IEMS 5730 Spring 2019 Homework #0

I declare that the assignment submitted on Elearning system is original except for source material explicitly acknowledged, and that the same or related material has not been previously submitted for another course. I also acknowledge that I am aware of University policy and regulations on honesty in academic work, and of the disciplinary guidelines and procedures applicable to breaches of such policy and regulations, as contained in the website http://www.cuhk.edu.hk/policy/academichonesty/.

Signed (Student Wenli SONG) Date: 17 JAN 2019

Name Wenli SONG SID 1155114524

Q1. Hadoop Cluster Setup

a. Single-node Hadoop

i. I setup a single-node Hadoop cluster in a pseudo-distributed mode. Below are printed pages of http://localhost:50070, I use external IP(34.73.64.109) to replace the localhost.

12/01/2019 Namenode information

Overview 'localhost:9000' (active)

Started:	Sat Jan 12 18:31:33 +0800 2019
Version:	2.9.2, r826afbeae31ca687bc2f8471dc841b66ed2c6704
Compiled:	Tue Nov 13 20:42:00 +0800 2018 by ajisaka from branch-2.9.2
Cluster ID:	CID-ce914393-f5a9-4f57-a5b4-e69d4c93a1f9
Block Pool ID:	BP-550933482-10.142.0.6-1547287791962

Summary

Security is off.

Safemode is off.

19 files and directories, 5 blocks = 24 total filesystem object(s).

Heap Memory used 91.35 MB of 214 MB Heap Memory. Max Heap Memory is 889 MB.

Non Heap Memory used 37.57 MB of 38.94 MB Commited Non Heap Memory. Max Non Heap Memory is 214 MB.

Configured Capacity:	98.4 GB
DFS Used:	9.85 MB (0.01%)
Non DFS Used:	3.18 GB
DFS Remaining:	91.09 GB (92.57%)
Block Pool Used:	9.85 MB (0.01%)
DataNodes usages% (Min/Median/Max/stdDev):	0.01% / 0.01% / 0.01% / 0.00%
Live Nodes	1 (Decommissioned: 0, In Maintenance: 0)
Dead Nodes	0 (Decommissioned: 0, In Maintenance: 0)
Decommissioning Nodes	0
Entering Maintenance Nodes	0
Total Datanode Volume Failures	0 (0 B)
Number of Under-Replicated Blocks	0
Number of Blocks Pending Deletion	0
Block Deletion Start Time	Sat Jan 12 18:31:33 +0800 2019
Last Checkpoint Time	Sat Jan 12 18:32:48 +0800 2019

NameNode Journal Status

Current transaction ID: 106

Journal Manager	State
FileJournalManager(root=/home/hduser/hdfs/namenode)	EditLogFileOutputStream(/home/hduser/hdfs/namenode/current/edits_inprogress_0000000000000000003)

http://34.73.64.109:50070/dfshealth.html#tab-overview 1/2

12/01/2019 Namenode information

NameNode Storage

Storage Directory	Type	State
/home/hduser/hdfs/namenode	IMAGE_AND_EDITS	Active

DFS Storage Types

Storage Type	Configured Capacity	Capacity Used	Capacity Remaining	Block Pool Used	Nodes In Service
DISK	98.4 GB	9.85 MB (0.01%)	91.09 GB (92.57%)	9.85 MB	1

Hadoop, 2018.

http://34.73.64.109:50070/dfshealth.html#tab-overview 2/2

ii. First, I generate some 10000 lines data.

songwenli — hduser@master-1: /usr/local/hadoop — ssh sfswl@34.73.64.109 — 78×39

```
19/01/12 11:42:44 INFO mapreduce.Job: Counters: 31
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=396336
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=164
                HDFS: Number of bytes written=10000000
                HDFS: Number of read operations=8
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Job Counters
                Launched map tasks=2
                Other local map tasks=2
                Total time spent by all maps in occupied slots (ms)=15007
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=15007
                Total vcore-milliseconds taken by all map tasks=15007
                Total megabyte-milliseconds taken by all map tasks=15367168
        Map-Reduce Framework
                Map input records=100000
                Map output records=100000
                Input split bytes=164
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=207
                CPU time spent (ms)=2620
                Physical memory (bytes) snapshot=385810432
                Virtual memory (bytes) snapshot=1675624448
                Total committed heap usage (bytes)=293601280
        org.apache.hadoop.examples.terasort.TeraGen$Counters
                CHECKSUM=214574985129000
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=10000000
```

