapreduce.Job:  map 100% reduce 65%
18/04/29 16:20:41 INFO mapreduce.Job:  map 100% reduce 67%
18/04/29 16:20:53 INFO mapreduce.Job:  map 100% reduce 68%
18/04/29 16:20:59 INFO mapreduce.Job:  map 100% reduce 70%
18/04/29 16:21:05 INFO mapreduce.Job:  map 100% reduce 72%
18/04/29 16:21:10 INFO mapreduce.Job:  map 100% reduce 74%
18/04/29 16:21:16 INFO mapreduce.Job:  map 100% reduce 75%
18/04/29 16:21:23 INFO mapreduce.Job:  map 100% reduce 76%
18/04/29 16:21:29 INFO mapreduce.Job:  map 100% reduce 77%
18/04/29 16:21:35 INFO mapreduce.Job:  map 100% reduce 79%
18/04/29 16:21:41 INFO mapreduce.Job:  map 100% reduce 80%
18/04/29 16:21:47 INFO mapreduce.Job:  map 100% reduce 82%
18/04/29 16:21:53 INFO mapreduce.Job:  map 100% reduce 83%
18/04/29 16:21:59 INFO mapreduce.Job:  map 100% reduce 84%
18/04/29 16:22:05 INFO mapreduce.Job:  map 100% reduce 86%
18/04/29 16:22:11 INFO mapreduce.Job:  map 100% reduce 87%
18/04/29 16:22:17 INFO mapreduce.Job:  map 100% reduce 88%
18/04/29 16:22:23 INFO mapreduce.Job:  map 100% reduce 90%
18/04/29 16:22:29 INFO mapreduce.Job:  map 100% reduce 92%
18/04/29 16:22:35 INFO mapreduce.Job:  map 100% reduce 93%
18/04/29 16:22:41 INFO mapreduce.Job:  map 100% reduce 95%
18/04/29 16:22:47 INFO mapreduce.Job:  map 100% reduce 97%
18/04/29 16:22:11 INFO mapreduce.Job:  map 100% reduce 87%
18/04/29 16:22:17 INFO mapreduce.Job:  map 100% reduce 88%
18/04/29 16:22:23 INFO mapreduce.Job:  map 100% reduce 90%
18/04/29 16:22:29 INFO mapreduce.Job:  map 100% reduce 92%
18/04/29 16:22:35 INFO mapreduce.Job:  map 100% reduce 93%
18/04/29 16:22:41 INFO mapreduce.Job:  map 100% reduce 95%
18/04/29 16:22:47 INFO mapreduce.Job:  map 100% reduce 97%
18/04/29 16:22:53 INFO mapreduce.Job:  map 100% reduce 98%
18/04/29 16:22:59 INFO mapreduce.Job:  map 100% reduce 100%
18/04/29 16:23:26 INFO mapreduce.Job: Job job_1524522050925_0292 completed
successfully
18/04/29 16:23:27 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=4865392093
                FILE: Number of bytes written=7094903869
```

FILE: Number of read operations=0
FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=8000500864
HDFS: Number of bytes written=8000000000
HDFS: Number of read operations=363
HDFS: Number of large read operations=0
HDFS: Number of write operations=2

Job Counters

Killed map tasks=2
Launched map tasks=121
Launched reduce tasks=1
Data-local map tasks=78
Rack-local map tasks=43
Total time spent by all maps in occupied slots (ms)=3000995
Total time spent by all reduces in occupied slots (ms)=814252
Total time spent by all map tasks (ms)=3000995
Total time spent by all reduce tasks (ms)=814252
Total vcore-milliseconds taken by all map tasks=3000995
Total vcore-milliseconds taken by all reduce tasks=814252
Total megabyte-milliseconds taken by all map tasks=3073018880
Total megabyte-milliseconds taken by all reduce tasks=833794048

Map-Reduce Framework

Map input records=80000000
Map output records=80000000
Map output bytes=8000000000
Map output materialized bytes=2205061594
Input split bytes=13440
Combine input records=0
Combine output records=0
Reduce input groups=80000000
Reduce shuffle bytes=2205061594
Reduce input records=80000000
Reduce output records=80000000
Spilled Records=254763951
Shuffled Maps =120
Failed Shuffles=0
Merged Map outputs=120
GC time elapsed (ms)=39780

CPU time spent (ms)=1980710
Physical memory (bytes) snapshot=35331608576
Virtual memory (bytes) snapshot=238297358336
Total committed heap usage (bytes)=24678236160
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=8000487424
File Output Format Counters
Bytes Written=8000000000
Spent 60ms computing base-splits.
Spent 5ms computing TeraScheduler splits.

hadoop20GB.log

Total Time= 1259.0 seconds
checksum 2626d6458319832

hadoop80GB.log

failed due to overtime in

spark8GB.log

2018-04-29 12:52:34 WARN NativeCodeLoader:62 - Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Spark context Web UI available at http://hadoop-d:4040
Spark context available as 'sc' (master = local[*], app id = local-1525006370032).
Spark session available as 'spark'.
Loading ./sparksort8GB.scala...
start: Long = 1525006372760
file: org.apache.spark.rdd.RDD[String] = /input/data-8GB MapPartitionsRDD[1] at textFile at <console>:24
file_sort: org.apache.spark.rdd.RDD[(String, String)] = MapPartitionsRDD[2] at map at <console>:25

sortkey: org.apache.spark.rdd.RDD[(String, String)] = ShuffledRDD[5] at sortByKey at <console>:26
keyval: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[6] at map at <console>:25
repartitioned: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[10] at repartition at <console>:25
end: Long = 1525006843562
Sorting time:-470802ms

Welcome to

