miniproject 2

zherui cao(zhc61@pitt.edu )

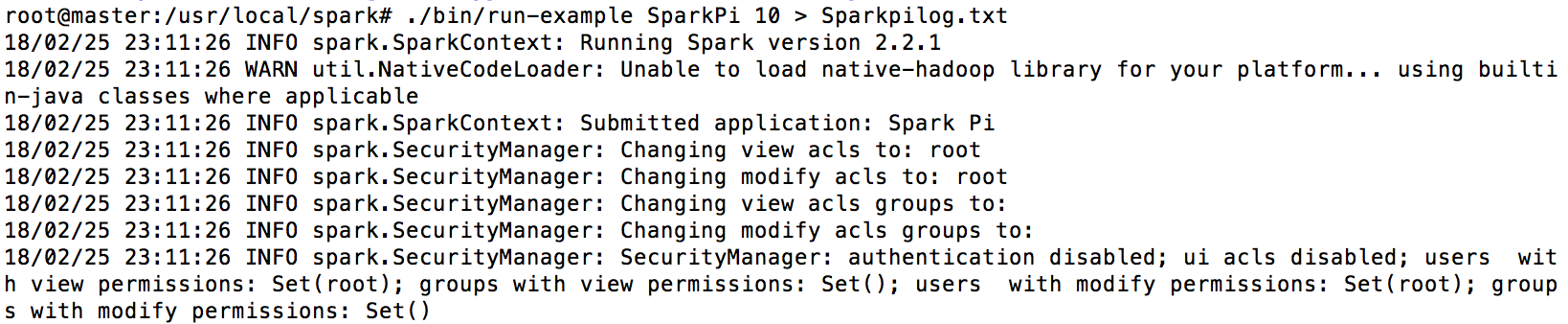
jiamin cheng ( jic122@pitt.edu)

Part 1: Setting up Spark

After downloading it, we modified the /usr/local/spark/conf/

running example time program

|  |
| --- |
| ./bin/run-example SparkPi 10 > Sparkpilog.txt |



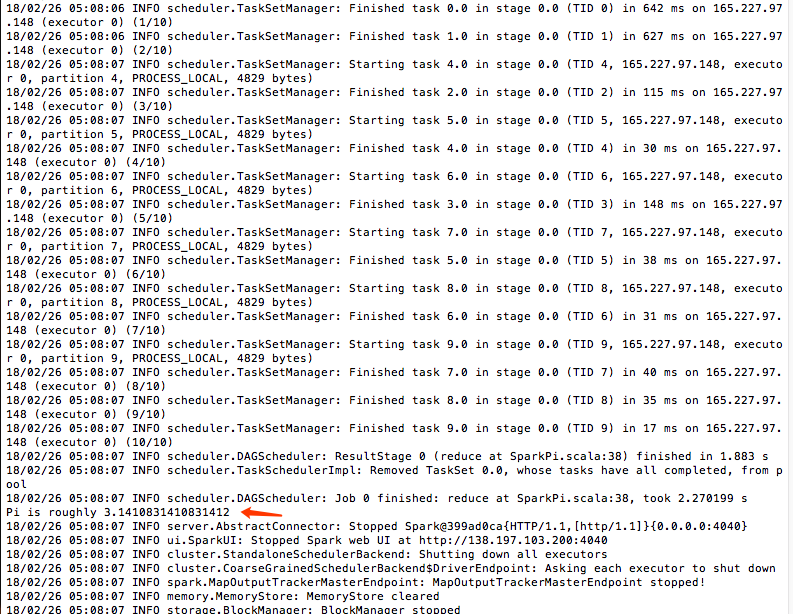
Then we can see the result



standalone mode

|  |
| --- |
| ./bin/spark-submit --class org.apache.spark.examples.SparkPi --master yarn --deploy-mode cluster --executor-memory 512M /usr/local/spark/examples/jars/spark-examples\_2.11-2.2.1.jar 10 |

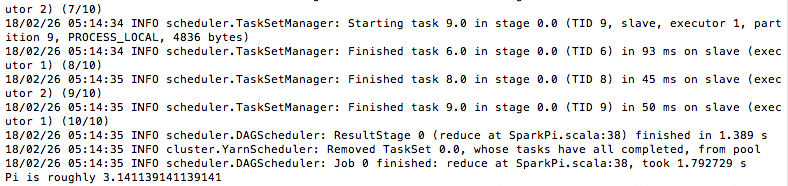
result:



yarn-client

|  |
| --- |
| ./bin/spark-submit --class org.apache.spark.examples.SparkPi --master yarn-client --executor-memory 512M /usr/local/spark/examples/jars/spark-examples\_2.11-2.2.1.jar 10 |