Then sort them by map reduce,

songwenli — hduser@master-1: /usr/local/hadoop — ssh sfswl@34.73.64.109 —...

```
19/01/12 11:45:11 INFO mapreduce.Job: Counters: 50
        File System Counters
                FILE: Number of bytes read=10400006
                FILE: Number of bytes written=21398673
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=10000252
                HDFS: Number of bytes written=10000000
                HDFS: Number of read operations=9
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        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)=17208
                Total time spent by all reduces in occupied slots (ms)=4617
                Total time spent by all map tasks (ms)=17208
                Total time spent by all reduce tasks (ms)=4617
                Total vcore-milliseconds taken by all map tasks=17208
                Total vcore-milliseconds taken by all reduce tasks=4617
                Total megabyte-milliseconds taken by all map tasks=17620992
                Total megabyte-milliseconds taken by all reduce tasks=4727808
        Map-Reduce Framework
                Map input records=100000
                Map output records=100000
                Map output bytes=10200000
                Map output materialized bytes=10400012
                Input split bytes=252
                Combine input records=0
                Combine output records=0
                Reduce input groups=100000
                Reduce shuffle bytes=10400012
                Reduce input records=100000
                Reduce output records=100000
                Spilled Records=200000
                Shuffled Maps =2
                Failed Shuffles=0
                Merged Map outputs=2
                GC time elapsed (ms)=287
                CPU time spent (ms)=7080
                Physical memory (bytes) snapshot=782544896
                Virtual memory (bytes) snapshot=2506321920
                Total committed heap usage (bytes)=548405248
        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=10000000
        File Output Format Counters
                Bytes Written=10000000
19/01/12 11:45:11 INFO terasort.TeraSort: done
```

At last, check the output,

```
songwenli — hduser@master-1: /usr/local/hadoop — ssh sfswl@34.73.64.109 — 88×57
19/01/12 12:56:43 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=92
                FILE: Number of bytes written=397321
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=10000127
                HDFS: Number of bytes written=22
                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)=4720
                Total time spent by all reduces in occupied slots (ms)=3719
                Total time spent by all map tasks (ms)=4720
                Total time spent by all reduce tasks (ms)=3719
                Total vcore-milliseconds taken by all map tasks=4720
                Total vcore-milliseconds taken by all reduce tasks=3719
                Total megabyte-milliseconds taken by all map tasks=9666560
                Total megabyte-milliseconds taken by all reduce tasks=7616512
        Map-Reduce Framework
                Map input records=100000
                Map output records=3
                Map output bytes=80
                Map output materialized bytes=92
                Input split bytes=127
                Combine input records=0
                Combine output records=0
                Reduce input groups=3
                Reduce shuffle bytes=92
                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)=112
                CPU time spent (ms)=2400
                Physical memory (bytes) snapshot=484450304
                Virtual memory (bytes) snapshot=1674027008
                Total committed heap usage (bytes)=351797248
        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=10000000
        File Output Format Counters
                Bytes Written=22
```

