

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

### Problem:

This programming assignment covers the Sort application implemented in 3 different ways: Java, Hadoop, and Spark. The sorting application can read large file, which is larger than system memory size and sort it in place. This assignment uses 2 datasets, 10 GB and 100GB dataset to benchmark the 3 approaches to sorting.

### Methodology:

**Shared-Memory Sort:** This program sorts data files of any size, which are larger than the memory. This program is implemented using Java language. This program implements external sorting algorithm that can handle massive amount of data. External sorting is used when data file is larger than main memory of computing device instead it uses external memory i.e. hard drive. This program uses multi-threading to take advantage of multi-core.

Shared-Memory external sorting program steps

1. Read input data file path and calculate size of file.
2. Find available main memory size of computing device.
3. Calculate data block size that fits into main memory
4. Read block size amount of data from input file and save it in external memory.
5. Follow step 4 up to end of input file.
6. Initialize the thread-pool with number of threads.
7. Assign each thread in thread pool with task in step 8 using executors.
8. Read each split file from step 4 and sort it using any  $n \log n$  algorithm, this program using merge sort algorithm to sort array of strings.
9. Save sorted file in external memory.
10. Merge sorted files in external memory and write it to output file.

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

**Hadoop Sort:** This program is implemented in java based on MapReduce programming model. This consists of map and reduce function which extends Hadoop mapper and reducer class respectively.

Implementation details of Hadoop sort are as follows.

1. Implement map function, which takes Key and value input as text format.
2. Implement reduce function, which takes key as text and List of values as value.
3. Set InputFormatClass as KeyValueTextInputFormat, This class brakes file into line and split each line into key and value based on separator.
4. Set mapper class to map function implemented in step1
5. Set reducer class to reduce function implemented in step2
6. Set OutputKeyClass and OutputValueClass as text.
7. Set FileInputFormatClass to input file path.
8. Set OutputFormatClass to output file path.
9. Run the sorting job and wait for completion.
10. Map function in step1 take its file each record as input in the form of key and value based on separator. Map functions outputs data file each record in form of key and value.
11. Hadoop framework implements sorting and shuffling after map function.
12. Reduce function take input from map function and gives output as key and value in sorted order.
13. Write sorted records to output file path.

Write a description of the function of each file, and what modifications you had to make to go from 1 node to multiple nodes

- 1) conf/master: This file lists IP address of machine runs secondary name nodes. Master file in master contains localhost.
- 2) conf/slaves: slave file have localhost entry by default. For master instance add IP address of master node and all slave node. For slave node add IP of its own in slaves file. This file is used to run data nodes and task trackers on slaves.

# **CS553 Programming Assignment #2**

## **Sort on Hadoop/Spark**

3) conf/core-site.xml: This file sets HDFS file system path and Hadoop temporary storage. This file informs Hadoop about name node execution. It contains configuration settings for HDFS I/O and port number to which name node daemon runs and listen.

Structure of file as follows

```
<configuration>
<property>
  <name>hadoop.tmp.dir</name>
  <value>/app/hadoop/tmp</value>
  <description>A base for other temporary directories.</description>
</property>
<property>
  <name>fs.default.name</name>
  <value>hdfs://MasterIP(ec2-52-87-241-215.compute-1.amazonaws.com:54310)</value>
</value>
</property>
</configuration>
```

4) conf/hdfs-site.xml: This file is used to configure HDFS daemons such as Name node, Data node and secondary name node. This file is used to set data block replication factor .i.e. number data replication per data block and setting HDFS permissions using Boolean value like true to enable permission, false to disable.

Structure of this file is as follows

```
<configuration>
<property>
  <name>dfs.replication</name>
  <value>1</value>
</property>
<property>
  <name>dfs.permissions</name>
  <value>false</value>
</property>
```

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

```
<property>
  <name>dfs.namenode.name.dir</name>
  <value>file:/usr/local/hadoop_store/hdfs/namenode</value>
</property>
<property>
  <name>dfs.datanode.data.dir</name>
  <value>file:/usr/local/hadoop_store/hdfs/datanode</value>
</property>
</configuration>
```

5) conf/mapred-site.xml: This file is used to configure map reduce framework and job tracker, task tracker. Map reduce job tracker runs at IP and port mentioned in mapreduce.jobtracker.address property. File can set map reduce framework name. This file is used to set different properties like mapred.system.dir, mapred.local.dir, mapred.tasktracker.{map|reduce}.tasks.maximum, fs.hosts/dfs.hosts.ex, mapred.hosts/mapred,

1) What is a Master node? What is a Slaves node?

Master node manages all other node such as data node, secondary nodes in cluster. Master node host various storage and processing services. Master manages following nodes to achieve fault tolerant

Name node: it manages HDFS storage.

Secondary name node:

Resource manager: This manages scheduling of tasks of Hadoop. It runs by yarn frame work.

Job tracker: It is used in 1 server job tracking.

Slave node is used to store and process data in Hadoop cluster. The different components involved in storage and computing of data are as follows.

# **CS553 Programming Assignment #2**

## **Sort on Hadoop/Spark**

Node manager: It manages individual slave node and reports back to resource manager.

Data node: This node is used to store the data in Hadoop cluster.

- 2) Why do we need to set unique available ports to those configuration files on a shared environment? What errors or side-effects will show if we use same port number for each user?

Unique ports are used to assign each port with unique functionality otherwise if we use same port for all services then all functionality get interrupt and communication problem between master and slave node. For example we use different port for HDFS filesystem, job tracker and Resource tracker.

- 3) How can we change the number of mappers and reducers from the configuration file?

Number of mappers and reducers can changed by setting “mapred.tasktracker.map.tasks.maximum” and “mapred.tasktracker.reduce.tasks.maximum” properties in mapred-site.xml

### **Spark Sort:**

This program is implemented using java. Sparks provides API to read data into sparks data structure called RDD (resilient distributed dataset). The map and reduce function are implemented on RDD. The in-memory primitives of sparks provides better performance compared to any other map-reduce framework.

Implementation details of Spark sort are as follows.

1. Read input data file into sparks RDD structure.
2. Implement Pair function which takes string as input and outputs key and value string as output. Pair function splits input string into two strings by splitting first 10 characters as key and remaining 90 characters as values. Pair function returns set of pair records.

# **CS553 Programming Assignment #2**

## **Sort on Hadoop/Spark**

3. Call sort by key on set of Pair records returned by step2 Pair function. This call sorts the set of pair records by key in ascending order.
4. The sorted set of record pair consists of parenthesis in each record. Remove parenthesis in each record by calling map function.
5. Save the sorted record set in text file format.
6. Parnellism and collect API calls can be used to collect parts of data into single dataset and save as single output text file.

### **Runtime environment settings**

#### **1. Shared Memory.**

1. Login to Amazon AWS account.
2. Launch c3.large instance using Ubuntu operating system AMI.
3. Install JDK on instance
4. Copy shared memory program to c3.large instance.
5. Compile program using javac compiler
6. Run program using java command

#### **2. Hadoop.**

##### **1. 1-Node**

1. Login to Amazon AWS account
2. Launch c3.large instance using Ubuntu operating system AMI.
3. Install updates
4. Install JDK on instance
5. Install SSH
6. Create and set-up SSH certificates to access its nodes, in case of 1 node localhost is its node.
7. Add newly created SSH certificate key to authorized keys.
8. Download Hadoop
9. Extract and move Hadoop folder /usr/local/Hadoop
10. Create name node and data node directory
11. Set up following configuration files.
  - a. ~./bashrc

# **CS553 Programming Assignment #2**

## **Sort on Hadoop/Spark**

Add following variables

```
export JAVA_HOME=JDK path (ex. /usr/lib/jvm/java-7-openjdk-amd64)
export HADOOP_INSTALL=Hadoop path (/usr/local/Hadoop)
export PATH=$PATH:$HADOOP_INSTALL/bin
export PATH=$PATH:$HADOOP_INSTALL/sbin
export HADOOP_MAPRED_HOME=$HADOOP_INSTALL
export HADOOP_COMMON_HOME=$HADOOP_INSTALL
export HADOOP_HDFS_HOME=$HADOOP_INSTALL
export YARN_HOME=$HADOOP_INSTALL
export
HADOOP_COMMON_LIB_NATIVE_DIR=$HADOOP_INSTALL/lib/native
export HADOOP_OPTS="-Djava.library.path=$HADOOP_INSTALL/lib"
```

b. Hadoop-env.sh

Set below variable

```
export JAVA_HOME=JDK Path (/usr/lib/jvm/java-7-openjdk-amd64)
```

c. core-site.xml

Create Hadoop temporary directory like /app/Hadoop/temp. Add below configuration in core-site.xml

```
<configuration>
<property>
  <name>hadoop.tmp.dir</name>
  <value>Directory Path (like /app/Hadoop/temp)</value>
</property>
<property>
  <name>fs.default.name</name>
  <value>Instance Public DNS of instance</value>
</property>
</configuration>
```

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

d. mapred-site.xml

Add below configuration

```
<configuration>
<property>
  <name>mapreduce.jobtracker.address</name>
  <value>hdfs://Public DNS of master instance</value>
</property>
<property>
  <name>mapreduce.framework.name</name>
  <value>yarn</value>
</property>
</configuration>
```

e. hdfs-site.xml

```
<configuration>
<property>
  <name>dfs.replication</name>
  <value>1</value>
</property>
<property>
  <name>dfs.permissions</name>
  <value>false</value>
</property>
<property>
  <name>dfs.namenode.name.dir</name>
  <value>file:/Name node directory (ex
usr/local/hadoop_store/hdfs/namenode) </value>
</property>
```

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

```
<property>
  <name>dfs.datanode.data.dir</name>
  <value>file:/Datanode directory (ex.
usr/local/hadoop_store/hdfs/datanode) </value>
</property>
</configuration>
```

12. Format HDFS file system
13. Start Hadoop by running start-all.sh
14. Run jps command to check Hadoop status like node manager, name node, resource manger , data node.

Command for above steps are mentioned in attached Hadoop\_install script.

### 2. Hadoop 17 node.

1. Follow steps mentioned in 1-node Hadoop set-up
2. Add PEM key using ssh-add command to secured communication between master and nodes.
3. Take the image of above instance
4. Launch other 16 c3.large instance using above image AMI.
5. Open slaves file in ..../hadoop/etc/hadoop/slaves in master. Add master public DNS IP and 16 slave node public DNS IP.
6. Open slaves file in ..../hadoop/etc/hadoop/slaves in each slave and add its own public DNS IP.
7. Start Hadoop in master by running ./start-all.sh

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

### 3. Spark

#### 1. 1 -Node.

1. Install 1- node Hadoop as mentioned above.
2. Download spark and extract
3. Can change sparks runtime setting in spark-defaults.conf file

```
Ex. spark.driver.memory      2g  
spark.executor.extraJavaOptions -XX:+UseConcMarkSweepGC  
spark.driver.maxResultSize   1500g  
spark.akka.frameSize        1000  
spark.default.parallelism   100  
spark.akka.timeout          100
```

4. Run start-all.sh in sparks sbin directory.
5. Run jps to check running status of sparks like worker node, master node.

#### 2. 17- Node

1. Create AMI of required storage space using mount command to club all space in instance.
2. Move to ec2 directory in spark directory
3. Set AWS\_ACCESS\_KEY\_ID using export command.
4. Set AWS\_SECRET\_ACCESS\_KEY using export command.
5. Run ./spark-ec2 -k <keypair> -i <key-file> -s <num-slaves> launch <cluster-name> to launch 17 node cluster.  
  
<keypair> is name of EC2 key pair  
<key-file> is the private key file for your key pair  
<num-slaves> is the number of slave nodes to launch  
<cluster-name> is the name of cluster.
6. Run ./spark-ec2 -k <keypair> -i <key-file> login <cluster-name> to login to cluster.  
  
where <keypair> is name of EC2 pair.  
<key-file> is private key file.

# **CS553 Programming Assignment #2**

## **Sort on Hadoop/Spark**

7. Run ./spark-submit <key-file> <jar-file> <hdfs input file path> <hdfs output file path>

### **Difficulties**

1. Spark 17 node cluster set up. Using of command Run ./spark-ec2 -k <keypair> -i <key-file> -s <num-slaves> launch <cluster-name> to launch 17 node set up fails with error connection refused or Public key access denied. This error occurs randomly.

OS used (Linux distribution, kernel): Ubuntu 14.04.3 LTS, Release: 14.04, Codename: trusty

ANT version: Apache Ant(TM) version 1.9.3 compiled on April 8 2014

Java version: java version "1.7.0\_95"

Hadoop version: 2.7.2

Spark version: 1.6.1

Draw graphs showing execution time data and speedup data. Please explain your results. Please include a comparison between the Shared Memory Sort compared to the Hadoop Sort and Spark Sort for 1 node and Hadoop and Spark at 16 nodes. Can you explain the difference in performance?

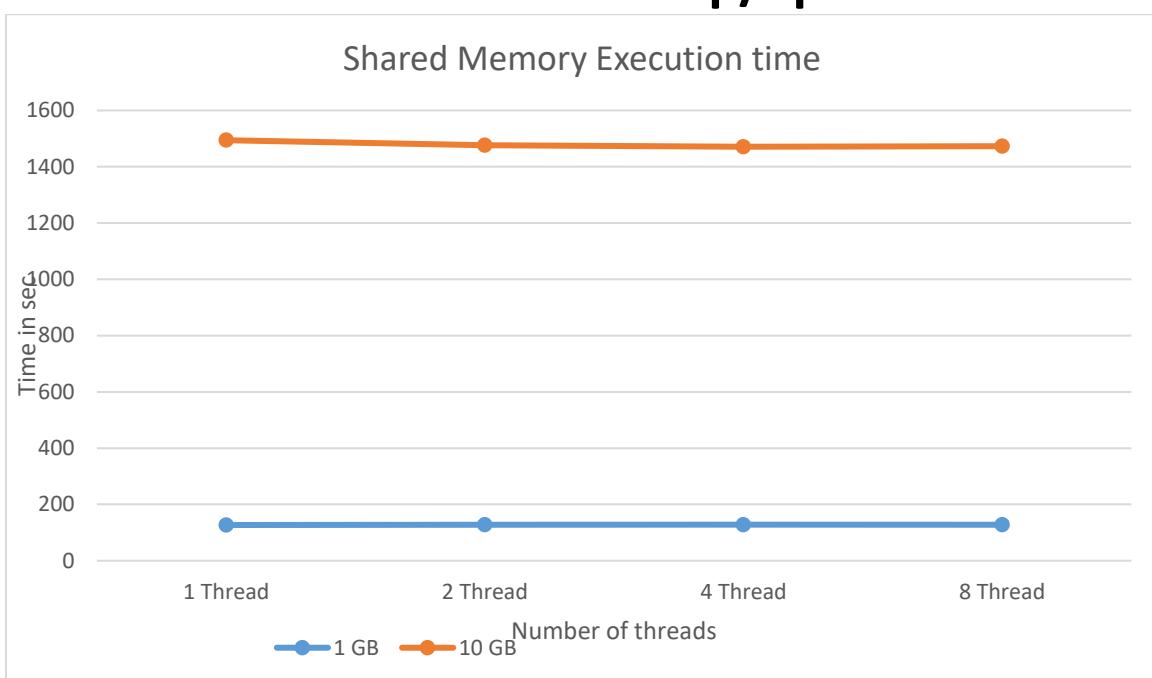
### **Performance**

#### **Shared Memory: Execution Time (sec)**

	1 Thread	2 Thread	4 Thread	8 Thread
1 GB	126.802	127.716	128.118	128.006
10 GB	1494.288	1476.089	1471.159	1473.229

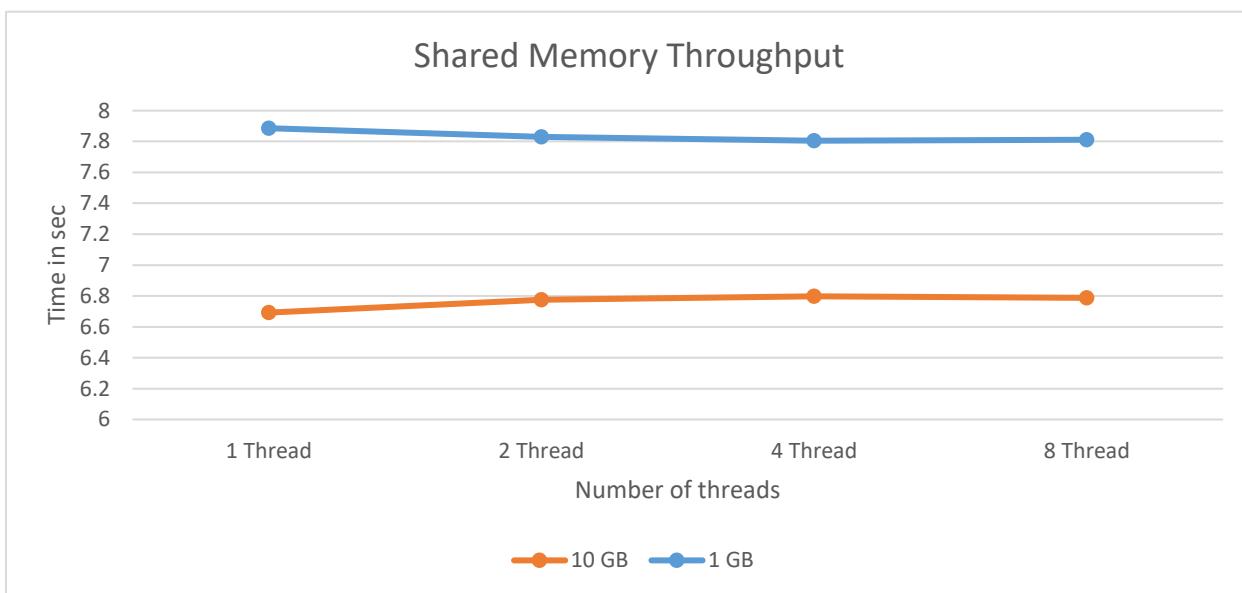
# CS553 Programming Assignment #2

## Sort on Hadoop/Spark



### Throughput ( mb/sec)

	1 Thread	2 Thread	4 Thread	8 Thread
1 GB	7.886	7.829	7.805	7.8121
10 GB	6.692	6.775	6.7974	6.7878

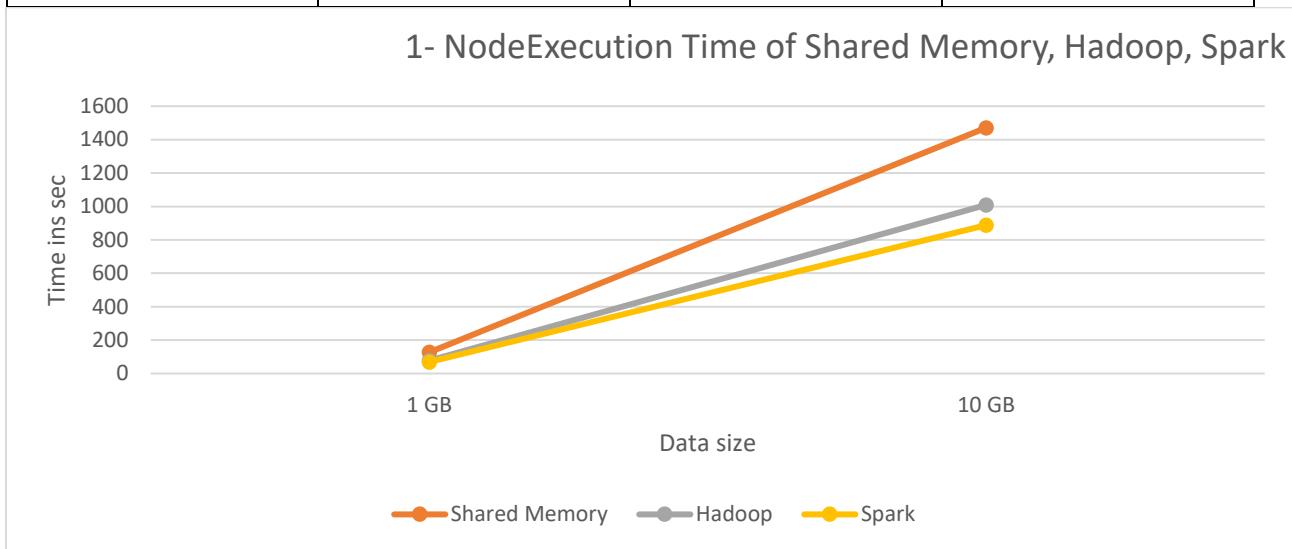


# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

1-Node execution time Shared memory, Hadoop, Spark.