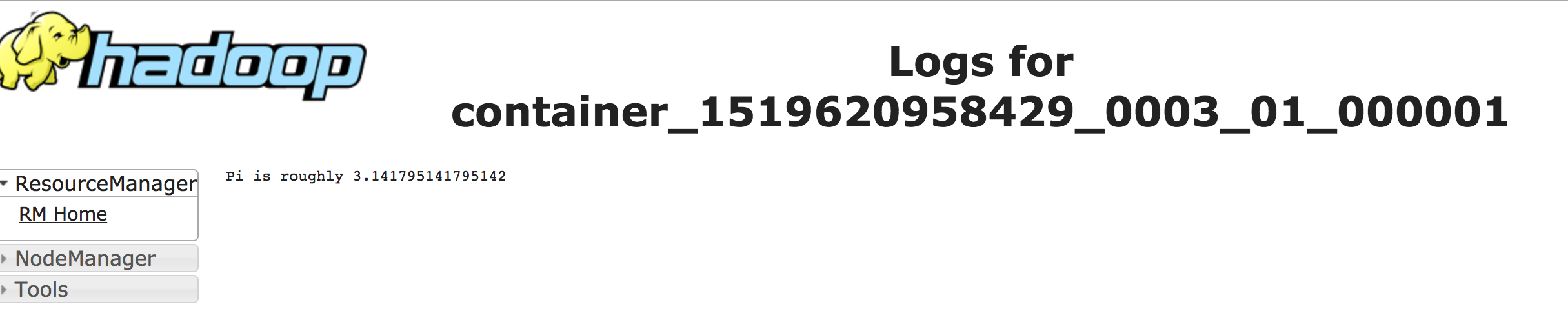
result:



yarn-cluster

|  |
| --- |
| ./bin/spark-submit --class org.apache.spark.examples.SparkPi --master yarn-cluster --executor-memory 512M /usr/local/spark/examples/jars/spark-examples\_2.11-2.2.1.jar 10 |

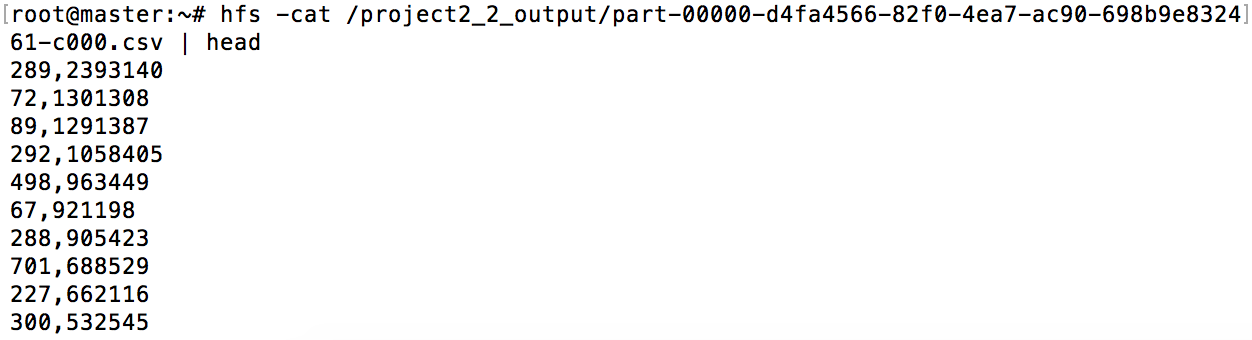
In this mode, the result will not be printed on the console. So we check it from the webUI.



