
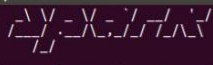


Time in seconds 2588.6 (splitting,sorting and splitting) to run 128 GB file (i3.large)

```
ubuntu@ip-172-31-25-221: ~  
gensort      hadoop      makefile     SharedMemory.class  spark-1.6.0-bin-hadoop2.6  
Gensort      input128.txt  scala        SharedMemory.java   spark-1.6.0-bin-hadoop2.6.tgz  
ubuntu@ip-172-31-25-221:~$ rm input2.txt  
ubuntu@ip-172-31-25-221:~$ rm input.txt  
ubuntu@ip-172-31-25-221:~$ vi scode.scala  
ubuntu@ip-172-31-25-221:~$ ./spark-1.6.0-bin-hadoop2.6/bin/spark-shell  
log4j:WARN No appenders could be found for logger (org.apache.hadoop.metrics2.lib.MutableMetricsFactory).  
log4j:WARN Please initialize the log4j system properly.  
log4j:WARN See http://logging.apache.org/log4j/1.2/faq.html#noconfig for more info.  
Using Spark's repl log4j profile: org/apache/spark/log4j-defaults-repl.properties  
To adjust logging level use sc.setLogLevel("INFO")  
Welcome to  
 version 1.6.0  
Using Scala version 2.10.5 (OpenJDK 64-Bit Server VM, Java 1.8.0_151)  
Type in expressions to have them evaluated.  
Type :help for more information.  
Spark context available as sc.  
17/12/02 06:38:09 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependencies)  
17/12/02 06:38:09 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependencies)  
17/12/02 06:38:14 WARN ObjectStore: Version information not found in metastore. hive.metastore.schema.verification is not enabled so recording the schema version 1.2.0  
17/12/02 06:38:14 WARN ObjectStore: Failed to get database default, returning NoSuchObjectException  
17/12/02 06:38:16 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependencies)  
17/12/02 06:38:16 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependencies)  
17/12/02 06:38:20 WARN ObjectStore: Version information not found in metastore. hive.metastore.schema.verification is not enabled so recording the schema version 1.2.0  
17/12/02 06:38:20 WARN ObjectStore: Failed to get database default, returning NoSuchObjectException  
SQL context available as sqlContext.  
scala> :load /home/ubuntu/scode.scala  
Loading /home/ubuntu/scode.scala...  
lines: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[1] at textFile at <console>:27  
t1: Long = 27005460953065  
sort: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[7] at map at <console>:29  
duration: Double = 2588.623730743  
2588.623730743  
scala>
```

Time in seconds 803.9 (splitting,sorting and splitting) to run 128 GB file (13.4x large)

```
ubuntu@ip-172-31-25-87: ~  
 version 1.6.0  
Using Scala version 2.10.5 (OpenJDK 64-Bit Server VM, Java 1.8.0_151)  
Type in expressions to have them evaluated.  
Type :help for more information.  
Spark context available as sc.  
17/12/02 07:32:42 WARN General: Plugin (Bundle) "org.datanucleus.api.jdo" is already registered. Ensure you dont have multiple JAR versions of the same plugin in the classpath. The URL "file:/home/ubuntu/spark-1.6.0-bin-hadoop2.6/lib/datanucleus-api-jdo-3.2.6.jar" is already registered, and you are trying to register an identical plugin located at URL "file:/home/ubuntu/spark/lib/datanucleus-api-jdo-3.2.6.jar."  
17/12/02 07:32:42 WARN General: Plugin (Bundle) "org.datanucleus" is already registered. Ensure you dont have multiple JAR versions of the same plugin in the classpath. The URL "file:/home/ubuntu/spark-1.6.0-bin-hadoop2.6/lib/datanucleus-core-3.2.10.jar" is already registered, and you are trying to register an identical plugin located at URL "file:/home/ubuntu/spark/lib/datanucleus-core-3.2.10.jar."  
17/12/02 07:32:42 WARN General: Plugin (Bundle) "org.datanucleus.store.rdbms" is already registered. Ensure you dont have multiple JAR versions of the same plugin in the classpath. The URL "file:/home/ubuntu/spark-1.6.0-bin-hadoop2.6/lib/datanucleus-rdbms-3.2.9.jar" is already registered, and you are trying to register an identical plugin located at URL "file:/home/ubuntu/spark-1.6.0-bin-hadoop2.6/lib/datanucleus-rdbms-3.2.9.jar."  
17/12/02 07:32:42 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependencies)  
17/12/02 07:32:42 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependencies)  
17/12/02 07:32:45 WARN ObjectStore: Version information not found in metastore. hive.metastore.schema.verification is not enabled so recording the schema version 1.2.0  
17/12/02 07:32:45 WARN ObjectStore: Failed to get database default, returning NoSuchObjectException  
17/12/02 07:32:46 WARN General: Plugin (Bundle) "org.datanucleus.store.rdbms" is already registered. Ensure you dont have multiple JAR versions of the same plugin in the classpath. The URL "file:/home/ubuntu/spark-1.6.0-bin-hadoop2.6/lib/datanucleus-rdbms-3.2.9.jar" is already registered, and you are trying to register an identical plugin located at URL "file:/home/ubuntu/spark-1.6.0-bin-hadoop2.6/lib/datanucleus-rdbms-3.2.9.jar."  
17/12/02 07:32:46 WARN General: Plugin (Bundle) "org.datanucleus.api.jdo" is already registered. Ensure you dont have multiple JAR versions of the same plugin in the classpath. The URL "file:/home/ubuntu/spark-1.6.0-bin-hadoop2.6/lib/datanucleus-api-jdo-3.2.6.jar" is already registered, and you are trying to register an identical plugin located at URL "file:/home/ubuntu/spark/lib/datanucleus-api-jdo-3.2.6.jar."  
17/12/02 07:32:46 WARN General: Plugin (Bundle) "org.datanucleus" is already registered. Ensure you dont have multiple JAR versions of the same plugin in the classpath. The URL "file:/home/ubuntu/spark-1.6.0-bin-hadoop2.6/lib/datanucleus-core-3.2.10.jar" is already registered, and you are trying to register an identical plugin located at URL "file:/home/ubuntu/spark/lib/datanucleus-core-3.2.10.jar."  
17/12/02 07:32:46 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependencies)  
17/12/02 07:32:46 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependencies)  
SQL context available as sqlContext.  
scala> :load /home/ubuntu/scode.scala  
Loading /home/ubuntu/scode.scala...  
lines: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[1] at textFile at <console>:27  
t1: Long = 28284321391908  
sort: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[7] at map at <console>:29  
duration: Double = 803.935325867  
803.935325867  
[Stage 1:] (0 + 16) / 4096
```

