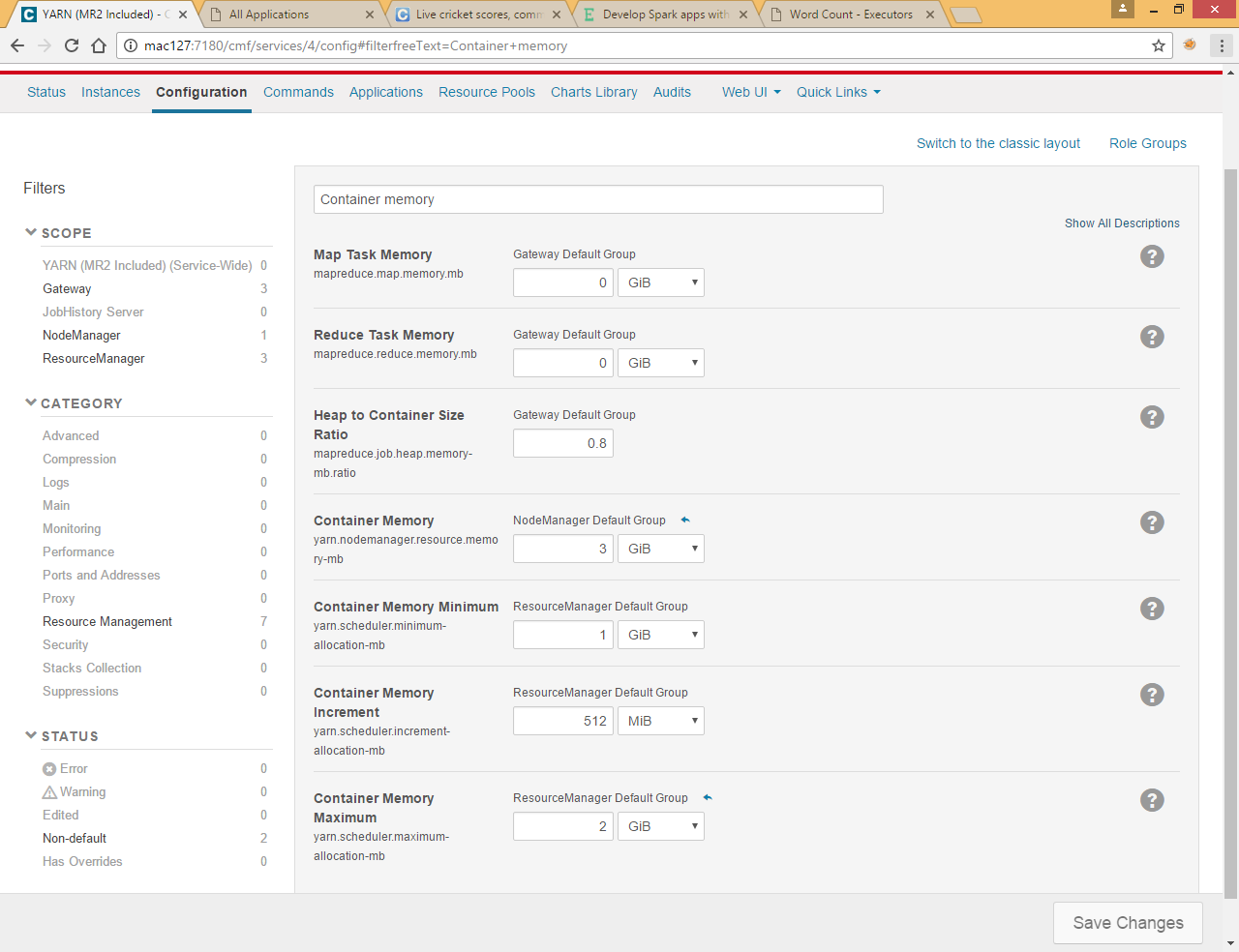
# Spark Configuration

**First install Spark on Cloudera cluster and run spark-shell from command prompt**

|  |
| --- |
| [root@mac127 ~]# spark-shell  Setting default log level to "WARN".  To adjust logging level use sc.setLogLevel(newLevel).  Welcome to  \_\_\_\_ \_\_  / \_\_/\_\_ \_\_\_ \_\_\_\_\_/ /\_\_  \_\ \/ \_ \/ \_ `/ \_\_/ '\_/  /\_\_\_/ .\_\_/\\_,\_/\_/ /\_/\\_\ version 1.6.0  /\_/  Using Scala version 2.10.5 (Java HotSpot(TM) 64-Bit Server VM, Java 1.7.0\_67)  Type in expressions to have them evaluated.  Type :help for more information.  17/01/13 16:25:54 ERROR spark.SparkContext: Error initializing SparkContext.  java.lang.IllegalArgumentException: Required executor memory (1024+384 MB) is above the max threshold (1024 MB) of this cluster! Please check the values of 'yarn.scheduler.maximum-allocation-mb' and/or 'yarn.nodemanager.resource.memory-mb'.  at org.apache.spark.deploy.yarn.Client.verifyClusterResources(Client.scala:284)  at org.apache.spark.deploy.yarn.Client.submitApplication(Client.scala:140)  at org.apache.spark.scheduler.cluster.YarnClientSchedulerBackend.start(YarnClientSchedulerBackend.scala:57)  at org.apache.spark.scheduler.TaskSchedulerImpl.start(TaskSchedulerImpl.scala:157)  at org.apache.spark.SparkContext.<init>(SparkContext.scala:542)  at org.apache.spark.repl.SparkILoop.createSparkContext(SparkILoop.scala:1022)  at $line3.$read$$iwC$$iwC.<init>(<console>:15)  at $line3.$read$$iwC.<init>(<console>:25)  at $line3.$read.<init>(<console>:27)  at $line3.$read$.<init>(<console>:31)  at $line3.$read$.<clinit>(<console>)  at $line3.$eval$.<init>(<console>:7)  at $line3.$eval$.<clinit>(<console>)  at $line3.$eval.$print(<console>)  at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)  at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)  at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)  at java.lang.reflect.Method.invoke(Method.java:606)  at org.apache.spark.repl.SparkIMain$ReadEvalPrint.call(SparkIMain.scala:1045)  at org.apache.spark.repl.SparkIMain$Request.loadAndRun(SparkIMain.scala:1326)  at org.apache.spark.repl.SparkIMain.loadAndRunReq$1(SparkIMain.scala:821)  at org.apache.spark.repl.SparkIMain.interpret(SparkIMain.scala:852)  at org.apache.spark.repl.SparkIMain.interpret(SparkIMain.scala:800)  at org.apache.spark.repl.SparkILoop.reallyInterpret$1(SparkILoop.scala:857)  at org.apache.spark.repl.SparkILoop.interpretStartingWith(SparkILoop.scala:902)  at org.apache.spark.repl.SparkILoop.command(SparkILoop.scala:814)  at org.apache.spark.repl.SparkILoopInit$$anonfun$initializeSpark$1.apply(SparkILoopInit.scala:125)  at org.apache.spark.repl.SparkILoopInit$$anonfun$initializeSpark$1.apply(SparkILoopInit.scala:124)  at org.apache.spark.repl.SparkIMain.beQuietDuring(SparkIMain.scala:305)  at