	Shared Memory	Hadoop	Spark
1 GB	126.802	77	69
10 GB	1471.159	1009	888

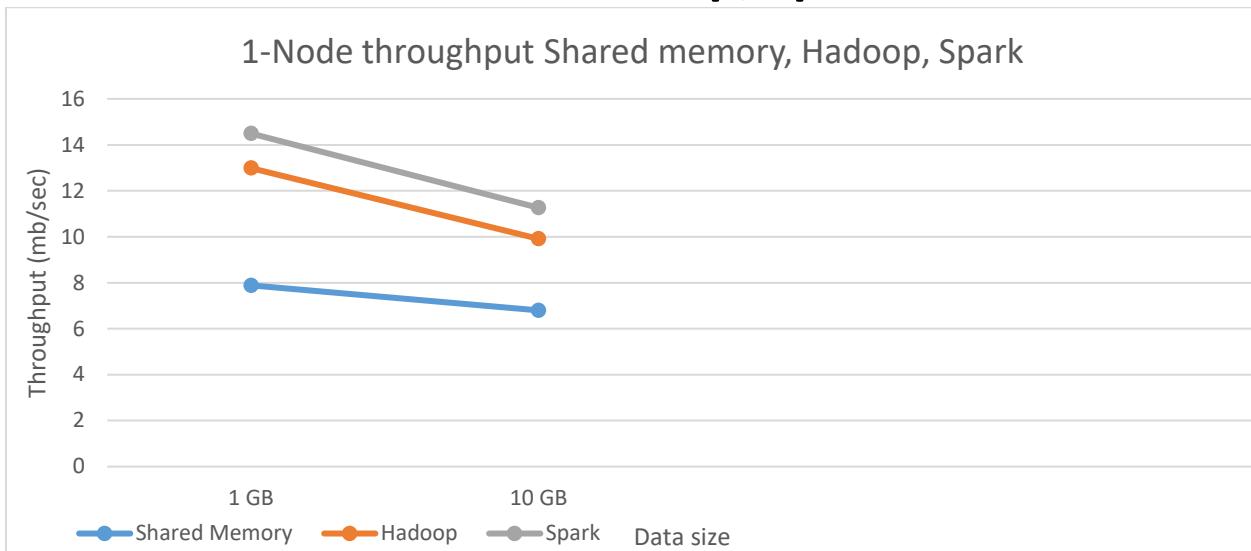


1-Node Throughput Shared memory, Hadoop, Spark

	Shared Memory	Hadoop	Spark
1 GB	7.886	12.987	14.493
10 GB	6.7974	9.911	11.2613

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark



### Comparison between shared memory, Hadoop, spark on 1 node

In all experiments, Spark is more efficient than Hadoop, shared memory because Spark in-memory primitive make it faster than Hadoop. Since my shared memory is not optimized code, Spark and Hadoop are faster i.e. taking less time to sort data than shared memory time.

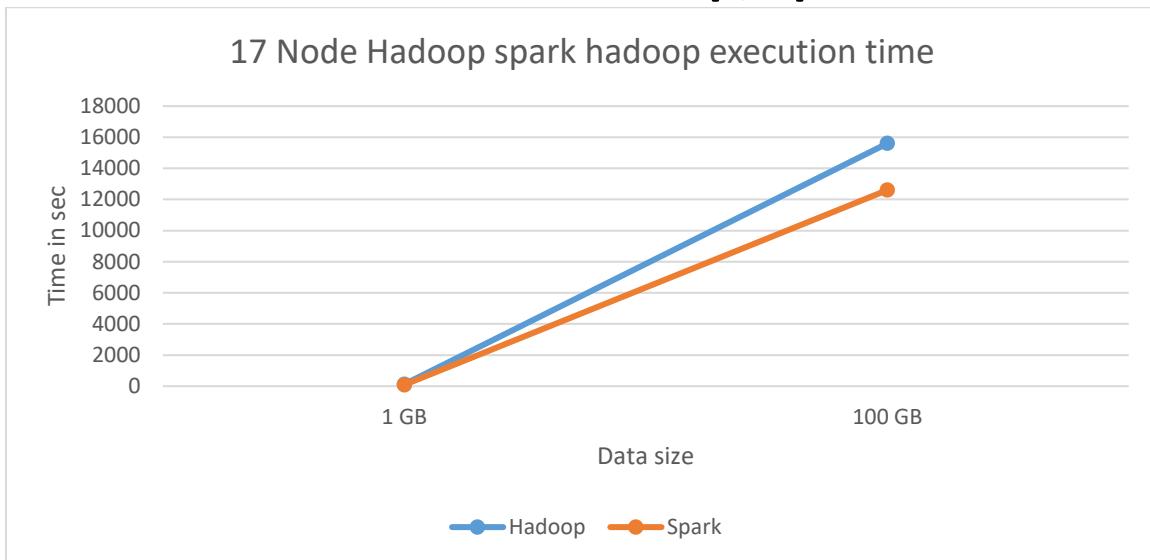
If shared memory code is optimized then it will be faster than Hadoop, spark because both Hadoop, spark takes some extra time schedule tasks and sorts data.

### 17- Node Hadoop, Spark execution time

	Hadoop	Spark
1 GB	127	91
100 GB	15600	12600

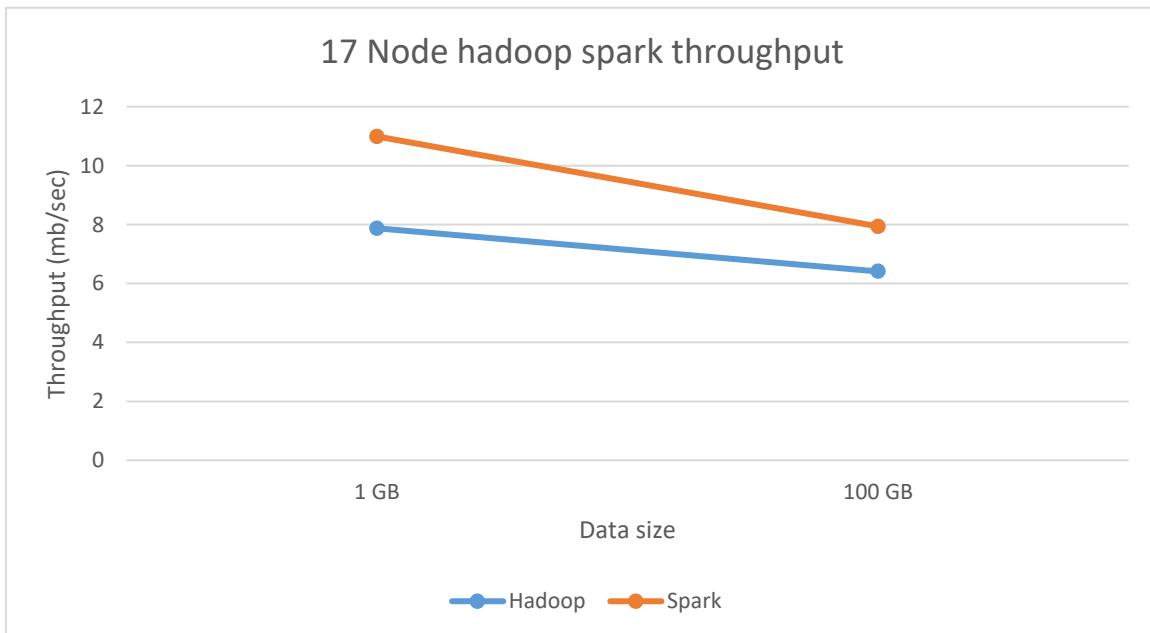
# CS553 Programming Assignment #2

## Sort on Hadoop/Spark



17 node Hadoop, spark throughput.

	Hadoop	Spark
1 GB	7.874	10.989
100 GB	6.41	7.9365



# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

Comparison between 1GB Shared memory at 1 node to Hadoop, Spark at 16 node

	Time (ms)
Shared Memory	126
Hadoop	127
Spark	91

Shared memory performance at 1 node is better than hadoop and spark performance at 16 node because in 16 node spark and Hadoop framework take time to schedule tasks, transfer data block between master and slave node, collect sorted data from master.

### Comparison between Hadoop, spark on 16 Node.

Spark performance is better than Hadoop in 16 node cluster because sparks API structure RDD (Resilient distributed data) is in-memory primitive so it sorts record more efficient than Hadoop. Hadoop splits data into block and stores each block of data as sorted file and then merges each sorted file so it takes more time than spark.

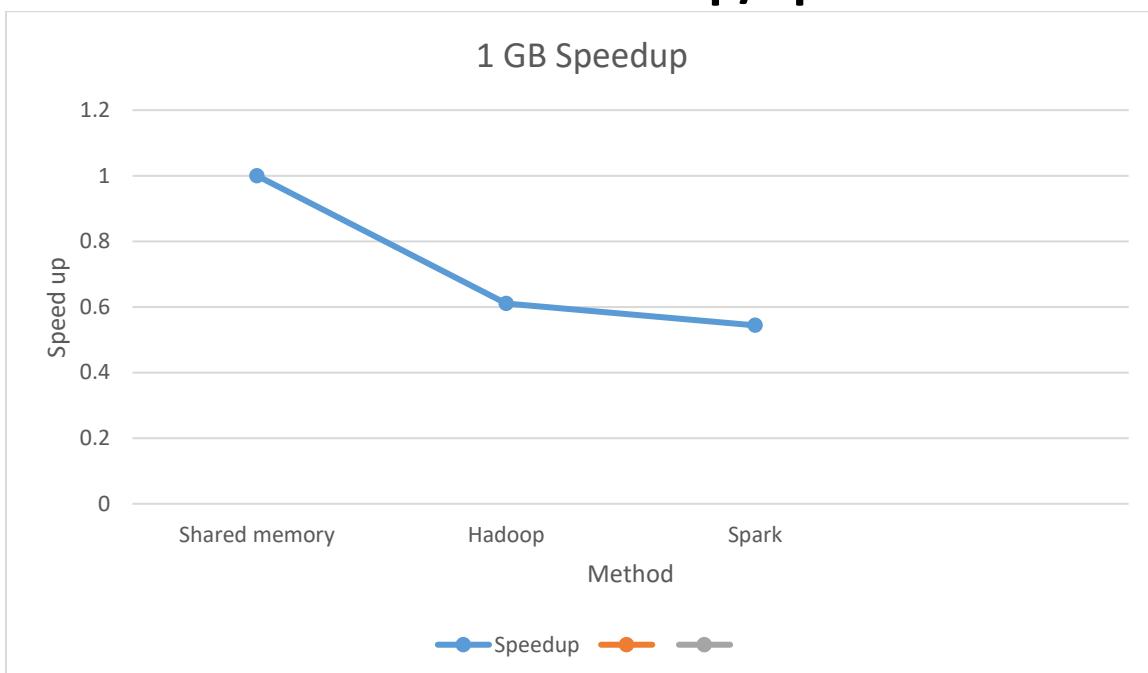
### Speedup

- 1 GB 1 Node Shared memory, Hadoop, Spark.

	Speedup	Time (in sec)
Shared memory	1	126.802
Hadoop	0.611	77
Spark	0.544	69

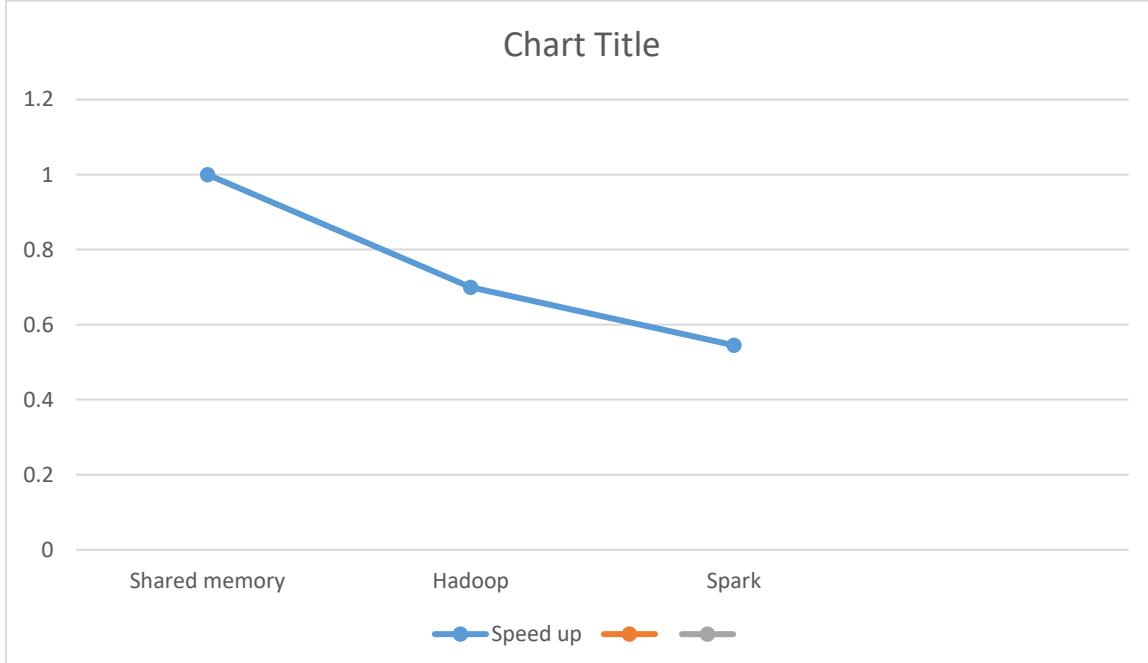
# CS553 Programming Assignment #2

## Sort on Hadoop/Spark



1 Node 10 GB speed up

	Speed up	Time
Shared memory	1	1471.159
Hadoop	0.70	1009
Spark	0.545	888

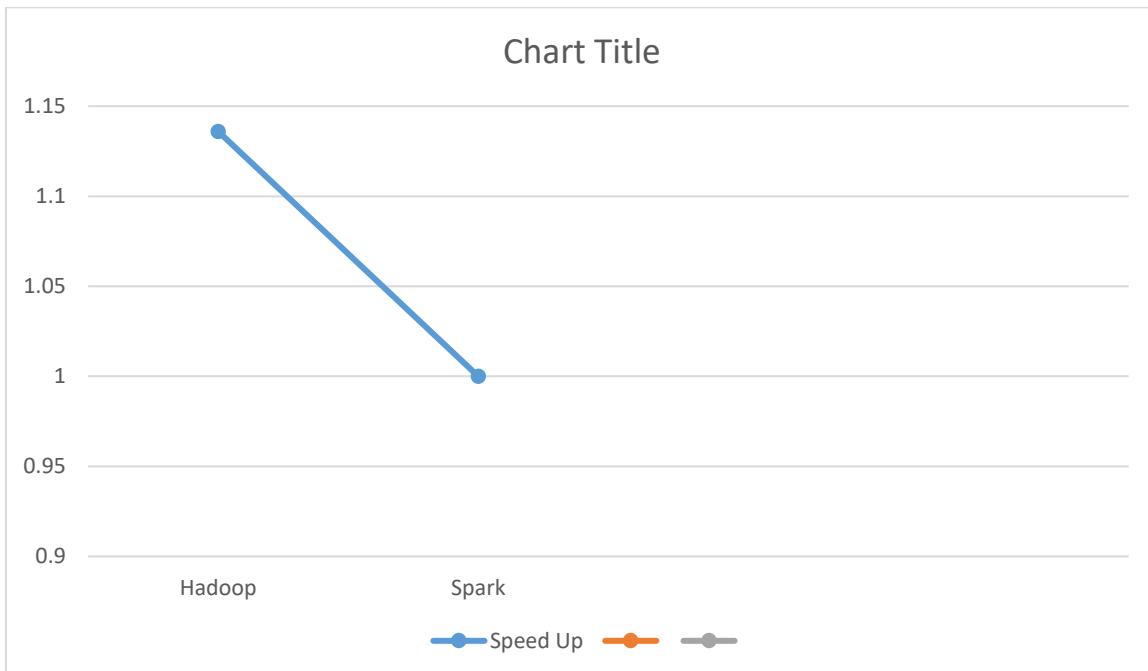


# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

1 Node 1GB Hadoop Spark

	Speed Up	Time
Hadoop	1.136	1009
Spark	1	888

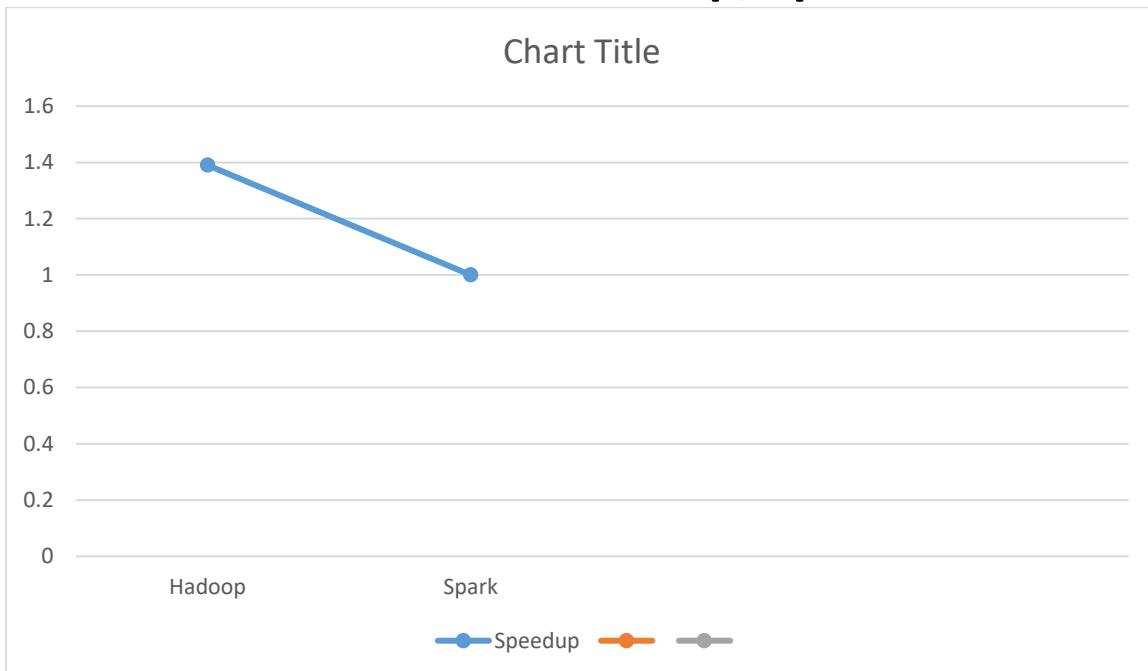


17 Node 1 GB Hadoop spark.

	Speedup	Time
Hadoop	1.39	127
Spark	1	91

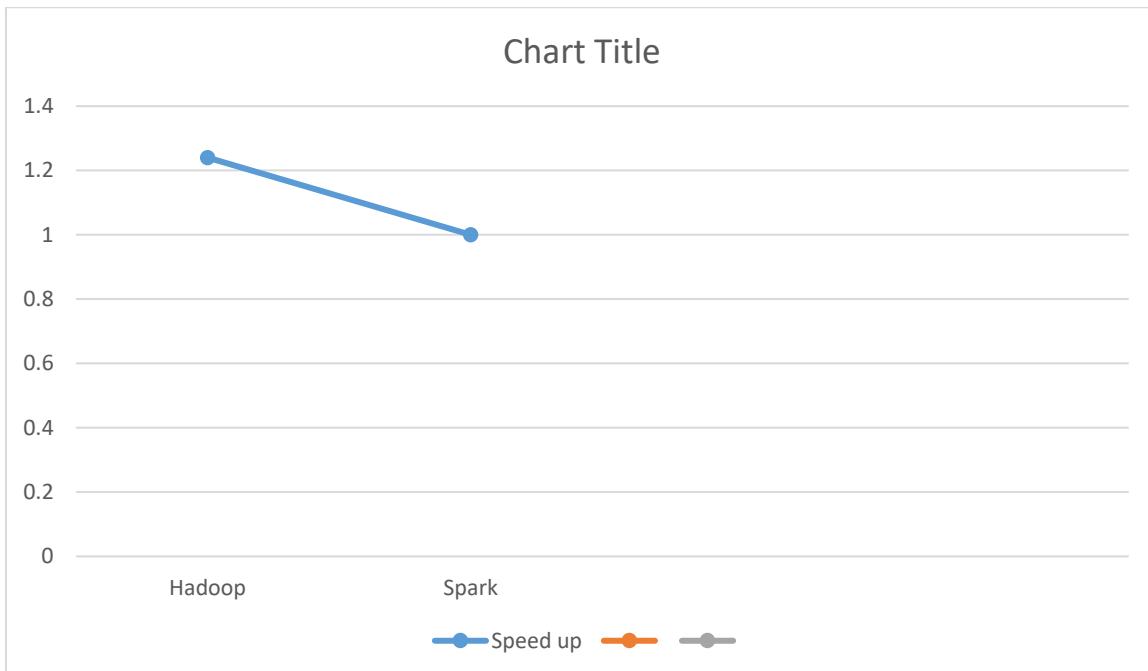
# CS553 Programming Assignment #2

## Sort on Hadoop/Spark



17 Node 100Gb Hadoop Spark speed up

	Speed up	Time	Nodes
Hadoop	1.24	15600	17
Spark	1	12600	17



# **CS553 Programming Assignment #2**

## **Sort on Hadoop/Spark**

### **Conclusion:**

Spark and Shared memory are good at 1 node scale. Spark is good at 16 nodes. Spark is good at 100 and 1000 node scales because of its in-memory primitive data structure RDD (Resilient distributed dataset).