Time in seconds 6282.47 sec (splitting, sorting and splitting) to run 1 TB file (13.4x large)

```
ubuntu@ip-172-31-25-87:~$  
of the same plugin in the classpath. The URL "file:/home/ubuntu/spark/lib/datanucleus-rdbms-3.2.9.jar" is already registered, and you are trying  
to register an identical plugin located at URL "file:/home/ubuntu/spark-1.6.0-bin-hadoop2.6/lib/datanucleus-rdbms-3.2.9.jar."  
17/12/02 12:06:32 WARN General: Plugin (Bundle) "org.datanucleus.api.jdo" is already registered. Ensure you dont have multiple JAR versions of t  
he same plugin in the classpath. The URL "file:/home/ubuntu/spark-1.6.0-bin-hadoop2.6/lib/datanucleus-api-jdo-3.2.6.jar" is already registered,  
and you are trying to register an identical plugin located at URL "file:/home/ubuntu/spark/lib/datanucleus-api-jdo-3.2.6.jar."  
17/12/02 12:06:32 WARN General: Plugin (Bundle) "org.datanucleus" is already registered. Ensure you dont have multiple JAR versions of the same  
plugin in the classpath. The URL "file:/home/ubuntu/spark-1.6.0-bin-hadoop2.6/lib/datanucleus-core-3.2.10.jar" is already registered, and you ar  
e trying to register an identical plugin located at URL "file:/home/ubuntu/spark/lib/datanucleus-core-3.2.10.jar."  
17/12/02 12:06:32 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependencies)  
17/12/02 12:06:32 WARN Connection: BoneCP specified but not present in CLASSPATH (or one of dependencies)  
SQL context available as sqlContext.  
  
scala> :load /home/ubuntu/scode.scala  
Loading /home/ubuntu/scode.scala...  
lines: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[1] at textFile at <console>:27  
ti: Long = 44707685489156  
sort: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[7] at map at <console>:29  
duration: Double = 6282.476592991  
0282.476592991  
  
scala> :q  
Stopping spark context.  
ubuntu@ip-172-31-25-87:~$ exit  
logout  
Connection to ec2-52-206-51-120.compute-1.amazonaws.com closed.  
lastwalker@chelsea:~/Documents/SEMESTER-3/Cloud_Computing/A2/SPARK/spark/ec2$ ssh -i spark.pem ubuntu@ec2-52-206-51-120.compute-1.amazonaws.com  
Welcome to Ubuntu 16.04.3 LTS (GNU/Linux 4.4.0-1041-aws x86_64)  
  
 * Documentation:  https://help.ubuntu.com  
 * Management:    https://landscape.canonical.com  
 * Support:       https://ubuntu.com/advantage  
  
Get cloud support with Ubuntu Advantage Cloud Guest:  
http://www.ubuntu.com/business/services/cloud  
  
0 packages can be updated.  
0 updates are security updates.  
  
Last login: Sat Dec  2 12:03:21 2017 from 208.59.154.186  
ubuntu@ip-172-31-25-87:~$ ls -sh input128.txt  
1001G input128.txt  
ubuntu@ip-172-31-25-87:~$
```