Part 2 Developing Spark programs

|  |
| --- |
| var a=spark.read.format("csv").option("header","true").option("delimiter","\t").csv("hdfs://master:9000/project2\_2/user\_artist.dat");  a=a.withColumn("weight", col("weight").cast("Integer"))  var sum = a.groupBy("artistID").sum("weight")  var sum2 = sum.select(sum("artistID"),sum("sum(weight)").alias("total")) sort(desc("total"))  sum2=sum2.repartition(1)  sum2.write.format("csv").save("hdfs://master:9000/project2\_2\_output") |

and the result is



(we set alias so that the hfs is the short cut of hadoop fs)

Part 3

question 1 to 3

|  |
| --- |
| var textFile = sc.textFile("hdfs://master:9000/access\_log.txt").cache()  var counts = textFile.map( line => (line.split(']')(1).split('\"')(1).split(' ')(1).split('?')(0),1))  counts=counts.reduceByKey( \_ + \_ ).sortBy(\_.\_2,false).repartition(1)  counts.saveAsTextFile("hdfs://master:9000/project2\_part3\_url\_output") |

question 1



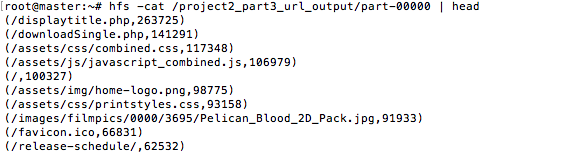
question 2



check the answers using wc



question 3





even though displaytitle.php shows 264106 times in this log file, there are many invalid request which we should ignores. So we only have 263725 valid requests.

for example:

|  |
| --- |
| 0.169.174.147 - - [07/Nov/2011:02:27:21 -0800] "GET /displaytitle.php%3fid=668/title.php?id=' HTTP/1.1" 404 231  10.118.250.30 - - [09/Nov/2011:14:16:30 -0800] "GET /displaytitle.phpid=3815 HTTP/1.1" 404 191  10.118.250.30 - - [09/Nov/2011:14:40:10 -0800] "GET /displaytitle.phpid=38Suspiria HTTP/1.1" 404 196 |

all the above are the invalid requests we ignored.

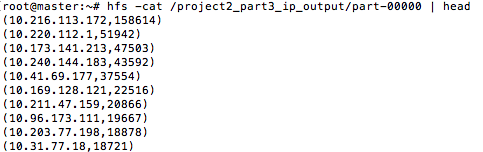
question 4

|  |
| --- |
| var textFile = sc.textFile("hdfs://master:9000/access\_log.txt").cache()  var counts=textFile.map(line=> (line.split (' ')(0),1))  .reduceByKey(\_+\_).sortBy(\_.\_2,false)  counts = counts.repartition(1)  counts.saveAsTextFile("hdfs://master:9000/project2\_part3\_ip\_output") |

then we type

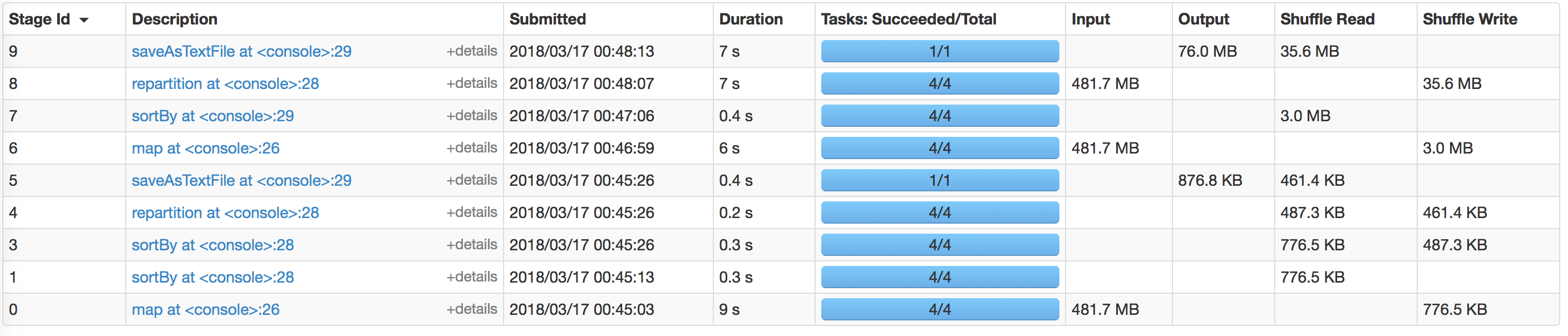
hfs -cat /project2\_3\_1\_output/part-00000 | head

to see the result



performance measurements:

using cached RDD in spark:



Using hadoop without cached RDD

Obviously,it will take much more time to finish.