### **Comparison with cloudsort**

#### **2014 winners**

Hadoop- 102.5 TB in 4,328 seconds

Throughput: 23682.99 mb/sec

2100 nodes (2 2.3Ghz hexcore Xeon E5-2630, 64 GB memory, 12x3TB disks)

Spark- 100 TB in 1,406 seconds

Throughput: 71123.75 mb/sec

207 Amazon EC2 i2.8xlarge nodes x

(32 vCores - 2.5Ghz Intel Xeon E5-2670 v2, 244GB memory, 8x800 GB SSD)

#### **2013 Winners**

Hadoop: 102.5 TB in 4,328 seconds

Throughput: 23682.99 mb/sec

2100 nodes x

(2 2.3Ghz hexcore Xeon E5-2630, 64 GB memory, 12x3TB disks)

2013 and 2014 winners have more throughput than our experiment throughput.

Because of system specifications .i.e. amount of memory, Disk type SSD.

Configurations settings.

### **Screen Shots**

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

## 1. Shared memory

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

ubuntu@ip-172-31-49-141:~\$  
ubuntu@ip-172-31-49-141:~\$

```
Thread: 12 working on file SplitFile2z731516791127830tmp
Thread: 11 done sorting on file SplitFile1z273151307830tmp
Thread: 12 working on file SplitFile1e5916663191398539tmp
Thread: 9 done sorting on file SplitFile1e631311801045528593tmp
Thread: 9 working on file SplitFile0231819507579873649tmp
Thread: 10 done sorting on file SplitFile2713252451580570080tmp
Thread: 10 working on file SplitFile4539825583019352216tmp
Thread: 11 done sorting on file SplitFile2872624858136808tmp
Thread: 12 working on file SplitFile1e599160663191398539tmp
Thread: 12 working on file SplitFile1218208551581747865tmp
Thread: 9 done sorting on file SplitFile1c31819507579873649tmp
Thread: 9 working on file SplitFile082116867701722816tmp
Thread: 10 working on file SplitFile1e599160663191398539tmp
Thread: 10 working on file SplitFile72835275904602017tmp
Thread: 12 done sorting on file SplitFile1j128208551581747865tmp
Thread: 11 done sorting on file SplitFile78826726248581362808tmp
Thread: 11 working on file SplitFile3993648861423621148tmp
Thread: 12 working on file SplitFile23917847615639923tmp
Thread: 12 done sorting on file SplitFile1e599160663191398539tmp
Thread: 9 working on file SplitFile47884726189196709844tmp
Thread: 12 done sorting on file SplitFile1e62391784513551039329tmp
Thread: 12 working on file SplitFile3271118361096680841tmp
Thread: 11 done sorting on file SplitFile1e5393648861423621148tmp
Thread: 11 working on file SplitFile949477785747883777tmp
Thread: 10 working on file SplitFile77404647530635095858tmp
Thread: 9 working on file SplitFile77404647530635095858tmp
Thread: 10 done sorting on file SplitFile7722852759046020317tmp
Thread: 10 working on file SplitFile1888081255461034977tmp
Thread: 12 done sorting on file SplitFile3271118361096680841tmp
Thread: 12 working on file SplitFile1e599160663191398539tmp
Thread: 12 working on file SplitFile1e599160663191398539tmp
Thread: 11 working on file SplitFile5876627161117089663tmp
Thread: 10 done sorting on file SplitFile1e888081255461613997tmp
Thread: 10 working on file SplitFile5132708945377419948tmp
Thread: 9 done sorting on file SplitFile77404647530635095858tmp
Thread: 9 working on file SplitFile1e5393648861423621148tmp
Thread: 12 working on file SplitFile1e599160663191398539tmp
Thread: 12 working on file SplitFile1e599160663191398539tmp
Thread: 11 done sorting on file SplitFile5876627161117089663tmp
Thread: 11 working on file SplitFile5429432495313844954tmp
Thread: 10 done sorting on file SplitFile5132708945377419948tmp
Thread: 10 working on file SplitFile1e599160663191398539tmp
Thread: 10 working on file SplitFile1e4161574962236181tmp
Thread: 9 working on file SplitFile7386582882526285638tmp
Thread: 12 done sorting on file SplitFile0072745931264732532tmp
Thread: 12 working on file SplitFile340218520370785080tmp
Thread: 12 done sorting on file SplitFile340218520370785080tmp
Thread: 12 working on file SplitFile1e599160663191398539tmp
Thread: 12 working on file SplitFile1e599160663191398539tmp
Thread: 10 working on file SplitFile1e505484561653329877tmp
Thread: 9 done sorting on file SplitFile7386582882526285638tmp
Thread: 9 working on file SplitFile3059588010610496210tmp
Thread: 11 done sorting on file SplitFile5429432495313844954tmp
Thread: 11 working on file SplitFile7960423438322455377tmp
```

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

## 2. Hadoop

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

```
hduser@ip-172-31-49-141:~/usr/local/hadoop
$ ./bin/native-code-loader.sh
16/03/28 21:21:28 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/03/28 21:21:29 INFO client.RMProxy: Connecting to ResourceManager at ec2-52-47-241-215.compute-1.amazonaws.com/172.31.49.141:8032
16/03/28 21:21:29 WARN mapred.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
16/03/28 21:21:30 INFO input.FileInputFormat: Total input paths to process : 1
16/03/28 21:21:30 INFO mapred.JobSubmitter: number of splits:1
16/03/28 21:21:30 INFO mapred.JobClient: Running job: job_1459128097292_0002
16/03/28 21:21:30 INFO mapred.YarnClientImpl: Submitted application application_1459128097292_0002
16/03/28 21:21:30 INFO mapred.Job: The url to track the job: http://ec2-52-87-241-215.compute-1.amazonaws.com:8088/proxy/application_1459128097292_0002/
16/03/28 21:21:31 INFO mapred.Job: Running Job: job_1459128097292_0002
16/03/28 21:21:37 INFO mapred.Job: Job job_1459128097292_0002 running in uber mode : false
16/03/28 21:21:37 INFO mapred.Job: map 0% reduce 0%
16/03/28 21:21:37 INFO mapred.Job: map 12% reduce 0%
16/03/28 21:21:37 INFO mapred.Job: map 34% reduce 0%
16/03/28 21:21:37 INFO mapred.Job: map 48% reduce 0%
16/03/28 21:21:37 INFO mapred.Job: map 52% reduce 0%
16/03/28 21:21:37 INFO mapred.Job: map 56% reduce 0%
16/03/28 21:21:37 INFO mapred.Job: map 59% reduce 0%
16/03/28 21:21:37 INFO mapred.Job: map 61% reduce 0%
16/03/28 21:21:37 INFO mapred.Job: map 67% reduce 0%
16/03/28 21:22:39 INFO mapred.Job: map 68% reduce 0%
16/03/28 21:22:39 INFO mapred.Job: map 75% reduce 0%
16/03/28 21:22:39 INFO mapred.Job: map 80% reduce 17%
16/03/28 21:22:39 INFO mapred.Job: map 83% reduce 31%
16/03/28 21:23:05 INFO mapred.Job: map 96% reduce 29%
16/03/28 21:23:05 INFO mapred.Job: map 100% reduce 29%
16/03/28 21:23:11 INFO mapred.Job: map 100% reduce 67%
16/03/28 21:23:14 INFO mapred.Job: map 100% reduce 72%
16/03/28 21:23:17 INFO mapred.Job: map 100% reduce 78%
16/03/28 21:23:23 INFO mapred.Job: map 100% reduce 113%
16/03/28 21:23:23 INFO mapred.Job: map 100% reduce 88%
16/03/28 21:23:26 INFO mapred.Job: map 100% reduce 94%
16/03/28 21:23:29 INFO mapred.Job: map 100% reduce 100%
16/03/28 21:23:31 INFO mapred.Job: Job job_1459128097292_0002 completed successfully
16/03/28 21:23:31 INFO mapred.Job: Counters: 50
File System Counters
  FILE: Number of bytes read=197834714
  FILE: Number of bytes written=29939373967
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=10000000000
  HDFS: Number of bytes written=10000000000
  HDFS: Number of read operations=27
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=2
Job Counters
  Killed map tasks=1
  Launched map tasks=9
  Launched reduce tasks=1
  Data-local map tasks=9
```

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

```
hadoop@ip-172-31-27-143:~/user/local/hadoop
```

x vishwanath@vishwanath-Q551L: ~/Documents

```
16/03/25 05:47:47 INFO reduce.OnDiskMapOutput: Read 136902060 bytes from map-output for attempt_local1144051581_0001_m_000055_0
16/03/25 05:47:47 INFO reduce.MergeManagerImpl: attempt_local1144051581_0001_m_000034_0: Shuffling to disk since 136902065 is greater than maxSingleShuffleLimit (93860656)
16/03/25 05:47:49 INFO mapred.LocalJobRunner: localfetcher#1 about to shuffle output for attempt_local1144051581_0001_m_000055_0: decomp: 136902064 len: 136902064 to DISK
16/03/25 05:47:49 INFO mapred.LocalJobRunner: reduce > copy taskattempt_local1144051581_0001_r_000000_0 finished merging 10 map output files on disk of total-size 1379716496. Local output file is /app/hadoop/tmp/mapred/local/runner/hduser/job_local1144051581_0001/_attempt_local1144051581_0001/r_000000_0/output.map2.out.merged.of.size.1369020954
16/03/25 05:47:50 INFO reduce.MergeManagerImpl: OnDiskMerger: We have 10 map outputs on disk. Triggering merge...
16/03/25 05:47:50 INFO mapred.Merger: Merging 10 sorted segments
16/03/25 05:47:50 INFO mapred.Merger: Map output 10 segments to the same key-pass, with 10 segments left of total size: 1369019856 bytes
16/03/25 05:47:50 INFO mapred.Merger: Job: map 100% reduce 31%
16/03/25 05:47:52 INFO reduce.OnDiskMapOutput: Read 136902060 bytes from map-output for attempt_local1144051581_0001_m_000087_0
16/03/25 05:47:52 INFO reduce.MergeManagerImpl: attempt_local1144051581_0001_m_000087_0: Shuffling to disk since 136902158 is greater than maxSingleShuffleLimit (93860656)
16/03/25 05:47:52 INFO mapred.LocalJobRunner: localfetcher#1 about to shuffle output of map attempt_local1144051581_0001_m_000087_0: decomp: 136902158 len: 136902158 to DISK
16/03/25 05:47:53 INFO mapred.LocalJobRunner: reduce > copy taskattempt_local1144051581_0001_m_000088_0 succeeded at 136959..98 MB/s Aggregated copy rate(70 of 75 at 907467.00 MB/s)
16/03/25 05:47:53 INFO mapred.LocalJobRunner: reduce > copy taskattempt_local1144051581_0001_m_000089_0 succeeded at 136959..98 MB/s Aggregated copy rate(70 of 75 at 907467.00 MB/s)
16/03/25 05:47:54 INFO reduce.OnDiskMapOutput: Read 136902162 bytes from map-output for attempt_local1144051581_0001_m_000087_0
16/03/25 05:47:54 INFO reduce.MergeManagerImpl: attempt_local1144051581_0001_m_000087_0: Shuffling to disk since 136902056 is greater than maxSingleShuffleLimit (93860656)
16/03/25 05:47:54 INFO mapred.LocalJobRunner: localfetcher#1 about to shuffle output of map attempt_local1144051581_0001_m_000086_0: decomp: 136902056 len: 136902056 to DISK
16/03/25 05:47:55 INFO mapred.LocalJobRunner: reduce > copy taskattempt_local1144051581_0001_m_000087_0 succeeded at 136956..07 MB/s Aggregated copy rate(71 of 75 at 9205238.00 MB/s)
16/03/25 05:48:02 INFO reduce.OnDiskMapOutput: Read 136902060 bytes from map-output for attempt_local1144051581_0001_m_000086_0
16/03/25 05:48:02 INFO reduce.MergeManagerImpl: attempt_local1144051581_0001_m_000086_0: Shuffling to disk since 136902056 is greater than maxSingleShuffleLimit (93860656)
16/03/25 05:48:02 INFO mapred.LocalJobRunner: localfetcher#1 about to shuffle output of map attempt_local1144051581_0001_m_000087_0: decomp: 136902056 len: 136902056 to DISK
16/03/25 05:48:04 INFO mapred.LocalJobRunner: reduce > copy taskattempt_local1144051581_0001_m_000087_0 succeeded at 136959..98 MB/s Aggregated copy rate(72 of 75 at 9335799.00 MB/s)
16/03/25 05:48:07 INFO mapred.LocalJobRunner: reduce > copy taskattempt_local1144051581_0001_m_000088_0 succeeded at 136959..98 MB/s Aggregated copy rate(73 of 75 at 9335799.00 MB/s)
16/03/25 05:48:07 INFO reduce.MergeManagerImpl: attempt_local1144051581_0001_m_000088_0: Shuffling to disk since 136902056 is greater than maxSingleShuffleLimit (93860656)
16/03/25 05:48:07 INFO mapred.LocalJobRunner: localfetcher#1 about to shuffle output of map attempt_local1144051581_0001_m_000089_0: decomp: 136902056 len: 136902056 to DISK
16/03/25 05:48:07 INFO mapred.LocalJobRunner: reduce > copy taskattempt_local1144051581_0001_m_000089_0 succeeded at 136959..98 MB/s Aggregated copy rate(73 of 75 at 9466358.00 MB/s)
16/03/25 05:48:10 INFO mapred.LocalJobRunner: reduce > copy taskattempt_local1144051581_0001_m_000090_0 succeeded at 136959..98 MB/s Aggregated copy rate(73 of 75 at 9466358.00 MB/s)
16/03/25 05:48:12 INFO reduce.OnDiskMapOutput: Read 136902060 bytes from map-output for attempt_local1144051581_0001_m_000090_0
16/03/25 05:48:12 INFO reduce.MergeManagerImpl: attempt_local1144051581_0001_m_000090_0: Shuffling to disk since 136902056 is greater than maxSingleShuffleLimit (93860656)
16/03/25 05:48:12 INFO mapred.LocalJobRunner: localfetcher#1 about to shuffle output of map attempt_local1144051581_0001_m_000091_0: decomp: 136902158 len: 136902158 to DISK
16/03/25 05:48:13 INFO mapred.LocalJobRunner: reduce > copy taskattempt_local1144051581_0001_m_000091_0 succeeded at 136959..98 MB/s Aggregated copy rate(74 of 75 at 9596918.00 MB/s)
16/03/25 05:48:14 INFO mapred.LocalJobRunner: map 100% reduce 33%
16/03/25 05:48:16 INFO mapred.LocalJobRunner: reduce > copy taskattempt_local1144051581_0001_m_000093_0 succeeded at 136959..98 MB/s Aggregated copy rate(74 of 75 at 9596918.00 MB/s)
16/03/25 05:48:18 INFO reduce.OnDiskMapOutput: Read 136902162 bytes from map-output for attempt_local1144051581_0001_m_000095_0
16/03/25 05:48:18 INFO mapred.LocalJobRunner: localfetcher#1 about to merge on disk map-output: Starting merge with 10 segments, while ignoring 9 segments
16/03/25 05:48:18 INFO reduce.EventFetcher: EventFetcher IS interrupted.. Returning
16/03/25 05:48:18 INFO mapred.LocalJobRunner: 75 / 75 copied.
16/03/25 05:48:19 INFO mapred.LocalJobRunner: reduce > sort
16/03/25 05:48:22 INFO mapred.LocalJobRunner: reduce > sort
16/03/25 05:48:25 INFO mapred.LocalJobRunner: reduce > sort
16/03/25 05:48:28 INFO mapred.LocalJobRunner: reduce > sort
16/03/25 05:48:29 INFO reduce.MergeManagerImpl: attempt_local1144051581_0001_r_000000_0 finished merging 10 map output files on disk of total-size 1379716390. Local output file is /app/hadoop/tmp/mapred/local/runner/hduser/job_local1144051581_0001/_attempt_local1144051581_0001_r_000000_0/output.map2.out.merged.of.size.1369020852
16/03/25 05:48:29 INFO reduce.MergeManagerImpl: OnDiskMerger: We have 10 map outputs on disk. Triggering merge...
16/03/25 05:48:29 INFO mapred.Merger: Merging 10 sorted segments
16/03/25 05:48:29 INFO mapred.Merger: Map output 10 segments to the same key-pass, with 10 segments left of total size: 1369019958 bytes
16/03/25 05:48:31 INFO mapred.LocalJobRunner: reduce > sort
16/03/25 05:48:34 INFO mapred.LocalJobRunner: reduce > sort
16/03/25 05:48:37 INFO mapred.LocalJobRunner: reduce > sort
16/03/25 05:48:40 INFO mapred.LocalJobRunner: reduce > sort
```

```
hadoop@ip-172-31-27-143: /usr/local/hadoop
hadoop@ip-172-31-27-143: /usr/local/hadoop
vishwanath@vishwanath-Q551LB: ~/Documents
```

16/03/25 05:40:34 INFO mapred.LocalJobRunner: Finishing task: attempt\_local1144051581\_0001\_m\_000046\_0  
16/03/25 05:40:34 INFO mapred.LocalJobRunner: Starting task: attempt\_local1144051581\_0001\_m\_000047\_0  
16/03/25 05:40:34 INFO mapred.Task: Map output collector class = org.apache.hadoop.mapred.MapOutputBuffer  
16/03/25 05:40:34 INFO mapred.Task: Using ResourceCalculatorProcessTree : []  
16/03/25 05:40:34 INFO mapred.MapTask: Processing split: hdfs://ec2-54-174-17-233.compute-1.amazonaws.com:54310/user/hduser/input/10gb\_data.txt:6308233216+134217728  
16/03/25 05:40:34 INFO mapred.MapTask: (EQUATOR) 0 kvl 26214396(104857584)  
16/03/25 05:40:34 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100  
16/03/25 05:40:34 INFO mapred.MapTask: soft limit at 83880080  
16/03/25 05:40:34 INFO mapred.MapTask: bufstart = 0; bufend = 104857600  
16/03/25 05:40:34 INFO mapred.MapTask: kvstart = 26214396; length = 6553600  
16/03/25 05:40:34 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask\$MapOutputBuffer  
16/03/25 05:40:35 INFO mapred.MapTask: Spilling map output  
16/03/25 05:40:35 INFO mapred.MapTask: bufstart = 0; bufend = 72315600; bufvoid = 104857600  
16/03/25 05:40:35 INFO mapred.MapTask: kvstart = 26214396; kvend = 23321776(93287104); length = 2892621/6553600  
16/03/25 05:40:35 INFO mapred.MapTask: (EQUATOR) 75280208 kvl 18802048(75208192)  
16/03/25 05:40:35 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100  
16/03/25 05:40:36 INFO mapred.MapTask: Finished spill 0  
16/03/25 05:40:36 INFO mapred.MapTask: (RESET) equator 75280208 kvl 18802048(75208192) kvl 18078904(72315616)  
16/03/25 05:40:37 INFO mapred.LocalJobRunner:  
16/03/25 05:40:37 INFO mapred.MapTask: Starting flush of map output  
16/03/25 05:40:37 INFO mapred.MapTask: Spilling map output  
16/03/25 05:40:37 INFO mapred.MapTask: bufstart = 75280208; bufend = 32252800; bufvoid = 104857500  
16/03/25 05:40:37 INFO mapred.MapTask: kvstart = 18802048(75208192); kvend = 16325968(65383872); length = 2476881/6553600  
16/03/25 05:40:37 INFO mapreduce.Job: map 63% reduce 0%  
16/03/25 05:40:38 INFO mapred.MapTask: Finished spill 1  
16/03/25 05:40:38 INFO mapred.Merger: Merging 2 sorted segments  
16/03/25 05:40:38 INFO mapred.Merger: Merged new large part, with 2 segments left of total size: 136991864 bytes  
16/03/25 05:40:38 INFO mapred.Task: Task attempt\_local1144051581\_0001\_m\_000047\_0 is done. And is in the process of committing  
16/03/25 05:40:40 INFO mapred.LocalJobRunner: hdfs://ec2-54-174-17-233.compute-1.amazonaws.com:54310/user/hduser/input/10gb\_data.txt:6308233216+134217728 > sort  
16/03/25 05:40:40 INFO mapred.Task: Task attempt\_local1144051581\_0001\_m\_000047\_0 done.  
16/03/25 05:40:40 INFO mapred.LocalJobRunner: Finishing task: attempt\_local1144051581\_0001\_m\_000047\_0  
16/03/25 05:40:40 INFO mapred.LocalJobRunner: Starting task: attempt\_local1144051581\_0001\_m\_000048\_0  
16/03/25 05:40:40 INFO mapred.Task: Using ResourceCalculatorProcessTree : []  
16/03/25 05:40:40 INFO mapred.MapTask: Processing split: hdfs://ec2-54-174-17-233.compute-1.amazonaws.com:54310/user/hduser/input/10gb\_data.txt:6442450944+134217728  
16/03/25 05:40:40 INFO mapred.MapTask: (EQUATOR) 0 kvl 26214396(104857584)  
16/03/25 05:40:40 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100  
16/03/25 05:40:40 INFO mapred.MapTask: soft limit at 83880080  
16/03/25 05:40:40 INFO mapred.MapTask: bufstart = 0; bufend = 104857600  
16/03/25 05:40:40 INFO mapred.MapTask: kvstart = 26214396; length = 6553600  
16/03/25 05:40:40 INFO mapred.MapTask: Map output collector class = org.apache.hadoop.mapred.MapTask\$MapOutputBuffer  
16/03/25 05:40:40 INFO mapreduce.Job: map 100% reduce 0%  
16/03/25 05:40:40 INFO mapred.MapTask: Spilling map output  
16/03/25 05:40:40 INFO mapred.MapTask: bufstart = 0; bufend = 72315600; bufvoid = 104857600  
16/03/25 05:40:40 INFO mapred.MapTask: kvstart = 26214396; kvend = 23321776(93287104); length = 2892621/6553600  
16/03/25 05:40:40 INFO mapred.MapTask: (EQUATOR) 75280208 kvl 18802048(75208192)  
16/03/25 05:40:40 INFO mapred.MapTask: mapreduce.task.io.sort.mb: 100  
16/03/25 05:40:42 INFO mapred.MapTask: Finished spill 0  
16/03/25 05:40:42 INFO mapred.MapTask: (RESET) equator 75280208 kvl 18802048(75208192) kvl 18078904(72315616)  
16/03/25 05:40:42 INFO mapred.LocalJobRunner:  
16/03/25 05:40:42 INFO mapred.MapTask: Starting flush of map output  
16/03/25 05:40:42 INFO mapred.MapTask: Spilling map output  
16/03/25 05:40:42 INFO mapred.MapTask: bufstart = 75280208; bufend = 32252800; bufvoid = 104857500  
16/03/25 05:40:42 INFO mapred.MapTask: kvstart = 18802048(75208192); kvend = 16325968(65383872); length = 2476881/6553600  
16/03/25 05:40:43 INFO mapred.LocalJobRunner: hdfs://ec2-54-174-17-233.compute-1.amazonaws.com:54310/user/hduser/input/10gb\_data.txt:6442450944+134217728 > sort  
16/03/25 05:40:43 INFO mapreduce.Job: map 65% reduce 0%