```
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.3.0
      /_/
```

Using Scala version 2.11.8 (OpenJDK 64-Bit Server VM, Java 1.8.0_162)
Type in expressions to have them evaluated.
Type :help for more information.

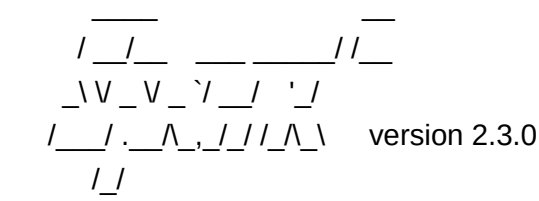
checksum 5f5e36b38987181

spark20GB.log

018-04-29 13:05:12 WARN NativeCodeLoader:62 - Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Spark context Web UI available at http://hadoop-d:4040
Spark context available as 'sc' (master = local[*], app id = local-1525007120259).
Spark session available as 'spark'.
Loading ./sparksort20GB.scala...
file: org.apache.spark.rdd.RDD[String] = /input/data-20GB MapPartitionsRDD[1] at textFile at <console>:24
file_sort: org.apache.spark.rdd.RDD[(String, String)] = MapPartitionsRDD[2] at map at <console>:25
sortkey: org.apache.spark.rdd.RDD[(String, String)] = ShuffledRDD[5] at sortByKey at <console>:26
keyval: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[6] at map at <console>:25
repartitioned: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[10] at repartition at <console>:25

end: Long = 1525008509400

Welcome to



Using Scala version 2.11.8 (OpenJDK 64-Bit Server VM, Java 1.8.0_162)
Type in expressions to have them evaluated.
Type :help for more information.

18/04/29 13:41:37 INFO mapreduce.Job: Running job: job_1524709778346_0362
18/04/29 13:41:46 INFO mapreduce.Job: Job job_1524709778346_0362 running in uber mode : false
18/04/29 13:41:46 INFO mapreduce.Job: map 0% reduce 0%
18/04/29 13:42:03 INFO mapreduce.Job: map 4% reduce 0%
18/04/29 13:42:09 INFO mapreduce.Job: map 6% reduce 0%
18/04/29 13:42:15 INFO mapreduce.Job: map 8% reduce 0%
18/04/29 13:42:21 INFO mapreduce.Job: map 11% reduce 0%
18/04/29 13:42:27 INFO mapreduce.Job: map 13% reduce 0%
18/04/29 13:42:33 INFO mapreduce.Job: map 15% reduce 0%
18/04/29 13:42:39 INFO mapreduce.Job: map 17% reduce 0%
18/04/29 13:42:45 INFO mapreduce.Job: map 19% reduce 0%
18/04/29 13:42:15 INFO mapreduce.Job: map 8% reduce 0%
18/04/29 13:42:21 INFO mapreduce.Job: map 11% reduce 0%
18/04/29 13:42:27 INFO mapreduce.Job: map 13% reduce 0%
18/04/29 13:42:33 INFO mapreduce.Job: map 15% reduce 0%
18/04/29 13:42:39 INFO mapreduce.Job: map 17% reduce 0%
18/04/29 13:42:45 INFO mapreduce.Job: map 19% reduce 0%
18/04/29 13:42:52 INFO mapreduce.Job: map 21% reduce 0%
18/04/29 13:42:58 INFO mapreduce.Job: map 24% reduce 0%
18/04/29 13:43:04 INFO mapreduce.Job: map 26% reduce 0%
18/04/29 13:43:10 INFO mapreduce.Job: map 28% reduce 0%
18/04/29 13:43:16 INFO mapreduce.Job: map 30% reduce 0%
18/04/29 13:43:22 INFO mapreduce.Job: map 32% reduce 0%
18/04/29 13:43:28 INFO mapreduce.Job: map 35% reduce 0%
18/04/29 13:43:34 INFO mapreduce.Job: map 37% reduce 0%