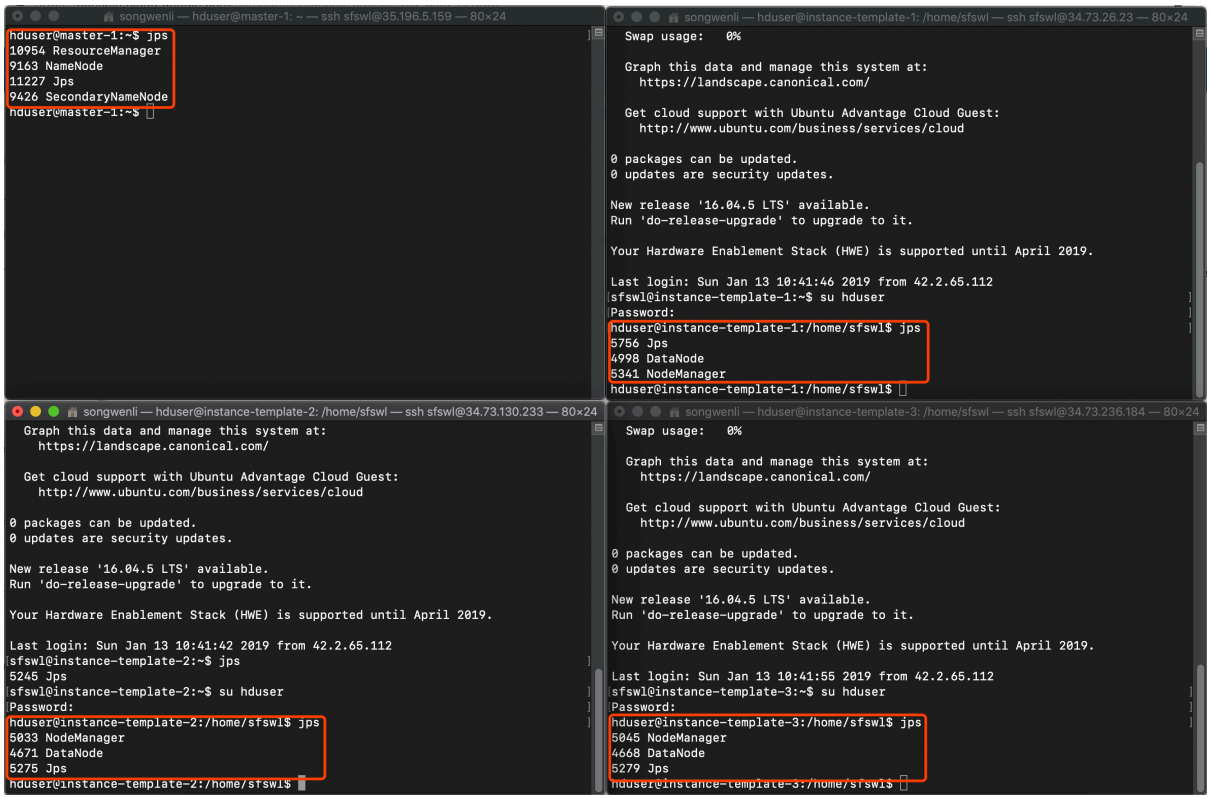
b. Multi-node Hadoop Cluster

i. First I create 4 VMs, install java and Hadoop.

Config the environment.

After start the HDFS, we can use command ‘jps’ to see processes.

We can see that ResourceManager, NameNode and SecondaryNameNode are running on the Master node, NodeManager, DataNode are running on the Slave node.



ii. When we use Teragen to generate data, 1 line is 100B.

Note:

In my assignment,

2GB = 2*1000*1000*10 lines * 100B / line

20GB = 20*1000*1000*10 lines * 100B / line

```
$ hadoop jar ~/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar
teragen 20000000 terasort/input-2G
```

Then sort them.

```
$ hadoop jar ~/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar
terasort terasort/input-2G terasort/output-2G
```

```
songwenli — ubuntu@ip-172-31-3-192: ~/hadoop — ssh -i .ssh/hadoop.pem ubu...
19/01/16 05:12:58 INFO mapreduce.Job:  map 100% reduce 91%
19/01/16 05:13:04 INFO mapreduce.Job:  map 100% reduce 100%
19/01/16 05:13:05 INFO mapreduce.Job: Job job_1547614870205_0002 completed succe
ssfully
19/01/16 05:13:05 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=4857932310
                FILE: Number of bytes written=6941323721
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2000002080
                HDFS: Number of bytes written=2000000000
                HDFS: Number of read operations=51
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=1
                Launched map tasks=17
                Launched reduce tasks=1
                Data-local map tasks=16
                Rack-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=366481
                Total time spent by all reduces in occupied slots (ms)=48024
                Total time spent by all map tasks (ms)=366481
                Total time spent by all reduce tasks (ms)=48024
                Total vcore-milliseconds taken by all map tasks=366481
                Total vcore-milliseconds taken by all reduce tasks=48024
                Total megabyte-milliseconds taken by all map tasks=375276544
                Total megabyte-milliseconds taken by all reduce tasks=49176576
        Map-Reduce Framework
                Map input records=20000000
                Map output records=20000000
                Map output bytes=2040000000
                Map output materialized bytes=2080000096
                Input split bytes=2080
                Combine input records=0
                Combine output records=0
                Reduce input groups=20000000
                Reduce shuffle bytes=2080000096
                Reduce input records=20000000
                Reduce output records=20000000
                Spilled Records=66710885
                Shuffled Maps =16
                Failed Shuffles=0
                Merged Map outputs=16
                GC time elapsed (ms)=3553
                CPU time spent (ms)=158960
                Physical memory (bytes) snapshot=4900782080
                Virtual memory (bytes) snapshot=33645268992
                Total committed heap usage (bytes)=3304587264
        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=2000000000
        File Output Format Counters
                Bytes Written=2000000000
19/01/16 05:13:05 INFO terasort.TeraSort: done
```

Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus
application_1547614870205_0002	ubuntu	TeraSort	MAPREDUCE	default	0	Wed Jan 16 13:11:51 +0800 2019	Wed Jan 16 13:13:03 +0800 2019	FINISHED	SUCCEEDED

From the screenshot we can see that it costs 1 mins 12 seconds.