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

```
hduser@ip-172-31-27-143:~/user/hduser$ ./sort_hadoop
16/03/25 05:52:35 INFO mapreduce.Job: map 100% reduce 99%
16/03/25 05:52:37 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:52:40 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:52:41 INFO mapreduce.Job: map 100% reduce 100%
16/03/25 05:52:43 INFO mapred.Task: Task attempt_local1144051581_0001_r_000000_0 is done. And is in the process of committing
16/03/25 05:52:43 INFO mapred.Task: Task attempt_local1144051581_0001_r_000000_0 is allowed to commit now
16/03/25 05:52:43 INFO mapred.FileOutputCommitter: Saved output of task 'attempt_local1144051581_0001_r_000000_0' to hdfs://ec2-54-174-17-233.compute-1.amazonaws.com:54310/user/hduser/output/_temporary/0/t
ask_local1144051581_0001_r_000000
16/03/25 05:52:43 INFO mapred.LocalJobRunner: reduce task executor completed
16/03/25 05:52:44 INFO mapreduce.Job: Job local1144051581_0001 completed successfully
16/03/25 05:52:45 INFO mapreduce.Job: Counters: 35
  File System Counters
    FILE: Number of bytes read=390439017948
    FILE: Number of bytes written=22547797212
    FILE: Number of read operations=0
    FILE: Number of large read operations=0
    FILE: Number of write operations=0
    HDFS: Number of bytes read=92466167808
    HDFS: Number of bytes written=108900000000
    HDFS: Number of read operations=0
    HDFS: Number of large read operations=0
    HDFS: Number of write operations=78
  Map-Reduce Framework
    Map input records=100000000
    Map output records=100000000
    Map output materialized bytes=10200000450
    Input split bytes=11325
    Combine input records=0
    Combine output records=0
    Reduce input records=100000000
    Reduce output file bytes=10200000450
    Reduce input records=100000000
    Reduce output records=100000000
    Spilled Records=395957882
    Shuffled Maps =75
    Failed Shuffles=0
    Merged Map outputs=75
    GC time elapsed (ms)=13261
    Total committed heap usage (bytes)=39611531264
  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=100000000000
  File Output Format Counters
    Bytes Written=100000000000
hduser@ip-172-31-27-143:~/user/hduser$
```

```
hduser@ip-172-31-27-143:~/user/hduser$ ./sort_hadoop
16/03/25 05:48:55 INFO mapred.LocalJobRunner: reduce > sort
16/03/25 05:48:56 INFO mapreduce.Job: map 100% reduce 41%
16/03/25 05:48:58 INFO mapred.LocalJobRunner: reduce > sort
16/03/25 05:48:59 INFO mapreduce.Job: map 100% reduce 61%
16/03/25 05:49:01 INFO mapred.LocalJobRunner: reduce > merge, performing the merge-pass, with 10 segments left of total size: 10199999010 bytes
16/03/25 05:49:01 INFO configuration.deprecation: mapred.skip.on is deprecated. Instead, use mapreduce.job.skiprecords
16/03/25 05:49:01 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:49:02 INFO mapreduce.Job: map 100% reduce 67%
16/03/25 05:49:04 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:49:05 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:49:06 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:49:08 INFO mapreduce.Job: map 100% reduce 68%
16/03/25 05:49:10 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:49:13 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:49:16 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:49:17 INFO mapreduce.Job: map 100% reduce 69%
16/03/25 05:49:18 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:49:22 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:49:23 INFO mapreduce.Job: map 100% reduce 70%
16/03/25 05:49:25 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:49:28 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:49:31 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:49:31 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:49:34 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:49:35 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:49:37 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:49:40 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:49:43 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:49:43 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:49:46 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:49:49 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:49:50 INFO mapreduce.Job: map 100% reduce 74%
16/03/25 05:49:52 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:49:53 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:49:56 INFO mapreduce.Job: map 100% reduce 75%
16/03/25 05:49:58 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:50:01 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:50:02 INFO mapreduce.Job: map 100% reduce 76%
16/03/25 05:50:04 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:50:07 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:50:08 INFO mapreduce.Job: map 100% reduce 77%
16/03/25 05:50:10 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:50:13 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:50:14 INFO mapreduce.Job: map 100% reduce 78%
16/03/25 05:50:16 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:50:19 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:50:22 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:50:23 INFO mapreduce.Job: map 100% reduce 79%
16/03/25 05:50:25 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:50:28 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:50:29 INFO mapreduce.Job: map 100% reduce 80%
16/03/25 05:50:31 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:50:34 INFO mapred.LocalJobRunner: reduce > reduce
16/03/25 05:50:35 INFO mapreduce.Job: map 100% reduce 81%
16/03/25 05:50:37 INFO mapred.LocalJobRunner: reduce > reduce
```

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

```
huser@lp-172-31-27-143: /usr/local/hadoop$ hadoop jar ./mapreduce-job.jar MapReduceJob
16/03/25 05:52:45 INFO mapreduce.Job: Counters: 35
File System Counters
  FILE: Number of bytes read=10434817948
  FILE: Number of bytes written=920547707212
  FILE: Number of read operations=0
  FILE: Number of large read operations=0
  FILE: Number of write operations=0
  HDFS: Number of bytes read=392460167888
  HDFS: Number of bytes written=1000000000000
  HDFS: Number of read operations=1
  HDFS: Number of large read operations=0
  HDFS: Number of write operations=78
Map-Reduce Framework
  Map input records=100000000
  Map output records=100000000
  Map output materialized bytes=10200000458
  Input split bytes=11325
  Combine Input records=0
  Combine output records=0
  Reduce input records=100000000
  Reduce shuffle bytes=10200000458
  Reduce input records=100000000
  Reduce output records=100000000
  Spilled Records=395957882
  Shuffled Maps=0
  Failed Maps=0
  Merged Map outputs=75
  GC time elapsed (ms)=13261
  Total committed heap usage (bytes)=39611531264
Shuffle Errors
  File not found=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAP=0
  WRONG_REDUCE=0
File Input Format Counters
  Bytes Read=10000303104
  File Output Format Counters
  Bytes Written=100000000000
huser@lp-172-31-27-143: /usr/local/hadoop$ cd ./hadoop-mapreduce-project/hadoop-mapreduce
huser@lp-172-31-27-143: /usr/local/hadoop$ ./bin/hdfs dfs -get output/part-r-00000
bash: ./bin/hdfs: No such file or directory
huser@lp-172-31-27-143: /usr/local/hadoop$ ./bin/hdfs dfs -get output/part-r-00000
huser@lp-172-31-27-143: /usr/local/hadoop$ ./bin/hdfs dfs -get output/part-r-00000
16/03/25 05:55:09 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
huser@lp-172-31-27-143: /usr/local/hadoop$ ls
32 bin grep hbase itunes-1.5.tar.gz hadoop-2.7.2 Hadoop-Sort.class Hadoop_SortSortReducer.class hs.jar lib LICENSE.txt NOTICE.txt README.txt share
hadoop-2.7.2.tar.gz hadoop-Sort.java Hadoop_SortTokenMapper.class include libexec logs part-r-00000 sbin
huser@lp-172-31-27-143: /usr/local/hadoop$ ./bin/valsort part-r-00000
Records: 100000000
Checksum: 2faed1ff18402d0
Duplicate keys: 0
SUCCESS - all records are in order
huser@lp-172-31-27-143: /usr/local/hadoop$
```

EC2 Management Console - Mozilla Firefox

Inbox (4,967) ... | inbox (661) - vbi ... | AWS Support D... | Illinois Institute ... | Home Page - 28 ... | https://m...dex.html | Facebook | EC2 Manage... | ConnectionRef... | Improving Map... | En | 10:19 AM | 12:19 AM

https://console.aws.amazon.com/ec2/v2/home?region=us-east-1#instances:sort=instanceState

Search

Vishwanath N. Virginia Support

AWS Services Edit

Launch Instance Connect Actions

Filter by tags and attributes or search by keyword

1 to 17 of 17

Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS	Public IP	Key Name	Monitoring
Slave1	i-4d576679	c3.large	us-east-1a	running	2/2 checks passed	None	ec2-52-164-130-18.com...	54.164.130.18	IronMan_Spark	disabled
Slave2	i-565766df	c3.large	us-east-1a	running	2/2 checks passed	None	ec2-52-87-246-148.com...	52.87.246.148	IronMan_Spark	disabled
<input checked="" type="checkbox"/> Master	i-01c7f382	c3.large	us-east-1d	running	2/2 checks passed	None	ec2-52-87-241-215.com...	52.87.241.215	IronMan_Spark	disabled
Slave3	i-15776665	c3.large	us-east-1a	running	2/2 checks passed	None	ec2-52-87-171-75.com...	54.164.138.108	IronMan_Spark	disabled
Slave4	i-f57667b	c3.large	us-east-1a	running	2/2 checks passed	None	ec2-52-85-171-75.com...	54.85.171.75	IronMan_Spark	disabled
Slave5	i-6546572	c3.large	us-east-1a	running	2/2 checks passed	None	ec2-52-85-195-210.com...	54.85.195.210	IronMan_Spark	disabled
Slave6	i-05776654	c3.large	us-east-1a	running	2/2 checks passed	None	ec2-52-91-58-91.compu...	52.91.58.91	IronMan_Spark	disabled
Slave7	i-657667a	c3.large	us-east-1a	running	2/2 checks passed	None	ec2-52-85-197-41.com...	54.85.197.41	IronMan_Spark	disabled
Slave8	i-45776678	c3.large	us-east-1a	running	2/2 checks passed	None	ec2-52-85-29-142.com...	54.85.29.142	IronMan_Spark	disabled
Slave9	i-95776612	c3.large	us-east-1a	running	2/2 checks passed	None	ec2-52-89-136-172.com...	54.89.136.172	IronMan_Spark	disabled
Slave10	i-55776668	c3.large	us-east-1a	running	2/2 checks passed	None	ec2-52-164-100-3.com...	54.164.100.3	IronMan_Spark	disabled
Slave11	i-55776689	c3.large	us-east-1a	running	2/2 checks passed	None	ec2-52-90-19-169.com...	52.90.19.169	IronMan_Spark	disabled
Slave12	i-1577667f	c3.large	us-east-1a	running	2/2 checks passed	None	ec2-52-152-106-167.co...	54.152.106.167	IronMan_Spark	disabled
Slave13	i-557766db	c3.large	us-east-1a	running	2/2 checks passed	None	ec2-52-91-91-86.compu...	52.91.91.86	IronMan_Spark	disabled
Slave14	i-567766da	c3.large	us-east-1a	running	2/2 checks passed	None	ec2-52-84-217-165.com...	54.84.217.165	IronMan_Spark	disabled
Slave15	i-557766de	c3.large	us-east-1a	running	2/2 checks passed	None	ec2-52-207-228-162.co...	52.207.228.162	IronMan_Spark	disabled
Slave16	i-5546571	c3.large	us-east-1a	running	2/2 checks passed	None	ec2-52-172-38-191.com...	54.172.38.191	IronMan_Spark	disabled

Instance: i-01c7f382 (Master) Public DNS: ec2-52-87-241-215.compute-1.amazonaws.com

Description Status Checks Monitoring Tags

Instance ID: i-01c7f382	Public DNS: ec2-52-87-241-215.compute-1.amazonaws.com
Instance state: running	Public IP: 52.87.241.215
Instance type: c3.large	Elastic IP: -
Private DNS: ip-172-31-49-141.ec2.internal	Availability zone: us-east-1d
Private IPs: 172.31.49.141	Security groups: launch-wizard-8, view rules
Secondary private IPs:	Scheduled events: No scheduled events

Feedback English

2008 - 2016, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

```
hduser@ip-172-31-49-141: /usr/local/hadoop
16/03/28 04:52:47 INFO Client.RMProxy: Connecting to ResourceManager at ec2-52-87-241-215.compute-1.amazonaws.com/172.31.49.141:8082
16/03/28 04:52:48 WARN mapreduce: JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with ToolRunner to remedy this.
16/03/28 04:52:48 INFO input.FileInputFormat: Total input paths to process : 1
16/03/28 04:52:50 INFO mapreduce.JobSubmitter: number of splits:745
16/03/28 04:52:50 INFO mapreduce.Job: 2014-03-28 04:52:50 INFO mapreduce.Job: Job running in uber mode - finalStatus from HDFS ApplicationMaster for job: application_1459128097292_0001
16/03/28 04:52:50 INFO mapreduce.Job: The url to track the job: http://ec2-52-87-241-215.compute-1.amazonaws.com:8088/proxy/application_1459128097292_0001/
16/03/28 04:52:50 INFO mapreduce.Job: Running job: job_1459128097292_0001
16/03/28 04:52:59 INFO mapreduce.Job: Job job_1459128097292_0001 running in uber mode : false
16/03/28 04:52:59 INFO mapreduce.Job: map 0% reduce 0%
16/03/28 04:53:48 INFO mapreduce.Job: map 1% reduce 0%
16/03/28 04:54:01 INFO mapreduce.Job: map 2% reduce 0%
16/03/28 04:54:44 INFO mapreduce.Job: map 3% reduce 0%
16/03/28 04:56:44 INFO mapreduce.Job: map 4% reduce 0%
16/03/28 04:57:53 INFO mapreduce.Job: map 4% reduce 0%
16/03/28 04:59:09 INFO mapreduce.Job: map 5% reduce 0%
16/03/28 05:00:26 INFO mapreduce.Job: map 6% reduce 0%
16/03/28 05:02:01 INFO mapreduce.Job: map 7% reduce 0%
16/03/28 05:03:44 INFO mapreduce.Job: map 8% reduce 0%
16/03/28 05:04:24 INFO mapreduce.Job: map 9% reduce 0%
16/03/28 05:05:41 INFO mapreduce.Job: map 10% reduce 0%
16/03/28 05:07:19 INFO mapreduce.Job: map 11% reduce 0%
16/03/28 05:08:25 INFO mapreduce.Job: map 12% reduce 0%
16/03/28 05:09:41 INFO mapreduce.Job: map 13% reduce 0%
16/03/28 05:10:58 INFO mapreduce.Job: map 14% reduce 0%
16/03/28 05:12:32 INFO mapreduce.Job: map 15% reduce 0%
16/03/28 05:13:37 INFO mapreduce.Job: map 16% reduce 0%
16/03/28 05:14:55 INFO mapreduce.Job: map 17% reduce 0%
16/03/28 05:16:07 INFO mapreduce.Job: map 18% reduce 0%
16/03/28 05:17:24 INFO mapreduce.Job: map 18% reduce 1%
16/03/28 05:17:52 INFO mapreduce.Job: map 19% reduce 1%
16/03/28 05:19:40 INFO mapreduce.Job: map 19% reduce 2%
16/03/28 05:20:06 INFO mapreduce.Job: map 20% reduce 2%
16/03/28 05:22:01 INFO mapreduce.Job: map 20% reduce 3%
16/03/28 05:23:39 INFO mapreduce.Job: map 21% reduce 3%
16/03/28 05:24:42 INFO mapreduce.Job: map 22% reduce 3%
16/03/28 05:25:42 INFO mapreduce.Job: map 22% reduce 4%
16/03/28 05:26:54 INFO mapreduce.Job: map 22% reduce 5%
16/03/28 05:28:38 INFO mapreduce.Job: map 23% reduce 5%
16/03/28 05:28:49 INFO mapreduce.Job: map 23% reduce 6%
16/03/28 05:30:51 INFO mapreduce.Job: map 24% reduce 6%
16/03/28 05:31:41 INFO mapreduce.Job: map 24% reduce 7%
16/03/28 05:33:41 INFO mapreduce.Job: map 25% reduce 7%
16/03/28 05:34:26 INFO mapreduce.Job: map 25% reduce 8%
16/03/28 05:35:50 INFO mapreduce.Job: map 26% reduce 8%
16/03/28 05:37:07 INFO mapreduce.Job: map 26% reduce 9%
16/03/28 05:39:24 INFO mapreduce.Job: map 27% reduce 9%
16/03/28 05:40:36 INFO mapreduce.Job: map 28% reduce 9%
16/03/28 05:41:55 INFO mapreduce.Job: map 28% reduce 9%
16/03/28 05:45:41 INFO mapreduce.Job: map 29% reduce 10%
16/03/28 05:46:50 INFO mapreduce.Job: map 30% reduce 10%
16/03/28 05:49:52 INFO mapreduce.Job: map 31% reduce 10%
16/03/28 05:51:23 INFO mapreduce.Job: map 32% reduce 10%
16/03/28 05:51:43 INFO mapreduce.Job: map 32% reduce 11%
16/03/28 05:53:54 INFO mapreduce.Job: map 33% reduce 11%
16/03/28 05:55:46 INFO mapreduce.Job: map 34% reduce 11%
```

```
hduser@ip-172-31-49-141: /usr/local/hadoop
16/03/28 08:11:54 INFO mapreduce.Job: map 92% reduce 31%
16/03/28 08:13:48 INFO mapreduce.Job: map 93% reduce 31%
16/03/28 08:15:55 INFO mapreduce.Job: map 94% reduce 31%
16/03/28 08:17:42 INFO mapreduce.Job: map 95% reduce 31%
16/03/28 08:19:34 INFO mapreduce.Job: map 95% reduce 32%
16/03/28 08:20:54 INFO mapreduce.Job: map 96% reduce 32%
16/03/28 08:24:54 INFO mapreduce.Job: map 97% reduce 32%
16/03/28 08:26:00 INFO mapreduce.Job: map 98% reduce 32%
16/03/28 08:26:39 INFO mapreduce.Job: map 98% reduce 33%
16/03/28 08:28:44 INFO mapreduce.Job: map 99% reduce 33%
16/03/28 08:29:54 INFO mapreduce.Job: map 99% reduce 33%
16/03/28 08:37:54 INFO mapreduce.Job: map 100% reduce 34%
16/03/28 08:38:06 INFO mapreduce.Job: map 100% reduce 35%
16/03/28 08:38:15 INFO mapreduce.Job: map 100% reduce 36%
16/03/28 08:38:24 INFO mapreduce.Job: map 100% reduce 37%
16/03/28 08:38:33 INFO mapreduce.Job: map 100% reduce 38%
16/03/28 08:38:45 INFO mapreduce.Job: map 100% reduce 39%
16/03/28 08:38:54 INFO mapreduce.Job: map 100% reduce 40%
16/03/28 08:39:06 INFO mapreduce.Job: map 100% reduce 41%
16/03/28 08:39:15 INFO mapreduce.Job: map 100% reduce 42%
16/03/28 08:39:25 INFO mapreduce.Job: map 100% reduce 43%
16/03/28 08:39:34 INFO mapreduce.Job: map 100% reduce 44%
16/03/28 08:39:46 INFO mapreduce.Job: map 100% reduce 45%
16/03/28 08:39:55 INFO mapreduce.Job: map 100% reduce 46%
16/03/28 08:40:07 INFO mapreduce.Job: map 100% reduce 47%
16/03/28 08:40:16 INFO mapreduce.Job: map 100% reduce 48%
16/03/28 08:40:25 INFO mapreduce.Job: map 100% reduce 49%
16/03/28 08:40:34 INFO mapreduce.Job: map 100% reduce 50%
16/03/28 08:40:46 INFO mapreduce.Job: map 100% reduce 51%
16/03/28 08:40:55 INFO mapreduce.Job: map 100% reduce 52%
16/03/28 08:41:07 INFO mapreduce.Job: map 100% reduce 53%
16/03/28 08:41:16 INFO mapreduce.Job: map 100% reduce 54%
16/03/28 08:41:25 INFO mapreduce.Job: map 100% reduce 55%
16/03/28 08:41:34 INFO mapreduce.Job: map 100% reduce 56%
16/03/28 08:41:46 INFO mapreduce.Job: map 100% reduce 57%
16/03/28 08:41:55 INFO mapreduce.Job: map 100% reduce 58%
16/03/28 08:42:07 INFO mapreduce.Job: map 100% reduce 59%
16/03/28 08:42:16 INFO mapreduce.Job: map 100% reduce 60%
16/03/28 08:42:28 INFO mapreduce.Job: map 100% reduce 61%
16/03/28 08:42:40 INFO mapreduce.Job: map 100% reduce 62%
16/03/28 08:42:46 INFO mapreduce.Job: map 100% reduce 63%
16/03/28 08:42:58 INFO mapreduce.Job: map 100% reduce 64%
16/03/28 08:43:07 INFO mapreduce.Job: map 100% reduce 65%
16/03/28 08:43:16 INFO mapreduce.Job: map 100% reduce 66%
16/03/28 08:43:28 INFO mapreduce.Job: map 100% reduce 67%
16/03/28 08:43:40 INFO mapreduce.Job: map 100% reduce 68%
16/03/28 08:45:08 INFO mapreduce.Job: map 100% reduce 69%
16/03/28 08:45:59 INFO mapreduce.Job: map 100% reduce 70%
16/03/28 08:46:54 INFO mapreduce.Job: map 100% reduce 71%
16/03/28 08:47:45 INFO mapreduce.Job: map 100% reduce 72%
16/03/28 08:48:40 INFO mapreduce.Job: map 100% reduce 73%
16/03/28 08:49:25 INFO mapreduce.Job: map 100% reduce 74%
16/03/28 08:50:25 INFO mapreduce.Job: map 100% reduce 75%
16/03/28 08:51:16 INFO mapreduce.Job: map 100% reduce 76%
16/03/28 08:52:11 INFO mapreduce.Job: map 100% reduce 77%
```