org.apache.spark.repl.SparkILoopInit$class.initializeSpark(SparkILoopInit.scala:124)  at org.apache.spark.repl.SparkILoop.initializeSpark(SparkILoop.scala:64)  at org.apache.spark.repl.SparkILoop$$anonfun$org$apache$spark$repl$SparkILoop$$process$1$$anonfun$apply$mcZ$sp$5.apply$mcV$sp(SparkILoop.scala:974)  at org.apache.spark.repl.SparkILoopInit$class.runThunks(SparkILoopInit.scala:160)  at org.apache.spark.repl.SparkILoop.runThunks(SparkILoop.scala:64)  at org.apache.spark.repl.SparkILoopInit$class.postInitialization(SparkILoopInit.scala:108)  at org.apache.spark.repl.SparkILoop.postInitialization(SparkILoop.scala:64)  at org.apache.spark.repl.SparkILoop$$anonfun$org$apache$spark$repl$SparkILoop$$process$1.apply$mcZ$sp(SparkILoop.scala:991)  at org.apache.spark.repl.SparkILoop$$anonfun$org$apache$spark$repl$SparkILoop$$process$1.apply(SparkILoop.scala:945)  at org.apache.spark.repl.SparkILoop$$anonfun$org$apache$spark$repl$SparkILoop$$process$1.apply(SparkILoop.scala:945)  at scala.tools.nsc.util.ScalaClassLoader$.savingContextLoader(ScalaClassLoader.scala:135)  at org.apache.spark.repl.SparkILoop.org$apache$spark$repl$SparkILoop$$process(SparkILoop.scala:945)  at org.apache.spark.repl.SparkILoop.process(SparkILoop.scala:1064)  at org.apache.spark.repl.Main$.main(Main.scala:35)  at org.apache.spark.repl.Main.main(Main.scala)  at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)  at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)  at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)  at java.lang.reflect.Method.invoke(Method.java:606)  at org.apache.spark.deploy.SparkSubmit$.org$apache$spark$deploy$SparkSubmit$$runMain(SparkSubmit.scala:731)  at org.apache.spark.deploy.SparkSubmit$.doRunMain$1(SparkSubmit.scala:181)  at org.apache.spark.deploy.SparkSubmit$.submit(SparkSubmit.scala:206)  at org.apache.spark.deploy.SparkSubmit$.main(SparkSubmit.scala:121)  at org.apache.spark.deploy.SparkSubmit.main(SparkSubmit.scala)  17/01/13 16:25:54 ERROR util.Utils: Uncaught exception in thread main  java.lang.NullPointerException  at org.apache.spark.network.shuffle.ExternalShuffleClient.close(ExternalShuffleClient.java:152)  at org.apache.spark.storage.BlockManager.stop(BlockManager.scala:1231)  at org.apache.spark.SparkEnv.stop(SparkEnv.scala:96)  at org.apache.spark.SparkContext$$anonfun$stop$12.apply$mcV$sp(SparkContext.scala:1768)  at org.apache.spark.util.Utils$.tryLogNonFatalError(Utils.scala:1230)  at org.apache.spark.SparkContext.stop(SparkContext.scala:1767)  at org.apache.spark.SparkContext.<init>(SparkContext.scala:614)  at org.apache.spark.repl.SparkILoop.createSparkContext(SparkILoop.scala:1022)  at $line3.$read$$iwC$$iwC.<init>(<console>:15)  at $line3.$read$$iwC.<init>(<console>:25)  at $line3.$read.<init>(<console>:27)  at $line3.$read$.<init>(<console>:31)  at $line3.$read$.