Last, we validate the output.

```
$ hadoop jar ~/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar
teravalidate terasort/output-2G terasort/check-2G
```

```
songwenli — ubuntu@ip-172-31-3-192: ~/hadoop — ssh -i .ssh/hadoop.pem ubu...
19/01/16 05:17:05 INFO mapreduce.Job:  map 100% reduce 0%
19/01/16 05:17:09 INFO mapreduce.Job:  map 100% reduce 100%
19/01/16 05:17:10 INFO mapreduce.Job: Job job_1547614870205_0003 completed succe
ssfully
19/01/16 05:17:10 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=94
                FILE: Number of bytes written=397159
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2000000131
                HDFS: Number of bytes written=24
                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
                Rack-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=18626
                Total time spent by all reduces in occupied slots (ms)=2147
                Total time spent by all map tasks (ms)=18626
                Total time spent by all reduce tasks (ms)=2147
                Total vcore-milliseconds taken by all map tasks=18626
                Total vcore-milliseconds taken by all reduce tasks=2147
                Total megabyte-milliseconds taken by all map tasks=19073024
                Total megabyte-milliseconds taken by all reduce tasks=2198528
        Map-Reduce Framework
                Map input records=20000000
                Map output records=3
                Map output bytes=82
                Map output materialized bytes=94
                Input split bytes=131
                Combine input records=0
                Combine output records=0
                Reduce input groups=3
                Reduce shuffle bytes=94
                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)=267
                CPU time spent (ms)=17040
                Physical memory (bytes) snapshot=479862784
                Virtual memory (bytes) snapshot=3948572672
                Total committed heap usage (bytes)=315621376
        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=2000000000
        File Output Format Counters
                Bytes Written=24
```

Generate 20G data

```
$ hadoop jar ~/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar
teragen 200000000 terasort/input-20G
```

songwenli — ubuntu@ip-172-31-3-192: ~/hadoop — ssh -i .ssh/hadoop.pem ubu...

```
19/01/16 05:21:04 INFO mapreduce.Job:  map 98% reduce 0%
19/01/16 05:21:07 INFO mapreduce.Job:  map 100% reduce 0%
19/01/16 05:21:07 INFO mapreduce.Job: Job job_1547614870205_0004 completed succe
ssfully
19/01/16 05:21:07 INFO mapreduce.Job: Counters: 31
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=396148
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=170
                HDFS: Number of bytes written=20000000000
                HDFS: Number of read operations=8
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=4
        Job Counters
                Launched map tasks=2
                Other local map tasks=2
                Total time spent by all maps in occupied slots (ms)=296692
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=296692
                Total vcore-milliseconds taken by all map tasks=296692
                Total megabyte-milliseconds taken by all map tasks=303812608
        Map-Reduce Framework
                Map input records=200000000
                Map output records=200000000
                Input split bytes=170
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=1868
                CPU time spent (ms)=237420
                Physical memory (bytes) snapshot=476225536
                Virtual memory (bytes) snapshot=3954708480
                Total committed heap usage (bytes)=285736960
        org.apache.hadoop.examples.terasort.TeraGen$Counters
                CHECKSUM=429525686602393848
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=20000000000
```

Sort them

```
$ hadoop jar ~/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar
terasort terasort/input-20G terasort/output-20G
```

```
songwenli — ubuntu@ip-172-31-3-192: ~/hadoop — ssh -i .ssh/hadoop.pem ubu...
19/01/16 05:41:27 INFO mapreduce.Job:  map 100% reduce 100%
19/01/16 05:41:27 INFO mapreduce.Job: Job job_1547614870205_0005 completed succe
ssfully
19/01/16 05:41:27 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=69937669844
                FILE: Number of bytes written=90767794775
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=20000019650
                HDFS: Number of bytes written=20000000000
                HDFS: Number of read operations=453
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=2
                Launched map tasks=152
                Launched reduce tasks=1
                Data-local map tasks=151
                Rack-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=6752658
                Total time spent by all reduces in occupied slots (ms)=937270
                Total time spent by all map tasks (ms)=6752658
                Total time spent by all reduce tasks (ms)=937270
                Total vcore-milliseconds taken by all map tasks=6752658
                Total vcore-milliseconds taken by all reduce tasks=937270
                Total megabyte-milliseconds taken by all map tasks=6914721792
                Total megabyte-milliseconds taken by all reduce tasks=959764480
        Map-Reduce Framework
                Map input records=200000000
                Map output records=200000000
                Map output bytes=20400000000
                Map output materialized bytes=20800000900
                Input split bytes=19650
                Combine input records=0
                Combine output records=0
                Reduce input groups=200000000
                Reduce shuffle bytes=20800000900
                Reduce input records=200000000
                Reduce output records=200000000
                Spilled Records=872477568
                Shuffled Maps =150
                Failed Shuffles=0
                Merged Map outputs=150
                GC time elapsed (ms)=48612
                CPU time spent (ms)=1729850
                Physical memory (bytes) snapshot=42722770944
                Virtual memory (bytes) snapshot=297758896128
                Total committed heap usage (bytes)=30028070912
        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=20000000000
19/01/16 05:41:27 INFO terasort.TeraSort: done
```

Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus
application_1547614870205_0005	ubuntu	TeraSort	MAPREDUCE	default	0	Wed Jan 16 13:24:53 +0800	Wed Jan 16 13:41:25 +0800	FINISHED	SUCCEEDED

It spent 16 mins 32 seconds.

Validate

```
$ hadoop jar ~/hadoop/share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.2.jar
teravalidate terasort/output-20G terasort/check-20G
```

songwenli — ubuntu@ip-172-31-3-192: ~/hadoop — ssh -i .ssh/hadoop.pem ubu...

```
19/01/16 05:47:33 INFO mapreduce.Job:  map 100% reduce 100%
19/01/16 05:47:33 INFO mapreduce.Job: Job job_1547614870205_0006 completed succe
ssfully
19/01/16 05:47:33 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=95
                FILE: Number of bytes written=397165
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=20000000132
                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
                Rack-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=184856
                Total time spent by all reduces in occupied slots (ms)=2257
                Total time spent by all map tasks (ms)=184856
                Total time spent by all reduce tasks (ms)=2257
                Total vcore-milliseconds taken by all map tasks=184856
                Total vcore-milliseconds taken by all reduce tasks=2257
                Total megabyte-milliseconds taken by all map tasks=189292544
                Total megabyte-milliseconds taken by all reduce tasks=2311168
        Map-Reduce Framework
                Map input records=200000000
                Map output records=3
                Map output bytes=83
                Map output materialized bytes=95
                Input split bytes=132
                Combine input records=0
                Combine output records=0
                Reduce input groups=3
                Reduce shuffle bytes=95
                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)=1517
                CPU time spent (ms)=144520
                Physical memory (bytes) snapshot=501800960
                Virtual memory (bytes) snapshot=3963731968
                Total committed heap usage (bytes)=298844160
        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
```

c. Running the Python Code on Hadoop

Create directory and put the dataset into DFS

```
$ hdfs dfs -mkdir datasets
$ hdfs dfs -put ~/Downloads/ pyfiles
$ hdfs dfs -put ~/Downloads/Large_Dataset.txt datasets
```

Run task

```
$ hadoop jar ~/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.9.2.jar -file
~/MapReduce_WordCount/mapper.py -mapper ~/MapReduce_WordCount/mapper.py -file
~/MapReduce_WordCount/reducer.py -reducer ~/MapReduce_WordCount/reducer.py -input
datasets/Large_Dataset.txt -output output-wordcount
```

```
songwenli — ubuntu@ip-172-31-3-192: ~ — ssh -i .ssh/hadoop.pem ubuntu@...
19/01/17 06:05:20 INFO mapreduce.Job:  map 100% reduce 95%
19/01/17 06:05:26 INFO mapreduce.Job:  map 100% reduce 98%
19/01/17 06:05:29 INFO mapreduce.Job:  map 100% reduce 100%
19/01/17 06:05:30 INFO mapreduce.Job: Job job_1547701681590_0006 completed succe
ssfully
19/01/17 06:05:30 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=1234056790
                FILE: Number of bytes written=1864044476
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=420032924
                HDFS: Number of bytes written=98812261
                HDFS: Number of read operations=15
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=3
                Launched map tasks=7
                Launched reduce tasks=1
                Data-local map tasks=4
                Rack-local map tasks=3
                Total time spent by all maps in occupied slots (ms)=474860
                Total time spent by all reduces in occupied slots (ms)=156358
                Total time spent by all map tasks (ms)=474860
                Total time spent by all reduce tasks (ms)=156358
                Total vcore-milliseconds taken by all map tasks=474860
                Total vcore-milliseconds taken by all reduce tasks=156358
                Total megabyte-milliseconds taken by all map tasks=486256640
                Total megabyte-milliseconds taken by all reduce tasks=160110592
        Map-Reduce Framework
                Map input records=3474848
                Map output records=53108124
                Map output bytes=522761584
                Map output materialized bytes=628977856
                Input split bytes=452
                Combine input records=0
                Combine output records=0
                Reduce input groups=7496007
                Reduce shuffle bytes=628977856
                Reduce input records=53108124
                Reduce output records=7496007
                Spilled Records=157593576
                Shuffled Maps =4
                Failed Shuffles=0
                Merged Map outputs=4
                GC time elapsed (ms)=1074
                CPU time spent (ms)=282450
                Physical memory (bytes) snapshot=1430495232
                Virtual memory (bytes) snapshot=9889742848
                Total committed heap usage (bytes)=986185728
        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=420032472
        File Output Format Counters
                Bytes Written=98812261
19/01/17 06:05:30 INFO streaming.StreamJob: Output directory: output-wordcount
```

Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus
application_1547701681590_0006	ubuntu	streamjob7823205603775034048.jar	MAPREDUCE	default	0	Thu Jan 17 14:02:37 +0800 2019	Thu Jan 17 14:05:29 +0800 2019	FINISHED	SUCCEEDED

d. Compiling the Java WordCount program for MapReduce

Add environment variable

```
export HADOOP_CLASSPATH=${JAVA_HOME}/lib/tools.jar
```

Compile jar file

```
$ hadoop com.sun.tools.javac.Main WordCount.java
$ jar cf wc.jar WordCount*.class
```

Run

```
$ hadoop jar wc.jar WordCount datasets/Large_Dataset.txt output-java-wordcount
```

```
songwenli — ubuntu@ip-172-31-3-192: ~ — ssh -i .ssh/hadoop.pem ubuntu@...
19/01/17 05:58:13 INFO mapreduce.Job:  map 100% reduce 98%
19/01/17 05:58:14 INFO mapreduce.Job:  map 100% reduce 100%
19/01/17 05:58:14 INFO mapreduce.Job: Job job_1547701681590_0005 completed succe
ssfully
19/01/17 05:58:14 INFO mapreduce.Job: Counters: 51
        File System Counters
                FILE: Number of bytes read=481892688
                FILE: Number of bytes written=668442877
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=420032976
                HDFS: Number of bytes written=98812261
                HDFS: Number of read operations=15
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=3
                Launched map tasks=7
                Launched reduce tasks=1
                Data-local map tasks=4
                Rack-local map tasks=3
                Total time spent by all maps in occupied slots (ms)=355373
                Total time spent by all reduces in occupied slots (ms)=80869
                Total time spent by all map tasks (ms)=355373
                Total time spent by all reduce tasks (ms)=80869
                Total vcore-milliseconds taken by all map tasks=355373
                Total vcore-milliseconds taken by all reduce tasks=80869
                Total megabyte-milliseconds taken by all map tasks=363901952
                Total megabyte-milliseconds taken by all reduce tasks=82809856
        Map-Reduce Framework
                Map input records=3474848
                Map output records=53108124
                Map output bytes=628977832
                Map output materialized bytes=185559069
                Input split bytes=504
                Combine input records=73371746
                Combine output records=32017963
                Reduce input groups=7496007
                Reduce shuffle bytes=185559069
                Reduce input records=11754341
                Reduce output records=7496007
                Spilled Records=43772304
                Shuffled Maps =4
                Failed Shuffles=0
                Merged Map outputs=4
                GC time elapsed (ms)=1253
                CPU time spent (ms)=154020
                Physical memory (bytes) snapshot=1382178816
                Virtual memory (bytes) snapshot=9855324160
                Total committed heap usage (bytes)=979369984
        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=420032472
        File Output Format Counters
                Bytes Written=98812261
```

Show 20 entries

ID	User	Name	Application Type	Queue	Application Priority	StartTime	FinishTime	State	FinalStatus
application_1547701681590_0005	ubuntu	word count	MAPREDUCE	default	0	Thu Jan 17 13:56:36 +0800 2019	Thu Jan 17 13:58:13 +0800 2019	FINISHED	SUCCEEDED

Running time comparison

	Time Cost
Python (Hadoop streaming)	2 mins 52 s = 172 s
Java	1 mins 37 s = 97 s