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

EC2 Management Console - Mozilla Firefox

java code to read in... EC2 For Complete ... GC: Collections - jav... EC2 Management C... Monitoring spark jo... Spark shell -Details for ... Inbox (692) -vbidari... Search

Vishwanath N. Virginia Support

AWS Services Edit

Launch Instance Connect Actions

Filter by tags and attributes or search by keyword

1 to 17 of 17 >>

Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS	Public IP	Key Name	Monitoring
Slave1	i-47e46c3	c3.large	us-east-1a	running	2/2 checks passed	None	ec2-54-152-53-64.comp...	54.152.53.64	IronMan_Spark	disabled
Slave2	i-90e46f14	c3.large	us-east-1a	running	2/2 checks passed	None	ec2-52-201-237-169.co...	52.201.237.169	IronMan_Spark	disabled
Slave3	i-24e7e5a0	c3.large	us-east-1a	running	2/2 checks passed	None	ec2-52-87-246-199.com...	52.87.246.199	IronMan_Spark	disabled
Slave4	i-25e7e5a1	c3.large	us-east-1a	running	2/2 checks passed	None	ec2-52-87-241-215.com...	52.87.241.215	IronMan_Spark	disabled
Master	i-01c7f382	c3.large	us-east-1d	running	2/2 checks passed	None	ec2-52-87-241-215.com...	52.87.241.215	IronMan_Spark	disabled
<input checked="" type="checkbox"/> Slave5	i-4be46cf	c3.large	us-east-1a	running	2/2 checks passed	None	ec2-52-90-128-62.compute...	52.90.128.62	IronMan_Spark	disabled
Slave6	i-48e46cc	c3.large	us-east-1a	running	2/2 checks passed	None	ec2-52-201-238-45.com...	52.201.238.45	IronMan_Spark	disabled
Slave7	i-4ae46ce	c3.large	us-east-1a	running	2/2 checks passed	None	ec2-52-207-225-234.co...	52.207.225.234	IronMan_Spark	disabled
Slave15	i-55e46d1	c3.large	us-east-1a	running	2/2 checks passed	None	ec2-54-175-231-31.com...	54.175.231.31	IronMan_Spark	disabled
Slave8	i-97e7e53d	c3.large	us-east-1a	running	2/2 checks passed	None	ec2-54-175-150-54.com...	54.173.150.54	IronMan_Spark	disabled
Slave9	i-22e468a0	c3.large	us-east-1a	running	2/2 checks passed	None	ec2-52-91-43-153.comp...	52.91.43.153	IronMan_Spark	disabled
Slave10	i-44eae8c0	c3.large	us-east-1a	running	2/2 checks passed	None	ec2-54-173-203-62.com...	54.173.203.62	IronMan_Spark	disabled
Slave16	i-20e8e84	c3.large	us-east-1a	running	2/2 checks passed	None	ec2-54-174-168-4.com...	54.174.168.4	IronMan_Spark	disabled
Slave11	i-43e46fc7	c3.large	us-east-1a	running	2/2 checks passed	None	ec2-54-89-136-239.com...	54.89.136.239	IronMan_Spark	disabled
Slave12	i-bea7e703e	c3.large	us-east-1a	running	2/2 checks passed	None	ec2-54-174-101.com...	54.84.174.101	IronMan_Spark	disabled
Slave13	i-23ee8e87	c3.large	us-east-1a	running	2/2 checks passed	None	ec2-54-174-234-116.co...	54.174.234.116	IronMan_Spark	disabled
Slave14	i-21ea8e85	c3.large	us-east-1a	running	2/2 checks passed	None	ec2-54-173-74-85.comp...	54.173.74.85	IronMan_Spark	disabled

Instance: i-4be46cf (Slave5) Public DNS: ec2-52-90-128-62.compute-1.amazonaws.com

Description Status Checks Monitoring Tags

Instance ID	i-4be46cf	Public DNS	ec2-52-90-128-02.compute-1.amazonaws.com
Instance state	running	Public IP	52.90.128.62
Instance type	c3.large	Elastic IP	-
Private DNS	ip-172-31-14-56.ec2.internal	Availability zone	us-east-1a

Feedback English

© 2008 - 2016, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

```
hduser@ip-172-31-49-141:~/usr/local/hadoop$ hdfs dfs -mkdir -p Input
16/03/31 10:02:04 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hduser@ip-172-31-49-141:~/usr/local/hadoop$ hdfs dfs -ls
16/03/31 10:02:08 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 1 items
drwxr-xr-x  - hduser  supergroup          0 2016-03-31 10:02 Input
hduser@ip-172-31-49-141:~/usr/local/hadoop$ ls
igbdata.txt.gz  gpl-2.0.txt  hadoop-2.7.2.tar.gz  Hadoop_Sort.java  Hadoop_SortTokenizerMapper.class  include  libexec  logs  output  sbin
hduser@ip-172-31-49-141:~/usr/local/hadoop$ bzip2 -d igbdata.txt.gz
hduser@ip-172-31-49-141:~/usr/local/hadoop$ hdfs dfs -put igbdata.txt Input
16/03/31 10:03:11 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
hduser@ip-172-31-49-141:~/usr/local/hadoop$ hdfs dfs -ls
16/03/31 10:03:46 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Found 1 items
-rw-r--r--  1 hduser  supergroup 10000000000 2016-03-31 10:03 Input/igbdata.txt
hduser@ip-172-31-49-141:~/usr/local/hadoop$ ./bin/hadoop jar hs.jar Hadoop_Sort Input/igbdata.txt Output
16/03/31 10:05:39 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
16/03/31 10:05:40 INFO Configuration.deprecation: Configuration item 'com.mapreduce.jobtracker.address' is deprecated. It has been replaced by 'mapreduce.jobtracker.address'. Please use the new name. This warning will be suppressed after 2014-08-01.
16/03/31 10:05:40 INFO mapred.JobClient: JobTracker at e2c-52-87-241-215.compute-1.amazonaws.com/172.31.49.141:8082
16/03/31 10:05:41 INFO mapred.JobClient: Hadoop version: 2.7.2
16/03/31 10:05:41 INFO InputFormat: Total input paths to process : 1
16/03/31 10:05:42 INFO mapreduce.JobSubmissionsumer: number of splits:8
16/03/31 10:05:42 INFO mapreduce.JobSubmissionsumer: Submitting tokens for job: job_1459418473485_0001
16/03/31 10:05:42 INFO mapred.YarnClientImpl: Submitted application application_1459418473485_0001
16/03/31 10:05:42 INFO mapred.YarnClientImpl: Application report for application_1459418473485_0001 from node-id: 1459418473485-0001
16/03/31 10:05:42 INFO mapred.YarnClientImpl: Application report for application_1459418473485_0001 from node-id: 1459418473485-0001
16/03/31 10:05:51 INFO mapreduce.Job: Job job_1459418473485_0001 running in uber mode : False
16/03/31 10:05:51 INFO mapreduce.Job: map 0% reduce 0%
16/03/31 10:06:18 INFO mapreduce.Job: map 2% reduce 0%
16/03/31 10:06:19 INFO mapreduce.Job: map 13% reduce 0%
16/03/31 10:06:20 INFO mapreduce.Job: map 28% reduce 0%
16/03/31 10:06:22 INFO mapreduce.Job: map 33% reduce 0%
16/03/31 10:06:23 INFO mapreduce.Job: map 38% reduce 0%
16/03/31 10:06:24 INFO mapreduce.Job: map 29% reduce 0%
16/03/31 10:06:25 INFO mapreduce.Job: map 36% reduce 0%
16/03/31 10:06:26 INFO mapreduce.Job: map 34% reduce 0%
16/03/31 10:06:27 INFO mapreduce.Job: map 32% reduce 0%
16/03/31 10:06:38 INFO mapreduce.Job: map 45% reduce 0%
16/03/31 10:06:39 INFO mapreduce.Job: map 47% reduce 0%
16/03/31 10:06:41 INFO mapreduce.Job: map 50% reduce 0%
16/03/31 10:06:47 INFO mapreduce.Job: map 51% reduce 0%
16/03/31 10:06:48 INFO mapreduce.Job: map 53% reduce 0%
16/03/31 10:06:49 INFO mapreduce.Job: map 54% reduce 0%
16/03/31 10:06:54 INFO mapreduce.Job: map 71% reduce 0%
16/03/31 10:06:55 INFO mapreduce.Job: map 72% reduce 0%
16/03/31 10:06:57 INFO mapreduce.Job: map 75% reduce 0%
16/03/31 10:07:16 INFO mapreduce.Job: map 89% reduce 13%
16/03/31 10:07:18 INFO mapreduce.Job: map 93% reduce 13%
16/03/31 10:07:20 INFO mapreduce.Job: map 97% reduce 7%
16/03/31 10:07:22 INFO mapreduce.Job: map 98% reduce 21%
16/03/31 10:07:23 INFO mapreduce.Job: map 100% reduce 21%
16/03/31 10:07:28 INFO mapreduce.Job: map 100% reduce 67%
16/03/31 10:07:28 INFO mapreduce.Job: map 100% reduce 69%
16/03/31 10:07:31 INFO mapreduce.Job: map 100% reduce 74%
16/03/31 10:07:34 INFO mapreduce.Job: map 100% reduce 78%
```

```
hadoop@ip-172-31-49-141:~/usr/local/hadoop$ ./bin/hadoop jar hs.jar Hadoop_Sort Input/igbdata.txt Output/
FILE: Number of bytes read=1978314714
FILE: Number of bytes written=2999373967
FILE: Number of read operations=0
FILE: Number of large read operations=0
HDFS: Number of bytes read=1000629864
HDFS: Number of bytes written=1000000000
HDFS: Number of read operations=27
HDFS: Number of large read operations=0
HDFS: Number of write operations=2
Job Counters
  Killed map tasks=1
  Launched map tasks=9
  Launched reduce tasks=1
  Data-local map tasks=1
  Total time spent by all maps in occupied slots (ms)=441040
  Total time spent by all reduces in occupied slots (ms)=45173
  Total time spent by all map tasks (ms)=441040
  Total time spent by all reduce tasks (ms)=45173
  Total vcore-milliseconds taken by all map tasks=441040
  Total vcore-milliseconds taken by all reduce tasks=45173
  Total megabyte-milliseconds taken by all map tasks=451624960
  Total megabyte-milliseconds taken by all reduce tasks=46257152
Map-Reduce Framework
  Map input records=10000000
  Map output records=10000000
  Map output bytes=1000000000
  Map output materialized bytes=1020000048
  Input split bytes=1192
  Combine input records=0
  Combine output records=0
  Reduce input groups=10000000
  Reducer input bytes=1020000048
  Reduce input records=10000000
  Reduce output records=10000000
  Spilled Records=29395241
  Shuffled Maps =0
  Failed Shuffles=0
  Merged Map outputs=0
  GC time elapsed (ms)=5530
  CPU time spent (ms)=119160
  Physical memory (bytes) snapshot=2240122880
  Virtual memory (bytes) snapshot=7458549760
  Total committed heap usage (bytes)=1658322944
Shuffle File IO Counters
  BAD_ID=0
  CONNECTION=0
  IO_ERROR=0
  WRONG_LENGTH=0
  WRONG_MAGIC=0
  WRONG_DUPLICATE=0
File Input Format Counters
  Bytes Read=1000028672
File Output Format Counters
  Bytes Written=1000000000
hadoop@ip-172-31-49-141:~/usr/local/hadoop$
```

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

### 3. Spark

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

```
hduser@ip-172-31-49-141:/usr/local/spark
[16/03/28 22:18:32 INFO SparkKafkaMapReduceUtil: attempt_201603282217_0002_n_000003_19: Committed
16/03/28 22:18:32 INFO Executor: Finished task 4.0 in stage 2.0 (TID 19). 2080 bytes result sent to driver
16/03/28 22:18:32 INFO TaskSetManager: Starting task 4.0 in stage 2.0 (TID 26, localhost, partition 4, NODE_LOCAL, 1961 bytes)
16/03/28 22:18:32 INFO TaskSetManager: Finished task 4.0 in stage 2.0 (TID 19) in 3894 ms on localhost (4/8)
16/03/28 22:18:32 INFO Executor: Running task 4.0 in stage 2.0 (TID 20)
16/03/28 22:18:32 INFO Executor: Getting 8 non-empty blocks out of 8 blocks
16/03/28 22:18:32 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
16/03/28 22:18:38 INFO FileOutputCommitter: Saved output of task 'attempt_201603282217_0002_m_000004_20' to hdfs://ec2-52-87-241-215.compute-1.amazonaws.com:54310/outputspark/_temporary/0/task_201603282217_0002_m_000004
16/03/28 22:18:38 INFO SparkKafkaMapReduceUtil: attempt_201603282217_0002_n_000004_20: Committed
16/03/28 22:18:38 INFO Executor: Finished task 4.0 in stage 2.0 (TID 20). 2080 bytes result sent to driver
16/03/28 22:18:38 INFO TaskSetManager: Starting task 4.0 in stage 2.0 (TID 21, localhost, partition 5, NODE_LOCAL, 1961 bytes)
16/03/28 22:18:38 INFO TaskSetManager: Finished task 4.0 in stage 2.0 (TID 20) in 6947 ms on localhost (5/8)
16/03/28 22:18:38 INFO Executor: Running task 5.0 in stage 2.0 (TID 21)
16/03/28 22:18:38 INFO Executor: Getting 8 non-empty blocks out of 8 blocks
16/03/28 22:18:38 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ns
16/03/28 22:18:38 INFO FileOutputCommitter: Saved output of task 'attempt_201603282217_0002_n_000005_21' to hdfs://ec2-52-87-241-215.compute-1.amazonaws.com:54310/outputspark/_temporary/0/task_201603282217_0002_n_000005
16/03/28 22:18:44 INFO SparkKafkaMapReduceUtil: attempt_201603282217_0002_n_000005_21: Committed
16/03/28 22:18:44 INFO Executor: Finished task 5.0 in stage 2.0 (TID 21). 2080 bytes result sent to driver
16/03/28 22:18:44 INFO TaskSetManager: Starting task 5.0 in stage 2.0 (TID 22, localhost, partition 6, NODE_LOCAL, 1961 bytes)
16/03/28 22:18:44 INFO TaskSetManager: Finished task 5.0 in stage 2.0 (TID 21) in 5579 ms on localhost (6/8)
16/03/28 22:18:44 INFO Executor: Running task 6.0 in stage 2.0 (TID 22)
16/03/28 22:18:44 INFO Executor: Getting 8 non-empty blocks out of 8 blocks
16/03/28 22:18:44 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ns
16/03/28 22:18:44 INFO FileOutputCommitter: Saved output of task 'attempt_201603282217_0002_m_000006_22' to hdfs://ec2-52-87-241-215.compute-1.amazonaws.com:54310/outputspark/_temporary/0/task_201603282217_0002_m_000006
16/03/28 22:18:50 INFO SparkKafkaMapReduceUtil: attempt_201603282217_0002_n_000006_22: Committed
16/03/28 22:18:50 INFO Executor: Finished task 6.0 in stage 2.0 (TID 22). 2080 bytes result sent to driver
16/03/28 22:18:50 INFO TaskSetManager: Starting task 6.0 in stage 2.0 (TID 23, localhost, partition 7, NODE_LOCAL, 1961 bytes)
16/03/28 22:18:50 INFO TaskSetManager: Finished task 6.0 in stage 2.0 (TID 22) in 6277 ms on localhost (7/8)
16/03/28 22:18:50 INFO Executor: Running task 7.0 in stage 2.0 (TID 23)
16/03/28 22:18:50 INFO ShuffleBlockFetcherIterator: Getting 8 non-empty blocks out of 8 blocks
16/03/28 22:18:50 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ns
16/03/28 22:18:50 INFO FileOutputCommitter: Saved output of task 'attempt_201603282217_0002_n_000007_23' to hdfs://ec2-52-87-241-215.compute-1.amazonaws.com:54310/outputspark/_temporary/0/task_201603282217_0002_n_000007
16/03/28 22:18:55 INFO SparkKafkaMapReduceUtil: attempt_201603282217_0002_n_000007_23: Committed
16/03/28 22:18:55 INFO Executor: Finished task 7.0 in stage 2.0 (TID 23). 2080 bytes result sent to driver
16/03/28 22:18:55 INFO TaskSetManager: Finished task 7.0 in stage 2.0 (TID 23) in 4739 ms on localhost (8/8)
16/03/28 22:18:55 INFO DAGScheduler: ResultStage 2 (saveAsTextFile at Tera_Sort.java:46) finished in 45.010 s
16/03/28 22:18:55 INFO DAGScheduler: Removed tasks because tasks have all completed, from pool
16/03/28 22:18:55 INFO DAGScheduler: Job 1 finished: saveAsTextFile at Tera_Sort.java:46, took 67.903165 s
16/03/28 22:18:55 INFO SparkUI: Stopped Spark web UI at http://172.31.49.141:4040
16/03/28 22:18:55 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
16/03/28 22:18:55 INFO MemoryStore: MemoryStore cleared
16/03/28 22:18:55 INFO BlockManager: BlockManager stopped
16/03/28 22:18:55 INFO BlockManager: BlockManagerMaster stopped
16/03/28 22:18:55 INFO OutputCommitCoordinator: OutputCommitCoordinator stopped!
16/03/28 22:18:55 INFO SparkContext: Successfully stopped SparkContext
16/03/28 22:18:55 INFO RemoteActorRefProvider$RemotingTerminator: Shutting down remote daemon.
16/03/28 22:18:55 INFO ShutdownHookManager: Shutdown hook called
16/03/28 22:18:55 INFO RemoteActorRefProvider$RemotingTerminator: Remote daemon shut down; proceeding with flushing remote transports.
16/03/28 22:18:55 INFO ShutdownHookManager: Deleting directory /tmp/spark-88498c91-3365-435e-a171-f4b218d74f32
16/03/28 22:18:55 INFO SparkUI: Stopped Spark web UI at http://172.31.49.141:4040
hduser@ip-172-31-49-141:/usr/local/sparks hdfs dfs -get outputspark
16/03/28 22:20:30 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
```

```
hduser@ip-172-31-49-141:/usr/local/spark/ec2
[root@ip-172-31-13-86 ephemeral-hdfs]$ ..spark/bin/spark-submit --class tera.Sort ..spark/bidarl-0.0.1-SNAPSHOT.jar spark input/gbdata.txt output
16/03/29 00:37:48 INFO spark.SparkContext: Running Spark version 1.6.1
16/03/29 00:37:48 WARN spark.SparkConf: SPARK_WORKER_INSTANCES was detected (set to '1').
This is deprecated in Spark 1.6+.

Please instead use:
- ./spark-submit with --num-executors to specify the number of executors
- Or see SPARK_EXECUTOR_INSTANCES
- spark.executor.instances to configure the number of instances in the spark config.

16/03/29 00:37:48 INFO spark.SecurityManager: Changing view acls to: root
16/03/29 00:37:48 INFO spark.SecurityManager: Changing modify acls to: root
16/03/29 00:37:48 INFO spark.SecurityManager: SecurityManager: authentication disabled; ul acls disabled; users wth view permissions: Set(root); users wth modify permissions: Set(root)
16/03/29 00:37:49 INFO util.Utils: Successfully started service 'sparkDriver' on port 44584.
16/03/29 00:37:49 INFO spark.SparkEnv: Environment: Staging complete.
16/03/29 00:37:49 INFO Remoting: Remoting started: listening on addresses :[akka.tcp://sparkDriverActorSystem@172.31.13.86:4080]
16/03/29 00:37:49 INFO util.Utils: Successfully started service 'sparkDriverActorSystem' on port 40805.
16/03/29 00:37:49 INFO spark.SparkEnv: Registering MapOutputTracker
16/03/29 00:37:49 INFO spark.SparkEnv: Registering BlockManagerMaster
16/03/29 00:37:49 INFO storage.BlockManagerMaster: Created local directory at /mnt/spark/blockmgr-fd38d24c-e938-48fd-a7bf-26bb0574d975
16/03/29 00:37:49 INFO storage.DiskBlockManager: Created local directory at /mnt2/spark/blockmgr-72f22ac-146b-4f0a-9c19-5126ab5938265
16/03/29 00:37:50 INFO spark.MemoryStore: MemoryStore started with capacity 511.5 MB
16/03/29 00:37:50 INFO spark.SparkEnv: Registering OutputCommitCoordinator
16/03/29 00:37:50 INFO server.Server: jetty-8.y.z-SNAPSHOT
16/03/29 00:37:50 INFO storage.AbstractConnector: Started SelectChannelConnector@0.0.0:4040
16/03/29 00:37:50 INFO spark.HttpFileServer: Starting HttpFileServer at http://172.31.13.86:4040
16/03/29 00:37:50 INFO spark.HttpFileServer: HTTP File server directory is /mnt/spark/675fc85-1264-4e77-9272-031a6b1507bd/httpd-12da1d7c-7f05-41eb-bdf2-fb0ae3d90d4e
16/03/29 00:37:50 INFO spark.HttpFileServer: Starting HTTP Server
16/03/29 00:37:50 INFO server.Server: jetty-8.y.z-SNAPSHOT
16/03/29 00:37:50 INFO server.AbstractConnector: Started SocketConnector@0.0.0:56607
16/03/29 00:37:50 INFO spark.SparkEnv: Added MR FileService at /ephemeral-hdfs//spark/bidarl-0.0.1-SNAPSHOT.jar at http://172.31.13.86:56607/jars/bidarl-0.0.1-SNAPSHOT.jar with timestamp 1459211870420
16/03/29 00:37:50 INFO executor.Executor: Starting executor ID driver on host localhost
16/03/29 00:37:50 INFO util.Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 33890.
16/03/29 00:37:50 INFO storage.BlockManagerMaster: Trying to register BlockManager
16/03/29 00:37:50 INFO storage.BlockManager: Registering block manager localhost:33890 with 511.5 MB RAM, BlockManagerId(driver, localhost, 33890)
16/03/29 00:37:51 INFO storage.BlockManagerMaster: Registered BlockManager
16/03/29 00:37:51 INFO storage.MemoryStore: Block broadcast_0 stored as Values in memory (estimated size 46.3 KB, free 46.3 KB)
16/03/29 00:37:51 INFO storage.MemoryStore: Block broadcast_0_piece0 stored as bytes in memory (estimated size 4.4 KB, free 50.7 KB)
16/03/29 00:37:51 INFO storage.BlockManagerInfo: Added broadcast_0_piece0 in memory on localhost:33890 (size: 4.4 KB, free: 511.5 MB)
16/03/29 00:37:51 INFO spark.SparkContext: Created broadcast 0 from textfile at tera.Sort.java:23
16/03/29 00:37:51 INFO util.MappedFileMapper: Mapped file /tmp/spark-88498c91-3365-435e-a171-f4b218d74f32/textfile to /tmp/spark/675fc85-1264-4e77-9272-031a6b1507bd/textfile for your platform... using builtin-java classes where applicable
16/03/29 00:37:51 WARN storage.LoadSnapshot: Snapshot native library not loaded
16/03/29 00:37:51 INFO spark.SparkContext: Starting job: sortByKey at tera.Sort.java:39
16/03/29 00:37:51 INFO scheduler.DAGScheduler: Got job 0 (sortByKey at tera.Sort.java:39) with 8 output partitions
16/03/29 00:37:51 INFO scheduler.DAGScheduler: Final Stage: ResultStage 0 (sortByKey at tera.Sort.java:39)
16/03/29 00:37:51 INFO scheduler.DAGScheduler: Submitting 1 pending stage (Stage 0)
16/03/29 00:37:51 INFO scheduler.DAGScheduler: Missing parents: List()
16/03/29 00:37:51 INFO scheduler.DAGScheduler: Submitting ResultStage 0 (MapPartitionsRDD[4] at sortByKey at tera.Sort.java:39), which has no missing parents
16/03/29 00:37:51 INFO storage.MemoryStore: Block broadcast_1 stored as values in memory (estimated size 4.1 KB, free 54.8 KB)
16/03/29 00:37:51 INFO storage.MemoryStore: Block broadcast_1_piece0 stored as bytes in memory (estimated size 2.3 KB, free 57.1 KB)
```