<clinit>(<console>)  at $line3.$eval$.<init>(<console>:7)  at $line3.$eval$.<clinit>(<console>)  at $line3.$eval.$print(<console>)  at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)  at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)  at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)  at java.lang.reflect.Method.invoke(Method.java:606)  at org.apache.spark.repl.SparkIMain$ReadEvalPrint.call(SparkIMain.scala:1045)  at org.apache.spark.repl.SparkIMain$Request.loadAndRun(SparkIMain.scala:1326)  at org.apache.spark.repl.SparkIMain.loadAndRunReq$1(SparkIMain.scala:821)  at org.apache.spark.repl.SparkIMain.interpret(SparkIMain.scala:852)  at org.apache.spark.repl.SparkIMain.interpret(SparkIMain.scala:800)  at org.apache.spark.repl.SparkILoop.reallyInterpret$1(SparkILoop.scala:857)  at org.apache.spark.repl.SparkILoop.interpretStartingWith(SparkILoop.scala:902)  at org.apache.spark.repl.SparkILoop.command(SparkILoop.scala:814)  at org.apache.spark.repl.SparkILoopInit$$anonfun$initializeSpark$1.apply(SparkILoopInit.scala:125)  at org.apache.spark.repl.SparkILoopInit$$anonfun$initializeSpark$1.apply(SparkILoopInit.scala:124)  at org.apache.spark.repl.SparkIMain.beQuietDuring(SparkIMain.scala:305)  at org.apache.spark.repl.SparkILoopInit$class.initializeSpark(SparkILoopInit.scala:124)  at org.apache.spark.repl.SparkILoop.initializeSpark(SparkILoop.scala:64)  at org.apache.spark.repl.SparkILoop$$anonfun$org$apache$spark$repl$SparkILoop$$process$1$$anonfun$apply$mcZ$sp$5.apply$mcV$sp(SparkILoop.scala:974)  at org.apache.spark.repl.SparkILoopInit$class.runThunks(SparkILoopInit.scala:160)  at org.apache.spark.repl.SparkILoop.runThunks(SparkILoop.scala:64)  at org.apache.spark.repl.SparkILoopInit$class.postInitialization(SparkILoopInit.scala:108)  at org.apache.spark.repl.SparkILoop.postInitialization(SparkILoop.scala:64)  at org.apache.spark.repl.SparkILoop$$anonfun$org$apache$spark$repl$SparkILoop$$process$1.apply$mcZ$sp(SparkILoop.scala:991)  at org.apache.spark.repl.SparkILoop$$anonfun$org$apache$spark$repl$SparkILoop$$process$1.apply(SparkILoop.scala:945)  at org.apache.spark.repl.SparkILoop$$anonfun$org$apache$spark$repl$SparkILoop$$process$1.apply(SparkILoop.scala:945)  at scala.tools.nsc.util.ScalaClassLoader$.savingContextLoader(ScalaClassLoader.scala:135)  at org.apache.spark.repl.SparkILoop.org$apache$spark$repl$SparkILoop$$process(SparkILoop.scala:945)  at org.apache.spark.repl.SparkILoop.process(SparkILoop.scala:1064)  at org.apache.spark.repl.Main$.main(Main.scala:35)  at org.apache.spark.repl.Main.main(Main.scala)  at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)  at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)  at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)  at java.lang.reflect.Method.invoke(Method.java:606)  at