```
18/04/29 13:43:40 INFO mapreduce.Job:  map 39% reduce 0%
18/04/29 13:43:46 INFO mapreduce.Job:  map 41% reduce 0%
18/04/29 13:43:52 INFO mapreduce.Job:  map 43% reduce 0%
18/04/29 13:43:58 INFO mapreduce.Job:  map 45% reduce 0%
18/04/29 13:44:04 INFO mapreduce.Job:  map 47% reduce 0%
18/04/29 13:44:10 INFO mapreduce.Job:  map 49% reduce 0%
18/04/29 13:44:16 INFO mapreduce.Job:  map 52% reduce 0%
18/04/29 13:44:22 INFO mapreduce.Job:  map 54% reduce 0%
18/04/29 13:44:28 INFO mapreduce.Job:  map 56% reduce 0%
18/04/29 13:44:34 INFO mapreduce.Job:  map 58% reduce 0%
18/04/29 13:44:40 INFO mapreduce.Job:  map 60% reduce 0%
18/04/29 13:44:46 INFO mapreduce.Job:  map 62% reduce 0%
18/04/29 13:44:52 INFO mapreduce.Job:  map 65% reduce 0%
18/04/29 13:44:58 INFO mapreduce.Job:  map 100% reduce 0%
18/04/29 13:45:04 INFO mapreduce.Job:  map 100% reduce 100%
18/04/29 13:45:04 INFO mapreduce.Job: Job job_1524709778346_0362 completed
successfully
18/04/29 13:45:04 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=91
                FILE: Number of bytes written=403655
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=20000000122
                HDFS: Number of bytes written=25
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
               File System Counters
                FILE: Number of bytes read=91
                FILE: Number of bytes written=403655
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=20000000122
                HDFS: Number of bytes written=25
                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
Total time spent by all maps in occupied slots (ms)=189656
Total time spent by all reduces in occupied slots (ms)=3546
Total time spent by all map tasks (ms)=189656
Total time spent by all reduce tasks (ms)=3546
Total vcore-milliseconds taken by all map tasks=189656
Total vcore-milliseconds taken by all reduce tasks=3546
Total megabyte-milliseconds taken by all map tasks=194207744
Total megabyte-milliseconds taken by all reduce tasks=3631104
Map-Reduce Framework
Map input records=200000000
Map output records=3
Map output bytes=79
Map output materialized bytes=91
Input split bytes=122
Combine input records=0
Combine output records=0
Reduce input groups=3
Reduce shuffle bytes=91
Reduce input records=3
Reduce output records=1
Spilled Records=6
Shuffled Maps =1
Failed Shuffles=0
Merged Map outputs=1
GC time elapsed (ms)=2802
CPU time spent (ms)=189660
Physical memory (bytes) snapshot=488460288
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=20000000000
File Output Format Counters
Bytes Written=25

checksum 5f5e36b38987181

spark80GB.log

failed due to out of space in phase 3.