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

```
hduser@ip-172-31-49-141: /usr/local/spark/ec2
hduser@ip-172-31-49-141: /usr/local/spark/ec2
root@ip-172-31-13-86: ~$ output|unxzdos file
unxzdos: converting file file to DOS format ...
root@ip-172-31-13-86: ~$ ls
file part-00000 part-00001 part-00002 part-00003 part-00004 part-00005 part-00006 part-00007 _SUCCESS
root@ip-172-31-13-86: ~$ cd /root/ephemeral-hdfs/_SUCCESS
root@ip-172-31-13-86: ~$ ls
hadoop-client-0.4.0.jar hadoop-micrcluster-1.0.4.jar ivy libexec output share
hadoop-core-1.0.4.jar hadoop-test-1.0.4.jar ivy.xml LICENSE.txt README.txt src
hadoop-examples-1.0.4.jar hadoop-tools-1.0.4.jar lib NOTICE.txt sbin webapps
root@ip-172-31-13-86: ~$ bin/CHANGES.txt docs hadoop-ant-1.0.4.jar
root@ip-172-31-13-86: ~$ ./valsort /root/ephemeral-hdfs/output/file
Records: 10000000
Checksum: 4c48a881c797d5
Duplicate keys: 0
SUCCESS - all records are in order
root@ip-172-31-13-86: ~$
```

```
hduser@ip-172-31-49-141: /usr/local/spark/ec2
hduser@ip-172-31-49-141: /usr/local/spark/ec2
vishwanath@vishwanath-Q551LB: ~/Documents
16/03/29 00:39:13 INFO mapred.FileOutputCommitter: Saved output of task 'attempt_201603290038_0002_m_000006_22' to hdfs://ec2-52-207-251-142.compute-1.amazonaws.com:9000/user/root/output
16/03/29 00:39:13 INFO executor.Executor: Flinshed task 6.0 in stage 2.0 (ID 22). 1165 bytes result sent to driver
16/03/29 00:39:13 INFO scheduler.TaskSetManager: Stopped task 6.0 in stage 2.0 (ID 23, localhost, partition 7, NODE_LOCAL, 1960 bytes)
16/03/29 00:39:13 INFO scheduler.TaskSetManager: Unfinished task 6.0 in stage 2.0 (ID 22) in 5328 ms on localhost (7/8)
16/03/29 00:39:13 INFO executor.Executor: Running task 7.0 in stage 2.0 (ID 23)
16/03/29 00:39:13 INFO storage.ShuffleFileLockFetcherIterator: Getting 8 non-empty blocks out of 8 blocks
16/03/29 00:39:13 INFO storage.ShuffleFileLockFetcherIterator: Started 0 remote fetches in 0 ms
16/03/29 00:39:18 INFO mapred.SparkHadoopMapRedUtil: Saved output of task 'attempt_201603290038_0002_m_000007_23' to hdfs://ec2-52-207-251-142.compute-1.amazonaws.com:9000/user/root/output
16/03/29 00:39:18 INFO mapred.SparkHadoopMapRedUtil: attempt_201603290038_0002_m_000007_23 Committed
16/03/29 00:39:18 INFO mapred.FileOutputCommitter: Saved output of task 'attempt_201603290038_0002_m_000007_23' to hdfs://ec2-52-207-251-142.compute-1.amazonaws.com:9000/user/root/output
16/03/29 00:39:18 INFO executor.Executor: Flinshed task 7.0 in stage 2.0 (ID 23) in 5444 ms on localhost (8/8)
16/03/29 00:39:18 INFO scheduler.TaskSetManager: Flinshed task 7.0 at tera.Sort.java:46) finished in 46.286 s
16/03/29 00:39:18 INFO scheduler.DAGScheduler: Job 1 finished saveAsTextFile at tera.Sort.java:46, took 72.492926 s
16/03/29 00:39:18 INFO handler.ContextHandler: stopped d.s.j.s.ServletContextHandler(/metrics/json,null)
16/03/29 00:39:18 INFO handler.ContextHandler: stopped d.s.j.s.ServletContextHandler(/stage/kill,null)
16/03/29 00:39:18 INFO handler.ContextHandler: stopped d.s.j.s.ServletContextHandler(/api,null)
16/03/29 00:39:18 INFO handler.ContextHandler: stopped d.s.j.s.ServletContextHandler(/null)
16/03/29 00:39:18 INFO handler.ContextHandler: stopped d.s.j.s.ServletContextHandler(/executors/threaddump,null)
16/03/29 00:39:18 INFO handler.ContextHandler: stopped d.s.j.s.ServletContextHandler(/executors/threaddump,null)
16/03/29 00:39:18 INFO handler.ContextHandler: stopped d.s.j.s.ServletContextHandler(/executors/threaddump,null)
16/03/29 00:39:18 INFO handler.ContextHandler: stopped d.s.j.s.ServletContextHandler(/executors,null)
16/03/29 00:39:18 INFO handler.ContextHandler: stopped d.s.j.s.ServletContextHandler(/environment/json,null)
16/03/29 00:39:18 INFO handler.ContextHandler: stopped d.s.j.s.ServletContextHandler(/storage/rd0/json,null)
16/03/29 00:39:18 INFO handler.ContextHandler: stopped d.s.j.s.ServletContextHandler(/storage/stage/json,null)
16/03/29 00:39:18 INFO handler.ContextHandler: stopped d.s.j.s.ServletContextHandler(/storage/null)
16/03/29 00:39:18 INFO handler.ContextHandler: stopped d.s.j.s.ServletContextHandler(/stage/pool/json,null)
16/03/29 00:39:18 INFO handler.ContextHandler: stopped d.s.j.s.ServletContextHandler(/stages/null)
16/03/29 00:39:18 INFO handler.ContextHandler: stopped d.s.j.s.ServletContextHandler(/stages/stage/json,null)
16/03/29 00:39:18 INFO handler.ContextHandler: stopped d.s.j.s.ServletContextHandler(/stages/stage/null)
16/03/29 00:39:18 INFO handler.ContextHandler: stopped d.s.j.s.ServletContextHandler(/stages/null)
16/03/29 00:39:18 INFO handler.ContextHandler: stopped d.s.j.s.ServletContextHandler(/jobs/job/json,null)
16/03/29 00:39:18 INFO handler.ContextHandler: stopped d.s.j.s.ServletContextHandler(/jobs/job,null)
16/03/29 00:39:18 INFO handler.ContextHandler: stopped d.s.j.s.ServletContextHandler(/jobs/json,null)
16/03/29 00:39:18 INFO handler.ContextHandler: stopped d.s.j.s.ServletContextHandler(/jobs/null)
16/03/29 00:39:18 INFO util.SparkUI: Stopped Spark web UI at http://ec2-52-207-251-142.compute-1.amazonaws.com:4040
16/03/29 00:39:18 INFO spark.MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
16/03/29 00:39:19 INFO storage.MemoryStore: MemoryStore cleared
16/03/29 00:39:19 INFO storage.BlockManager: BlockManager stopped
16/03/29 00:39:19 INFO storage.BlockManagerMaster: BlockManagerMaster stopped
16/03/29 00:39:19 INFO storage.DiskBlockResolver: DiskBlockResolver stopped
16/03/29 00:39:19 INFO spark.SparkContext: Successfully started SparkContext
16/03/29 00:39:19 INFO remote.RemoteActorRefProvider$RemoteTerminator: Shutting down remote daemon.
16/03/29 00:39:19 INFO remote.RemoteActorRefProvider$RemoteTerminator: Remote daemon shut down; proceeding with flushing remote transports.
16/03/29 00:39:19 INFO util.ShutdownHookManager: Shutdown hook called
16/03/29 00:39:19 INFO util.ShutdownHookManager: Deleting directory /mnt/spark/spark-675f0c85-1264-4e77-9272-031a6b1507bd
16/03/29 00:39:19 INFO util.ShutdownHookManager: Deleting directory /mnt/spark/spark-5eddf92-0ba4-463b-abcf-f6599b8e2c3b7
16/03/29 00:39:19 INFO util.ShutdownHookManager: Deleting directory /mnt/spark/spark-675f0c85-1264-4e77-9272-031a6b1507bd/httpd-12da1d7c-7f05-41eb-bdf2-fb0ae3d90d4e
16/03/29 00:39:19 INFO remote.RemoteActorRefProvider$RemoteTerminator: Renoting shut down.
root@ip-172-31-13-86: ~$ ./bin/hadoop dfs -ls output
root@ip-172-31-13-86: ~$
```

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

```
hduser@ip-172-31-27-143: /usr/local/spark
vishwanath@vishwanath-Q551LB:~/Documents
hduser@ip-172-31-27-143:/usr/local/hadoop$ ./bin/spark-submit --class tera_Sort --master spark://ip-172-31-27-143:7077 bldarl-0.0.1-SNAPSHOT.jar hdfs://ec2-54-174-17-233.compute-1.amazonaws.com:54310/user/hduser@ip-172-31-27-143:/usr/local/hadoop cd ..
hduser@ip-172-31-27-143:/usr/local/hadoop
hduser@ip-172-31-27-143:/usr/local/hadoop$ ./bin/spark-submit --class tera_Sort --master spark://ip-172-31-27-143:7077 bldarl-0.0.1-SNAPSHOT.jar hdfs://ec2-54-174-17-233.compute-1.amazonaws.com:54310/user/hduser@ip-172-31-27-143:/usr/local/hadoop
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
16/03/26 02:50:29 INFO SparkContext: Running Spark version 1.6.1
16/03/26 02:50:30 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-Java classes where applicable
16/03/26 02:50:30 INFO SparkConf: Detected deprecated memory fraction settings: [spark.storage.memoryFraction]. As of Spark 1.6, execution and storage memory management are unified. All memory fractions used by executors now refer to the old memory manager, so you may explicitly enable 'spark.memory.useLegacyMode' (not recommended).
16/03/26 02:50:30 INFO SecurityManager: Changing view acls to: hduser
16/03/26 02:50:30 INFO SecurityManager: Changing modify acls to: hduser
16/03/26 02:50:30 INFO SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(hduser); users with modify permissions: Set(hduser)
16/03/26 02:50:31 INFO Sif4Logger: Sif4Logger started
16/03/26 02:50:31 INFO TaskSetManager: Registered executor 0 on ip-172-31-27-143
16/03/26 02:50:32 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriverActorSystem@172.31.27.143:49590]
16/03/26 02:50:32 INFO Utils: Successfully started service 'sparkDriverActorSystem' on port 49590.
16/03/26 02:50:32 INFO SparkEnv: Registering BlockManagerMaster
16/03/26 02:50:32 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-c35ade90-3b11-481d-adda-9db4578e4b6c
16/03/26 02:50:32 INFO BlockManager: BlockManager started with capacity 1247.6 MB
16/03/26 02:50:32 INFO SparkEnv: Registering OutputCommitCoordinator
16/03/26 02:50:32 INFO Utils: Successfully started service 'SparkUI' on port 4040.
16/03/26 02:50:32 INFO SparkUI: Started SparkUI at http://172.31.27.143:4040
16/03/26 02:50:32 INFO HttpFileServer: HTTP file server directory is /tmp/spark-40093be0-5b05-4a7e-b621-58c1b48b272e/httpd-2d19ddd4-5acb-466a-8b44-fbd7eed0328e
16/03/26 02:50:32 INFO HttpServer: Starting HTTP Server
16/03/26 02:50:32 INFO Utils: Successfully started service 'HTTP file server' on port 52087.
16/03/26 02:50:32 INFO SparkContext: Added JAR file:/usr/local/spark/bldarl-0.0.1-SNAPSHOT.jar at http://172.31.27.143:52087/jars/bldarl-0.0.1-SNAPSHOT.jar with timestamp 1458906632747
```

```
hduser@ip-172-31-27-143: /usr/local/spark
vishwanath@vishwanath-Q551LB:~/Documents
hduser@ip-172-31-27-143:/usr/local/spark
16/03/26 02:50:50 INFO Executor: Finished task 7.0 in stage 0.0 (TID 7). 3068 bytes result sent to driver
16/03/26 02:50:50 INFO Executor: Starting task 8.0 in stage 0.0 (TID 8)
16/03/26 02:50:50 INFO TaskSetManager: Finishing task 7.0 in stage 0.0 (TID 7) in 2350 ms on localhost (8/75)
16/03/26 02:50:53 INFO Executor: Finished task 8.0 in stage 0.0 (TID 8). 3068 bytes result sent to driver
16/03/26 02:50:53 INFO Executor: Starting task 9.0 in stage 0.0 (TID 9)
16/03/26 02:50:53 INFO TaskSetManager: Finished task 8.0 in stage 0.0 (TID 8) in 2231 ms on localhost (9/75)
16/03/26 02:50:53 INFO HadoopRDD: Input split: hdfs://ec2-54-174-17-233.compute-1.amazonaws.com:54310/user/hduser/input/10gb_data.txt::1073741824+134217728
16/03/26 02:50:53 INFO Executor: Running task 9.0 in stage 0.0 (TID 9)
16/03/26 02:50:53 INFO TaskSetManager: Finished task 9.0 in stage 0.0 (TID 9) in 2261 ms on localhost (10/75)
16/03/26 02:50:55 INFO Executor: Running task 10.0 in stage 0.0 (TID 10)
16/03/26 02:50:55 INFO TaskSetManager: Finished task 10.0 in stage 0.0 (TID 10) in 2261 ms on localhost (10/75)
16/03/26 02:50:55 INFO HadoopRDD: Input split: hdfs://ec2-54-174-17-233.compute-1.amazonaws.com:54310/user/hduser/input/10gb_data.txt::134217728+134217728
16/03/26 02:50:55 INFO Executor: Running task 10.0 in stage 0.0 (TID 10)
16/03/26 02:50:55 INFO TaskSetManager: Finished task 10.0 in stage 0.0 (TID 10) in 2261 ms on localhost (11/75)
16/03/26 02:50:55 INFO HadoopRDD: Input split: hdfs://ec2-54-174-17-233.compute-1.amazonaws.com:54310/user/hduser/input/10gb_data.txt::1207959552+134217728
16/03/26 02:50:55 INFO Executor: Running task 11.0 in stage 0.0 (TID 11)
16/03/26 02:50:55 INFO TaskSetManager: Starting task 11.0 in stage 0.0 (TID 11)
16/03/26 02:50:55 INFO HadoopRDD: Input split: hdfs://ec2-54-174-17-233.compute-1.amazonaws.com:54310/user/hduser/input/10gb_data.txt::1470395008+134217728
16/03/26 02:50:55 INFO Executor: Finished task 11.0 in stage 0.0 (TID 11) in 2231 ms on localhost (12/75)
16/03/26 02:50:55 INFO TaskSetManager: Finished task 11.0 in stage 0.0 (TID 11) in 2231 ms on localhost (12/75)
16/03/26 02:50:55 INFO HadoopRDD: Input split: hdfs://ec2-54-174-17-233.compute-1.amazonaws.com:54310/user/hduser/input/10gb_data.txt::134217728+134217728
16/03/26 02:51:01 INFO Executor: Running task 12.0 in stage 0.0 (TID 12)
16/03/26 02:51:01 INFO TaskSetManager: Starting task 12.0 in stage 0.0 (TID 12)
16/03/26 02:51:01 INFO HadoopRDD: Input split: hdfs://ec2-54-174-17-233.compute-1.amazonaws.com:54310/user/hduser/input/10gb_data.txt::1610612736+134217728
16/03/26 02:51:01 INFO Executor: Finished task 12.0 in stage 0.0 (TID 12) in 2235 ms on localhost (13/75)
16/03/26 02:51:01 INFO TaskSetManager: Finished task 12.0 in stage 0.0 (TID 12) in 2235 ms on localhost (13/75)
16/03/26 02:51:01 INFO HadoopRDD: Input split: hdfs://ec2-54-174-17-233.compute-1.amazonaws.com:54310/user/hduser/input/10gb_data.txt::1744830464+134217728
16/03/26 02:51:01 INFO Executor: Finished task 13.0 in stage 0.0 (TID 13)
16/03/26 02:51:01 INFO TaskSetManager: Starting task 13.0 in stage 0.0 (TID 13)
16/03/26 02:51:01 INFO HadoopRDD: Input split: hdfs://ec2-54-174-17-233.compute-1.amazonaws.com:54310/user/hduser/input/10gb_data.txt::1744830464+134217728
16/03/26 02:51:01 INFO Executor: Finished task 13.0 in stage 0.0 (TID 13) in 2238 ms on localhost (14/75)
16/03/26 02:51:01 INFO TaskSetManager: Finished task 13.0 in stage 0.0 (TID 13) in 2238 ms on localhost (14/75)
16/03/26 02:51:01 INFO HadoopRDD: Input split: hdfs://ec2-54-174-17-233.compute-1.amazonaws.com:54310/user/hduser/input/10gb_data.txt::1879048192+134217728
16/03/26 02:51:06 INFO Executor: Finished task 14.0 in stage 0.0 (TID 14)
16/03/26 02:51:06 INFO TaskSetManager: Starting task 14.0 in stage 0.0 (TID 14)
16/03/26 02:51:06 INFO HadoopRDD: Input split: hdfs://ec2-54-174-17-233.compute-1.amazonaws.com:54310/user/hduser/input/10gb_data.txt::1610612736+134217728
16/03/26 02:51:06 INFO Executor: Finished task 14.0 in stage 0.0 (TID 14) in 2238 ms on localhost (15/75)
16/03/26 02:51:06 INFO TaskSetManager: Finished task 14.0 in stage 0.0 (TID 14) in 2238 ms on localhost (15/75)
16/03/26 02:51:06 INFO HadoopRDD: Input split: hdfs://ec2-54-174-17-233.compute-1.amazonaws.com:54310/user/hduser/input/10gb_data.txt::1747483648+134217728
16/03/26 02:51:08 INFO Executor: Running task 15.0 in stage 0.0 (TID 15)
16/03/26 02:51:08 INFO TaskSetManager: Finished task 15.0 in stage 0.0 (TID 15) in 2233 ms on localhost (16/75)
16/03/26 02:51:08 INFO HadoopRDD: Input split: hdfs://ec2-54-174-17-233.compute-1.amazonaws.com:54310/user/hduser/input/10gb_data.txt::1744830464+134217728
16/03/26 02:51:08 INFO Executor: Running task 15.0 in stage 0.0 (TID 15) in 2233 ms on localhost (16/75)
16/03/26 02:51:08 INFO TaskSetManager: Finished task 15.0 in stage 0.0 (TID 15) in 2233 ms on localhost (17/75)
16/03/26 02:51:08 INFO HadoopRDD: Input split: hdfs://ec2-54-174-17-233.compute-1.amazonaws.com:54310/user/hduser/input/10gb_data.txt::1747483648+134217728
16/03/26 02:51:10 INFO Executor: Running task 16.0 in stage 0.0 (TID 16)
16/03/26 02:51:10 INFO TaskSetManager: Finished task 16.0 in stage 0.0 (TID 16) in 2258 ms on localhost (17/75)
16/03/26 02:51:10 INFO HadoopRDD: Input split: hdfs://ec2-54-174-17-233.compute-1.amazonaws.com:54310/user/hduser/input/10gb_data.txt::2281701376+134217728
16/03/26 02:51:13 INFO Executor: Finished task 17.0 in stage 0.0 (TID 17)
16/03/26 02:51:13 INFO TaskSetManager: Starting task 17.0 in stage 0.0 (TID 17)
16/03/26 02:51:13 INFO HadoopRDD: Input split: hdfs://ec2-54-174-17-233.compute-1.amazonaws.com:54310/user/hduser/input/10gb_data.txt::1879048192+134217728
16/03/26 02:51:13 INFO Executor: Running task 18.0 in stage 0.0 (TID 18)
16/03/26 02:51:13 INFO TaskSetManager: Finished task 17.0 in stage 0.0 (TID 17) in 2223 ms on localhost (18/75)
16/03/26 02:51:13 INFO HadoopRDD: Input split: hdfs://ec2-54-174-17-233.compute-1.amazonaws.com:54310/user/hduser/input/10gb_data.txt::2415919104+134217728
```

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

```
hdsuser@ip-172-31-27-143:/usr/local/spark
vishwanath@vishwanath-0551LB:~Documents
x hdsuser@ip-172-31-27-143:/usr/local/spark

[...]
16/03/26 03:04:53 INFO SparkKafkaInputFormat: Finished task 70.0 in stage 2.0 (TID 220). 2080 bytes result sent to driver
16/03/26 03:04:53 INFO TaskSetManager: Finished task 70.0 in stage 2.0 (TID 221), localhost, partition 71, NODE_LOCAL, 1961 bytes)
16/03/26 03:04:53 INFO Executor: Running task 71.0 in stage 2.0 (TID 221)
16/03/26 03:04:53 INFO ShuffleBlockFetcherIterator: Getting 75 non-empty blocks out of 75 blocks
16/03/26 03:04:53 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
16/03/26 03:05:01 INFO FileOutputCommitter: Saved output of task 'attempt_201603260253_0002_m_000071_221' to hdfs://ec2-54-174-17-233.compute-1.amazonaws.com:54310/output_spark_final/_temporary/0/task_201603260253_0002_m_000071_221
16/03/26 03:05:01 INFO SparkKafkaInputFormat: Attempted attempt_201603260253_0002_m_000071_221: Committed
16/03/26 03:05:01 INFO Executor: Finished task 71.0 in stage 2.0 (TID 221). 2080 bytes result sent to driver
16/03/26 03:05:01 INFO TaskSetManager: Starting task 72.0 in stage 2.0 (TID 222, localhost, partition 72, NODE_LOCAL, 1961 bytes)
16/03/26 03:05:01 INFO TaskSetManager: Finished task 71.0 in stage 2.0 (TID 221) in 8306 ms on localhost (73/75)
16/03/26 03:05:01 INFO Executor: Running task 72.0 in stage 2.0 (TID 222)
16/03/26 03:05:01 INFO ShuffleBlockFetcherIterator: Getting 75 non-empty blocks out of 75 blocks
16/03/26 03:05:01 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
16/03/26 03:05:07 INFO FileOutputCommitter: Saved output of task 'attempt_201603260253_0002_m_000072_222' to hdfs://ec2-54-174-17-233.compute-1.amazonaws.com:54310/output_spark_final/_temporary/0/task_201603260253_0002_m_000072_222
16/03/26 03:05:07 INFO SparkKafkaInputFormat: Attempted attempt_201603260253_0002_m_000072_222: Committed
16/03/26 03:05:07 INFO Executor: Finished task 72.0 in stage 2.0 (TID 222). 2080 bytes result sent to driver
16/03/26 03:05:07 INFO TaskSetManager: Starting task 73.0 in stage 2.0 (TID 223, localhost, partition 73, NODE_LOCAL, 1961 bytes)
16/03/26 03:05:07 INFO TaskSetManager: Finished task 72.0 in stage 2.0 (TID 222) in 8306 ms on localhost (73/75)
16/03/26 03:05:07 INFO Executor: Running task 73.0 in stage 2.0 (TID 223)
16/03/26 03:05:07 INFO ShuffleBlockFetcherIterator: Getting 75 non-empty blocks out of 75 blocks
16/03/26 03:05:07 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
16/03/26 03:05:07 INFO FileOutputCommitter: Saved output of task 'attempt_201603260253_0002_m_000073_223' to hdfs://ec2-54-174-17-233.compute-1.amazonaws.com:54310/output_spark_final/_temporary/0/task_201603260253_0002_m_000073_223
16/03/26 03:05:13 INFO SparkKafkaInputFormat: Attempted attempt_201603260253_0002_m_000073_223: Committed
16/03/26 03:05:13 INFO Executor: Finished task 73.0 in stage 2.0 (TID 223). 2080 bytes result sent to driver
16/03/26 03:05:13 INFO TaskSetManager: Starting task 74.0 in stage 2.0 (TID 224, localhost, partition 74, NODE_LOCAL, 1961 bytes)
16/03/26 03:05:13 INFO TaskSetManager: Finished task 73.0 in stage 2.0 (TID 223) in 5541 ms on localhost (74/75)
16/03/26 03:05:13 INFO Executor: Running task 74.0 in stage 2.0 (TID 224)
16/03/26 03:05:13 INFO ShuffleBlockFetcherIterator: Getting 75 non-empty blocks out of 75 blocks
16/03/26 03:05:13 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
16/03/26 03:05:17 INFO FileOutputCommitter: Saved output of task 'attempt_201603260253_0002_m_000074_224' to hdfs://ec2-54-174-17-233.compute-1.amazonaws.com:54310/output_spark_final/_temporary/0/task_201603260253_0002_m_000074_224
16/03/26 03:05:17 INFO SparkKafkaInputFormat: Attempted attempt_201603260253_0002_m_000074_224: Committed
16/03/26 03:05:17 INFO Executor: Finished task 74.0 in stage 2.0 (TID 224). 2080 bytes result sent to driver
16/03/26 03:05:17 INFO TaskSetManager: Finished task 74.0 in stage 2.0 (TID 225) in 456 ms on localhost (75/75)
16/03/26 03:05:17 INFO DAGScheduler: ResultStage 2 (saveAsTextFile at tera.SortJava46) finished in 455.366 s
16/03/26 03:05:17 INFO TaskSchedulerImpl: Removed Taskset 2.0, whose tasks have all completed, from pool
16/03/26 03:05:17 INFO DAGScheduler: Job 1 finished: saveAsTextFile at tera.SortJava46, took 717.520384 s
16/03/26 03:05:18 INFO SparkUptimeTrackerMasterEndpoint: UptimeTrackerMasterEndpoint stopped!
16/03/26 03:05:18 INFO SparkUptimeTrackerMaster: Stopped
16/03/26 03:05:18 INFO BlockManager: BlockManager stopped
16/03/26 03:05:18 INFO BlockManagerMaster: BlockManagerMaster stopped
16/03/26 03:05:18 INFO OutputCommitCoordinator: OutputCommitCoordinator stopped!
16/03/26 03:05:18 INFO SparkContext: Successfully stopped SparkContext
16/03/26 03:05:18 INFO DAGScheduler: Stopped DAGScheduler
16/03/26 03:05:18 INFO ShutdownHookHandler: Deleting directory /tmp/spark-40093be0-5b65-4a7e-be21-58c1b48b272e
16/03/26 03:05:18 INFO RemoteActorRefProvider$RemoteTerminator: Shutting down remote daemon.
16/03/26 03:05:18 INFO ShutdownHookManager: Deleting directory /tmp/spark-40093be0-5b65-4a7e-be21-58c1b48b272e/httpd-2d19ddda-5ac6-4b6a-8b44-fbd7eed0328e
16/03/26 03:05:18 INFO RemoteActorRefProvider$RemoteTerminator: Remote daemon shut down; proceeding with flushing remote transports.
16/03/26 03:05:18 INFO RemoteActorRefProvider$RemoteTerminator: Remoting shut down.
hdsuser@ip-172-31-27-143:/usr/local/sparks ]
```

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

```
hadoop@ip-172-31-40-249:/mnt/raid/64 x hduuser@vishwanath-Q551LB: /home/vishwanath/workspace/bidari x vishwanath@vishwanath-Q551LB: ~/Documents x
[Output of previous command continues]
16/03/28 04:54:34 INFO TaskSetManager: Finished task 29.0 in stage 2.0 (TID 89) in 1385 ms on localhost (30/30)
16/03/28 04:54:34 INFO TaskSchedulerImpl: Removed Taskset 2.0, whose tasks have all completed, from pool
16/03/28 04:54:34 INFO DAGScheduler: Job 1 finished; collect at tera_Sort.java:46
16/03/28 04:54:34 INFO SparkContext: Starting job: saveAsTextFile at tera_Sort.java:46
16/03/28 04:54:34 INFO DAGScheduler: Got job 2 (saveAsTextFile at tera_Sort.java:46) with 1 output partitions
16/03/28 04:54:34 INFO DAGScheduler: Final stage: ResultStage 3 (saveAsTextFile at tera_Sort.java:46)
16/03/28 04:54:34 INFO DAGScheduler: Parents of final stage: List()
16/03/28 04:54:34 INFO DAGScheduler: Submitting ResultStage 3 (MapPartitionsRDD[8] at saveAsTextFile at tera_Sort.java:46), which has no missing parents
16/03/28 04:54:34 INFO MemoryStore: Block broadcast_4 stored as values in memory (estimated size 63.7 KB, free 237.7 KB)
16/03/28 04:54:34 INFO MemoryStore: Block broadcast_4_piece0 stored as bytes in memory (estimated size 21.6 KB, free 259.3 KB)
16/03/28 04:54:34 INFO BlockManagerInfo: Added broadcast_4_piece0 in memory localhost:59531 (size: 21.6 KB, free: 5.5 GB)
16/03/28 04:54:34 INFO SparkContext: Created broadcast 4 from broadcast at DAGScheduler.scala:1006
16/03/28 04:54:34 INFO DAGScheduler: Submitted 1 missing tasks in stage ResultStage 3 (MapPartitionsRDD[8] at saveAsTextFile at tera_Sort.java:46)
16/03/28 04:54:34 INFO DAGScheduler: Adding missing 1 pending tasks
16/03/28 04:54:45 INFO BlockManagerInfo: Removed broadcast_3_piece0 on localhost:59531 in memory (size: 2.5 KB, free: 5.5 GB)
16/03/28 04:54:48 INFO ContextCleaner: Cleamed accumulator 3
16/03/28 04:54:48 INFO ContextCleaner: Cleamed accumulator 2
16/03/28 04:54:48 WARN TaskSetManager: Stage 3 contains a task of very large size (986330 KB). The maximum recommended task size is 100 KB.
16/03/28 04:54:48 INFO TaskSetManager: Starting stage 3.0 (TID 98, localhost, partition 0,PROCESS_LOCAL, 1010602147 bytes)
16/03/28 04:54:48 INFO Executor: Running task 0.0 in stage 3.0 (TID 98)
16/03/28 04:55:40 INFO FileoutputCommitter: Saved output of task attempt_201603200454_0003_m_000000_90 to file:/usr/local/spark-1.6.1-bin-hadoop2.6/output/_temporary/0/task_201603200454_0003_m_000000
16/03/28 04:55:40 INFO SparkHadoopMapRedUtil: attempt_201603200454_0003_m_000000_90: Committed
16/03/28 04:55:40 INFO Executor: Finished task 0.0 in stage 3.0 (TID 98). 1864 bytes result sent to driver
16/03/28 04:55:40 INFO BlockManager: Finished task 0.0 in stage 3.0 (TID 98) in 6505 ms on localhost (1/1)
16/03/28 04:55:40 INFO DAGScheduler: Redistributing tasks after shuffle (stage 3.0) (0 partitions) finished in 65.05 s
16/03/28 04:55:40 INFO TaskSchedulerImpl: Removed Taskset 3.0, whose tasks have all completed, from pool
16/03/28 04:55:40 INFO DAGScheduler: Job 2 (finished: saveAsTextFile at tera_Sort.java:46) took 0:33:113 s
16/03/28 04:55:40 INFO SparkContext: Invoking stop() from shutdown hook
16/03/28 04:55:40 INFO SparkUI: Stopped Spark web UI at http://172.31.31.40:4040
16/03/28 04:55:40 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
16/03/28 04:55:40 INFO BlockManager: BlockManager stopped
16/03/28 04:55:40 INFO BlockManagerMaster: BlockManagerMaster stopped
16/03/28 04:55:40 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
16/03/28 04:55:40 INFO RemoteActorRefProvider$RemoteTerminator: Shutting down remote daemon.
16/03/28 04:55:40 INFO SparkContext: Accessory stopped SparkContext
16/03/28 04:55:40 INFO ShutdownHookManager: Shutting down
16/03/28 04:55:40 INFO ShutdownHookManager: Deleting directory /tmp/spark-42ff385e-b713-42ef-9043-891205b72a5d/httpd-d6188adcfcea-4adb-a3ef-1e57fdda2f11
16/03/28 04:55:40 INFO RemoteActorRefProvider$RemoteTerminator: Remote daemon shut down; proceeding with flushing remote transports.
16/03/28 04:55:40 INFO ShutdownHookManager: Deleting directory /tmp/spark-42ff385e-b713-42ef-9043-891205b72a5d
hadoop@ip-172-31-40-249:/usr/local/spark-1.6.1-bin-hadoop0.5 $ ls
bidari-0.1-SNAPSHOT.jar bin CHANGES.txt conf data ec2 executors lib LICENSE license logs NOTICE output2 python R README.md RELEASE sbin small_data.txt test.txt work
hadoop@ip-172-31-40-249:/usr/local/spark-1.6.1-bin-hadoop0.5 $ ls output2/
hadoop@ip-172-31-40-249:/usr/local/spark-1.6.1-bin-hadoop0.5 $ ls output2/
hadoop@ip-172-31-40-249:/mnt/raid/645 sudo ./valsort /usr/local/spark-1.6.1-bin-hadoop2.6/output2/part-00000
Records: 10000000
Checksum: 4c48a881c779ds
Duplicate keys: 0
SUCCESS - all records are in order
hadoop@ip-172-31-40-249:/mnt/raid/645 |
```

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

EC2 Management Console - Mozilla Firefox

(no subject)-vishw... All Mail - vbdari@h... EC2 Management C... Running Spark on Ec... AWS Support Dash... Facebook Sort Benchmark Dat... 100 gb to bytes - Go... + 4:23 PM

https://console.aws.amazon.com/ec2/v2/home?region=us-east-1#instances:sort=instanceId

Search

Vishwanath N. Virginia Support

EC2 Dashboard Services Edit

Launch Instance Connect Actions

Filter by tags and attributes or search by keyword

1 to 17 of 17 >

Name	Instance ID	Instance Type	Availability Zone	Instance State	Status Checks	Alarm Status	Public DNS	Public IP	Key Name	Monitoring	Launch Time
i-409bbdc3	c3.large	us-east-1d	running	Initializing	None	ec2-52-207-234-21.com...	52.207.234.21	IronMan_Spark	disabled	March 24, 2016 at 4:20...	
i-419bbdc2	c3.large	us-east-1d	running	Initializing	None	ec2-54-164-94-72.com...	54.164.94.72	IronMan_Spark	disabled	March 24, 2016 at 4:20...	
i-429bbdc1	c3.large	us-east-1d	running	Initializing	None	ec2-52-87-161-46.com...	52.87.161.46	IronMan_Spark	disabled	March 24, 2016 at 4:20...	
i-439bbdc0	c3.large	us-east-1d	running	Initializing	None	ec2-54-86-202-41.com...	54.86.202.41	IronMan_Spark	disabled	March 24, 2016 at 4:20...	
<input checked="" type="checkbox"/> i-6bcfe9e8	c3.large	us-east-1d	running	2/2 checks ...	None	ec2-52-90-9-69.compute...	52.90.9.69	IronMan_Spark	disabled	March 24, 2016 at 4:08...	
i-bf9abc37	c3.large	us-east-1d	running	Initializing	None	ec2-52-23-183-238.com...	52.23.183.238	IronMan_Spark	disabled	March 24, 2016 at 4:20...	
i-b59abc36	c3.large	us-east-1d	running	Initializing	None	ec2-52-23-167-187.com...	52.23.167.187	IronMan_Spark	disabled	March 24, 2016 at 4:20...	
i-b69abc35	c3.large	us-east-1d	running	Initializing	None	ec2-52-207-234-191.co...	52.207.234.191	IronMan_Spark	disabled	March 24, 2016 at 4:20...	
i-b79abc34	c3.large	us-east-1d	running	Initializing	None	ec2-52-90-176-135.com...	52.90.176.135	IronMan_Spark	disabled	March 24, 2016 at 4:20...	
i-b89abc3b	c3.large	us-east-1d	running	Initializing	None	ec2-52-207-212-226.co...	52.207.212.226	IronMan_Spark	disabled	March 24, 2016 at 4:20...	
i-b99abc3a	c3.large	us-east-1d	running	Initializing	None	ec2-54-98-124-240.com...	54.86.124.240	IronMan_Spark	disabled	March 24, 2016 at 4:20...	
i-ba9abc39	c3.large	us-east-1d	running	Initializing	None	ec2-52-201-216-28.com...	52.201.216.28	IronMan_Spark	disabled	March 24, 2016 at 4:20...	
i-bb9abc38	c3.large	us-east-1d	running	Initializing	None	ec2-52-207-219-92.com...	52.207.219.92	IronMan_Spark	disabled	March 24, 2016 at 4:20...	
i-bc9abc3f	c3.large	us-east-1d	running	Initializing	None	ec2-52-87-187-76.com...	52.87.187.76	IronMan_Spark	disabled	March 24, 2016 at 4:20...	
i-bd9abc3e	c3.large	us-east-1d	running	Initializing	None	ec2-54-88-33-225.com...	54.88.33.225	IronMan_Spark	disabled	March 24, 2016 at 4:20...	
i-be9abc3d	c3.large	us-east-1d	running	Initializing	None	ec2-54-85-163-112.com...	54.85.163.112	IronMan_Spark	disabled	March 24, 2016 at 4:20...	
i-bf9abc3c	c3.large	us-east-1d	running	Initializing	None	ec2-52-87-247-102.com...	52.87.247.102	IronMan_Spark	disabled	March 24, 2016 at 4:20...	

Instance: i-6bcfe9e8 Public DNS: ec2-52-90-9-69.compute-1.amazonaws.com

Feedback English

© 2008 - 2016, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

```
root@ip-172-31-52-97:/usr/local/spark/ec2          x  ubuntu@ip-172-31-52-97:/usr/local/spark/ec2          x  root@ip-172-31-52-97:/usr/local/spark/ec2
[...]
Deploying Spark config files...
RSYNCing /root/spark to slaves...
ec2-54-174-6-218.compute-1.amazonaws.com
ec2-54-174-182-179.compute-1.amazonaws.com
ec2-54-86-164-118.compute-1.amazonaws.com
ec2-52-201-246-219.compute-1.amazonaws.com
ec2-54-174-213-138.compute-1.amazonaws.com
ec2-54-172-101-191.compute-1.amazonaws.com
ec2-54-84-198-197.compute-1.amazonaws.com
ec2-52-87-184-98.compute-1.amazonaws.com
ec2-52-90-3-31.compute-1.amazonaws.com
ec2-52-90-251-120.compute-1.amazonaws.com
ec2-52-90-228-120.compute-1.amazonaws.com
ec2-52-207-225-183.compute-1.amazonaws.com
ec2-54-172-101-191.compute-1.amazonaws.com
ec2-52-207-225-189.compute-1.amazonaws.com
ec2-52-90-54-28.compute-1.amazonaws.com
ec2-52-207-241-80.compute-1.amazonaws.com
Setting up scala...
RSYNCing /root/scala to slaves...
ec2-54-174-6-218.compute-1.amazonaws.com
ec2-54-174-182-179.compute-1.amazonaws.com
ec2-54-86-164-118.compute-1.amazonaws.com
ec2-52-201-246-219.compute-1.amazonaws.com
ec2-54-174-213-138.compute-1.amazonaws.com
ec2-54-172-101-191.compute-1.amazonaws.com
ec2-54-84-198-197.compute-1.amazonaws.com
ec2-52-87-184-98.compute-1.amazonaws.com
ec2-52-90-3-31.compute-1.amazonaws.com
ec2-52-90-251-120.compute-1.amazonaws.com
ec2-52-90-228-120.compute-1.amazonaws.com
ec2-54-174-213-163.compute-1.amazonaws.com
ec2-54-172-101-191.compute-1.amazonaws.com
ec2-52-207-225-189.compute-1.amazonaws.com
ec2-52-90-54-28.compute-1.amazonaws.com
ec2-52-207-241-80.compute-1.amazonaws.com
Setting up spark...
RSYNCing /root/spark to slaves...
ec2-54-174-6-218.compute-1.amazonaws.com
ec2-54-174-182-179.compute-1.amazonaws.com
ec2-54-86-164-118.compute-1.amazonaws.com
ec2-52-201-246-219.compute-1.amazonaws.com
ec2-54-174-213-138.compute-1.amazonaws.com
ec2-54-173-174-212.compute-1.amazonaws.com
ec2-54-84-198-197.compute-1.amazonaws.com
ec2-52-87-184-98.compute-1.amazonaws.com
ec2-52-90-3-31.compute-1.amazonaws.com
ec2-52-90-251-120.compute-1.amazonaws.com
ec2-51-90-228-117.compute-1.amazonaws.com
ec2-54-174-6-183.compute-1.amazonaws.com
ec2-54-172-101-191.compute-1.amazonaws.com
ec2-52-207-225-189.compute-1.amazonaws.com
ec2-52-90-54-28.compute-1.amazonaws.com
ec2-52-207-241-80.compute-1.amazonaws.com
```

```
root@ip-172-31-52-97: /usr/local/spark/ec2
ubuntu@ip-172-31-52-97: /usr/local/spark/ec2
x  ubuntu@ip-172-31-52-97: ~
x  root@ip-172-31-52-97: /usr/local/spark/ec2
x


ec2-52-90-54-20.compute.1.amazonaws.com
ec2-52-207-241-80.compute.1.amazonaws.com
ec2-52-207-241-80.compute.1.amazonaws.com: no org.apache.spark.deploy.worker.Worker to stop
ec2-52-207-241-80.compute.1.amazonaws.com: no org.apache.spark.deploy.worker.Worker to stop
ec2-52-90-251-120.compute.1.amazonaws.com: no org.apache.spark.deploy.worker.Worker to stop
ec2-52-90-54-20.compute.1.amazonaws.com: no org.apache.spark.deploy.worker.Worker to stop
ec2-52-174-6-218.compute.1.amazonaws.com: no org.apache.spark.deploy.worker.Worker to stop
ec2-54-174-6-183.compute.1.amazonaws.com: no org.apache.spark.deploy.worker.Worker to stop
ec2-54-174-6-183.compute.1.amazonaws.com: no org.apache.spark.deploy.worker.Worker to stop
ec2-54-174-181-193.compute.1.amazonaws.com: no org.apache.spark.deploy.worker.Worker to stop
ec2-52-207-225-189.compute.1.amazonaws.com: no org.apache.spark.deploy.worker.Worker to stop
ec2-52-87-184-98.compute.1.amazonaws.com: no org.apache.spark.deploy.worker.Worker to stop
ec2-52-90-3-31.compute.1.amazonaws.com: no org.apache.spark.deploy.worker.Worker to stop
ec2-52-90-228-177.compute.1.amazonaws.com: no org.apache.spark.deploy.worker.Worker to stop
ec2-54-174-164-219.compute.1.amazonaws.com: no org.apache.spark.deploy.worker.Worker to stop
ec2-54-174-174-211.compute.1.amazonaws.com: no org.apache.spark.deploy.worker.Worker to stop
ec2-52-201-246-219.compute.1.amazonaws.com: no org.apache.spark.deploy.worker.Worker to stop
ec2-54-174-182-179.compute.1.amazonaws.com: no org.apache.spark.deploy.worker.Worker to stop
org.apache.spark.deploy.master.Master to stop
starting org.apache.spark.deploy.master.Master to log to /root/spark/logs/spark-root.org.apache.spark.deploy.master.Master-1-ip-172-31-7-161.ec2.internal.out
ec2-52-90-228-177.compute.1.amazonaws.com: starting org.apache.spark.deploy.worker.Worker, logging to /root/spark/logs/spark-root.org.apache.spark.deploy.worker.Worker-1-ip-172-31-2-154.ec2.internal.out
ec2-52-90-228-177.compute.1.amazonaws.com: starting org.apache.spark.deploy.worker.Worker, logging to /root/spark/logs/spark-root.org.apache.spark.deploy.worker.Worker-1-ip-172-31-7-172.ec2.internal.out
ec2-54-174-6-218.compute.1.amazonaws.com: starting org.apache.spark.deploy.worker.Worker, logging to /root/spark/logs/spark-root.org.apache.spark.deploy.worker.Worker-1-ip-172-31-7-27.ec2.internal.out
ec2-52-90-3-31.compute.1.amazonaws.com: starting org.apache.spark.deploy.worker.Worker, logging to /root/spark/logs/spark-root.org.apache.spark.deploy.worker.Worker-1-ip-172-31-1-89.ec2.internal.out
ec2-52-90-54-20.compute.1.amazonaws.com: starting org.apache.spark.deploy.worker.Worker, logging to /root/spark/logs/spark-root.org.apache.spark.deploy.worker.Worker-1-ip-172-31-5-14.ec2.internal.out
ec2-52-207-241-80.compute.1.amazonaws.com: starting org.apache.spark.deploy.worker.Worker, logging to /root/spark/logs/spark-root.org.apache.spark.deploy.worker.Worker-1-ip-172-31-8-101.ec2.internal.out
ec2-52-207-241-80.compute.1.amazonaws.com: starting org.apache.spark.deploy.worker.Worker, logging to /root/spark/logs/spark-root.org.apache.spark.deploy.worker.Worker-1-ip-172-31-8-102.ec2.internal.out
ec2-52-207-275-189.compute.1.amazonaws.com: starting org.apache.spark.deploy.worker.Worker, logging to /root/spark/logs/spark-root.org.apache.spark.deploy.worker.Worker-1-ip-172-31-6-60.ec2.internal.out
ec2-54-174-182-179.compute.1.amazonaws.com: starting org.apache.spark.deploy.worker.Worker, logging to /root/spark/logs/spark-root.org.apache.spark.deploy.worker.Worker-1-ip-172-31-9-192.ec2.internal.out
ec2-54-174-174-211.compute.1.amazonaws.com: starting org.apache.spark.deploy.worker.Worker, logging to /root/spark/logs/spark-root.org.apache.spark.deploy.worker.Worker-1-ip-172-31-9-193.ec2.internal.out
ec2-52-201-246-219.compute.1.amazonaws.com: starting org.apache.spark.deploy.worker.Worker, logging to /root/spark/logs/spark-root.org.apache.spark.deploy.worker.Worker-1-ip-172-31-13-209.ec2.internal.out
ec2-54-84-184-197.compute.1.amazonaws.com: starting org.apache.spark.deploy.worker.Worker, logging to /root/spark/logs/spark-root.org.apache.spark.deploy.worker.Worker-1-ip-172-31-13-53.ec2.internal.out
ec2-54-86-164-118.compute.1.amazonaws.com: starting org.apache.spark.deploy.worker.Worker, logging to /root/spark/logs/spark-root.org.apache.spark.deploy.worker.Worker-1-ip-172-31-10-55.ec2.internal.out
ec2-54-174-6-183.compute.1.amazonaws.com: starting org.apache.spark.deploy.worker.Worker, logging to /root/spark/logs/spark-root.org.apache.spark.deploy.worker.Worker-1-ip-172-31-14-167.ec2.internal.out
[timing] spark-standalone setup: 00h 05s
Setting up tachyon
open http://tachyon-slaves...
ec2-54-174-6-218.compute.1.amazonaws.com
ec2-54-174-182-179.compute.1.amazonaws.com
ec2-54-86-164-118.compute.1.amazonaws.com
ec2-52-201-246-219.compute.1.amazonaws.com
ec2-54-174-213-138.compute.1.amazonaws.com
ec2-54-174-174-211.compute.1.amazonaws.com
ec2-54-84-184-197.compute.1.amazonaws.com
ec2-54-84-184-197.compute.1.amazonaws.com
ec2-52-87-184-98.compute.1.amazonaws.com
ec2-52-90-3-31.compute.1.amazonaws.com
ec2-50-295-120.compute.1.amazonaws.com
ec2-54-174-6-183.compute.1.amazonaws.com
ec2-54-172-181-191.compute.1.amazonaws.com
ec2-52-207-225-189.compute.1.amazonaws.com
ec2-52-90-54-20.compute.1.amazonaws.com
ec2-52-207-241-80.compute.1.amazonaws.com
```

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

```
root@ip-172-31-52-97:/usr/local/spark/ec2
ubuntu@ip-172-31-52-97:/usr/local/spark/ec2          x  ubuntu@ip-172-31-52-97:~
Starting GANGLIA gmond: [ OK ]
Connection to ec2-54-174-182-179.compute-1.amazonaws.com closed. [FAILED]
Shutting down GANGLIA gmond: [ OK ]
Starting GANGLIA gmond: [ OK ]
Connection to ec2-54-86-164-118.compute-1.amazonaws.com closed. [FAILED]
Shutting down GANGLIA gmond: [ OK ]
Starting GANGLIA gmond: [ OK ]
Connection to ec2-52-201-246-219.compute-1.amazonaws.com closed. [FAILED]
Shutting down GANGLIA gmond: [ OK ]
Starting GANGLIA gmond: [ OK ]
Connection to ec2-54-174-213-138.compute-1.amazonaws.com closed. [FAILED]
Shutting down GANGLIA gmond: [ OK ]
Starting GANGLIA gmond: [ OK ]
Connection to ec2-54-173-174-212.compute-1.amazonaws.com closed. [FAILED]
Shutting down GANGLIA gmond: [ OK ]
Starting GANGLIA gmond: [ OK ]
Connection to ec2-54-194-198-197.compute-1.amazonaws.com closed. [FAILED]
Shutting down GANGLIA gmond: [ OK ]
Starting GANGLIA gmond: [ OK ]
Connection to ec2-52-87-184-98.compute-1.amazonaws.com closed. [FAILED]
Shutting down GANGLIA gmond: [ OK ]
Starting GANGLIA gmond: [ OK ]
Connection to ec2-54-176-3-31.compute-1.amazonaws.com closed. [FAILED]
Shutting down GANGLIA gmond: [ OK ]
Starting GANGLIA gmond: [ OK ]
Connection to ec2-52-90-251-120.compute-1.amazonaws.com closed. [FAILED]
Shutting down GANGLIA gmond: [ OK ]
Starting GANGLIA gmond: [ OK ]
Connection to ec2-50-228-177.compute-1.amazonaws.com closed. [FAILED]
Shutting down GANGLIA gmond: [ OK ]
Starting GANGLIA gmond: [ OK ]
Connection to ec2-54-174-6-183.compute-1.amazonaws.com closed. [FAILED]
Shutting down GANGLIA gmond: [ OK ]
Starting GANGLIA gmond: [ OK ]
Connection to ec2-52-172-161-193.compute-1.amazonaws.com closed. [FAILED]
Shutting down GANGLIA gmond: [ OK ]
Starting GANGLIA gmond: [ OK ]
Connection to ec2-52-207-225-189.compute-1.amazonaws.com closed. [FAILED]
Shutting down GANGLIA gmond: [ OK ]
Starting GANGLIA gmond: [ OK ]
Connection to ec2-50-54-20.compute-1.amazonaws.com closed. [FAILED]
Shutting down GANGLIA gmond: [ OK ]
Starting GANGLIA gmond: [ OK ]
Connection to ec2-52-207-241-80.compute-1.amazonaws.com closed. [FAILED]
Shutting down GANGLIA gmetad: [ OK ]
Starting GANGLIA gmetad: [ OK ]
Stopping httpd: [OK]
starting httpd: Syntax error on line 154 of /etc/httpd/conf/httpd.conf: Cannot load /etc/httpd/modules/mod_authz_core.so into server: /etc/httpd/modules/mod_authz_core.so: cannot open shared object file: No such file or directory [FAILED]
[time] ganglia setup: 00h 00m 12s
Connection to ec2-54-175-44-165.compute-1.amazonaws.com closed.
Spark standalone cluster started at http://ec2-54-175-44-165.compute-1.amazonaws.com:8080
Data started at http://ec2-54-175-44-165.compute-1.amazonaws.com:5000/ganglia
done!
root@ip-172-31-52-97:/usr/local/spark/ec2#
```

TeraSort - Spark Jobs - Mozilla Firefox 12:09 PM

Spark 1.8.1 Jobs Stages Storage Environment Executors TeraSort application UI

<http://ec2-54-85-192-104.compute-1.amazonaws.com:4040/jobs/>

**Spark Jobs (2)**

Total Uptime: 39 min  
Scheduling Mode: FIFO  
**Active Jobs:** 1  
**Completed Jobs:** 1

Event Timeline  Enable zooming

Executors  
Added    Removed    never added

Jobs  
Succeeded    Failed    Running

sortByKey at tera\_Sort.java:39 (Job 0)    saveAsTextFile at tera\_Sort.java:46 (Job 1)

11:30 11:35 11:40 11:45 11:50 11:55 12:00 12:05 12:10  
Tue 29 March

**Active Jobs (1)**

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
1	saveAsTextFile at tera_Sort.java:46	2016/03/29 16:57:33	11 min	0/2	117/1490

**Completed Jobs (1)**

Job Id	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
0	sortByKey at tera_Sort.java:39	2016/03/29 16:30:26	27 min	1/1	745/745

# CS553 Programming Assignment #2

## Sort on Hadoop/Spark

The screenshot shows a terminal window titled 'vishwanath@vishwanath-Q551LB: ~/Documents'. The command run is 'hduser@ip-172-31-49-141: /usr/local/spark/ec2\$ ./bin/spark-submit --class tera\_Sort /bdarl-0.0.1-SNAPSHOT.jar SparkInput Output'. The terminal output is as follows:

```
hduser@ip-172-31-49-141: /usr/local/spark/ec2$ ./bin/spark-submit --class tera_Sort /bdarl-0.0.1-SNAPSHOT.jar SparkInput Output
x vishwanath@vishwanath-Q551LB: ~/Documents
root@ip-172-31-34-88 spark$ ./bin/spark-submit --class tera_Sort /bdarl-0.0.1-SNAPSHOT.jar SparkInput Output
16/03/29 16:30:23 INFO spark.SparkContext: Running Spark version 1.6.1
16/03/29 16:30:23 WARN spark.SparkConf:
SPARK_WORKER_INSTANCES was detected (set to '1').
This is deprecated in Spark 1.6+.

Please instead use:
- ./spark-submit with --num-executors to specify the number of executors
- Or set SPARK_EXECUTOR_INSTANCES
- spark.executor.instances to configure the number of instances in the spark config.

16/03/29 16:30:23 INFO spark.SecurityManager: Changing view acls to: root
16/03/29 16:30:23 INFO spark.SecurityManager: Changing modify acls to: root
16/03/29 16:30:23 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users wth view permissions: Set(root); users with modify permissions: Set(root)
16/03/29 16:30:23 INFO util.Utils: Successfully started service 'sparkDriver' on port 53071.
16/03/29 16:30:24 INFO sif4j.Slf4jLogger: Slf4jLogger started
16/03/29 16:30:24 INFO Remoting: Starting remoting
16/03/29 16:30:24 INFO Slf4jLogger: Slf4jLogger started listening on addresses :[akka.tcp://sparkDriverActorSystem@172.31.34.88:53235]
16/03/29 16:30:24 INFO util.Utils: Successfully started service 'sparkDriverActorSystem' on port 53235.
16/03/29 16:30:24 INFO spark.SparkEnv: Registering MapOutputTracker
16/03/29 16:30:24 INFO spark.SparkEnv: Registering BlockManagerMaster
16/03/29 16:30:24 INFO storage.DiskBlockManager: Created local directory at /mnt/spark/blockmgr-0c266d56-026a-479c-9636-3e4a3238b2394
16/03/29 16:30:24 INFO storage.DiskBlockManager: Created local directory at /mnt2/spark/blockmgr-dsf035d6-7d30-4383-a0e3-3968cadfe74
16/03/29 16:30:24 INFO spark.SparkEnv: Registered executor 511.5 MB
16/03/29 16:30:24 INFO spark.SparkEnv: Registering OutputCommitCoordinator
16/03/29 16:30:24 INFO server.Server: jetty-8.y.z-SNAPSHOT
16/03/29 16:30:24 INFO server.AbstractConnector: Started SelectChannelConnector@0.0.0.0:4040
16/03/29 16:30:24 INFO util.Utils: Successfully started service 'SparkUI' on port 4040.
16/03/29 16:30:24 INFO util.Utils: Started SparkUI at http://ec2-54-85-192-104.compute-1.amazonaws.com:4040
16/03/29 16:30:25 INFO spark.HttpFileServer: File server directory is /mnt/spark/spark-3432d0fd-0e70-473e-86f5-64a25f6bd92b/httpd-ac10b85a-32d1-4bb4-a51e-591ce5153918
16/03/29 16:30:25 INFO spark.HttpServer: Starting HTTP Server
16/03/29 16:30:25 INFO server.Server: jetty-8.y.z-SNAPSHOT
16/03/29 16:30:25 INFO server.AbstractConnector: Started SocketConnector@0.0.0.0:4056
16/03/29 16:30:25 INFO util.Utils: Successfully started service 'HTTP file server' on port 4056.
16/03/29 16:30:25 INFO spark.PeerContext: Starting executors alive on host localhost
16/03/29 16:30:25 INFO util.Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 52042.
16/03/29 16:30:25 INFO storage.BlockManagerMasterEndpoint: Registering block manager localhost:52042 with 511.5 MB RAM, BlockManagerId(driver, localhost, 52042)
16/03/29 16:30:25 INFO storage.BlockManagerMaster: Trying to register BlockManager
16/03/29 16:30:25 INFO storage.BlockManagerMaster: Registered BlockManager
16/03/29 16:30:25 INFO storage.BroadCastBlockManager: Added broadcast 0 piece(s) stored as bytes in memory (estimated size 46.3 KB, free 46.3 KB)
16/03/29 16:30:25 INFO storage.BroadCastBlockManager: Added broadcast 0 piece(s) stored as bytes in memory (estimated size 4.4 KB, free 50.7 KB)
16/03/29 16:30:25 INFO storage.BroadCastBlockManager: Added broadcast 0 piece(s) stored as bytes in memory (estimated size 4.4 KB, free: 511.5 MB)
16/03/29 16:30:25 INFO spark.SparkContext: Created broadcast 0 from textfile at tera_Sort.java:23
16/03/29 16:30:26 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-Java classes where applicable
16/03/29 16:30:26 WARN snappyCompression: Snappy compression library not found
16/03/29 16:30:26 INFO spark.SparkContext: Total input paths to process: 1
16/03/29 16:30:26 INFO spark.SparkContext: Starting job: sortByKey at tera_Sort.java:39
16/03/29 16:30:26 INFO scheduler.DAGScheduler: Got job 0 (sortByKey at tera_Sort.java:39) with 745 output partitions
16/03/29 16:30:26 INFO scheduler.DAGScheduler: Final stage: ResultStage 0 (sortByKey at tera_Sort.java:39)
16/03/29 16:30:26 INFO scheduler.DAGScheduler: Parents of final stage: List()
16/03/29 16:30:26 INFO scheduler.DAGScheduler: Missing parents: List()
16/03/29 16:30:26 INFO scheduler.DAGScheduler: Submitting ResultStage 0 (MapPartitionsRDD[1] at sortByKey at tera_Sort.java:39), which has no missing parents
16/03/29 16:30:26 INFO scheduler.DAGScheduler: Submitting 1 missing tasks from ResultStage 0 (MapPartitionsRDD[1] at sortByKey at tera_Sort.java:39)
```