org.apache.spark.deploy.SparkSubmit$.org$apache$spark$deploy$SparkSubmit$$runMain(SparkSubmit.scala:731)  at org.apache.spark.deploy.SparkSubmit$.doRunMain$1(SparkSubmit.scala:181)  at org.apache.spark.deploy.SparkSubmit$.submit(SparkSubmit.scala:206)  at org.apache.spark.deploy.SparkSubmit$.main(SparkSubmit.scala:121)  at org.apache.spark.deploy.SparkSubmit.main(SparkSubmit.scala)  java.lang.IllegalArgumentException: Required executor memory (1024+384 MB) is above the max threshold (1024 MB) of this cluster! Please check the values of 'yarn.scheduler.maximum-allocation-mb' and/or 'yarn.nodemanager.resource.memory-mb'.  at org.apache.spark.deploy.yarn.Client.verifyClusterResources(Client.scala:284)  at org.apache.spark.deploy.yarn.Client.submitApplication(Client.scala:140)  at org.apache.spark.scheduler.cluster.YarnClientSchedulerBackend.start(YarnClientSchedulerBackend.scala:57)  at org.apache.spark.scheduler.TaskSchedulerImpl.start(TaskSchedulerImpl.scala:157)  at org.apache.spark.SparkContext.<init>(SparkContext.scala:542)  at org.apache.spark.repl.SparkILoop.createSparkContext(SparkILoop.scala:1022)  at $iwC$$iwC.<init>(<console>:15)  at $iwC.<init>(<console>:25)  at <init>(<console>:27)  at .<init>(<console>:31)  at .<clinit>(<console>)  at .<init>(<console>:7)  at .<clinit>(<console>)  at $print(<console>)  at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)  at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)  at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)  at java.lang.reflect.Method.invoke(Method.java:606)  at org.apache.spark.repl.SparkIMain$ReadEvalPrint.call(SparkIMain.scala:1045)  at org.apache.spark.repl.SparkIMain$Request.loadAndRun(SparkIMain.scala:1326)  at org.apache.spark.repl.SparkIMain.loadAndRunReq$1(SparkIMain.scala:821)  at org.apache.spark.repl.SparkIMain.interpret(SparkIMain.scala:852)  at org.apache.spark.repl.SparkIMain.interpret(SparkIMain.scala:800)  at org.apache.spark.repl.SparkILoop.reallyInterpret$1(SparkILoop.scala:857)  at org.apache.spark.repl.SparkILoop.interpretStartingWith(SparkILoop.scala:902)  at org.apache.spark.repl.SparkILoop.command(SparkILoop.scala:814)  at org.apache.spark.repl.SparkILoopInit$$anonfun$initializeSpark$1.apply(SparkILoopInit.scala:125)  at org.apache.spark.repl.SparkILoopInit$$anonfun$initializeSpark$1.apply(SparkILoopInit.scala:124)  at org.apache.spark.repl.SparkIMain.beQuietDuring(SparkIMain.scala:305)  at org.apache.spark.repl.SparkILoopInit$class.initializeSpark(SparkILoopInit.scala:124)  at org.apache.spark.repl.SparkILoop.initializeSpark(SparkILoop.scala:64)  at org.apache.spark.repl.SparkILoop$$anonfun$org$apache$spark$repl$SparkILoop$$process$1$$anonfun$apply$mcZ$sp$5.apply$mcV$sp(SparkILoop.scala:974)  at org.apache.spark.repl.SparkILoopInit$class.runThunks(SparkILoopInit.scala:160)  at org.apache.spark.repl.SparkILoop.runThunks(SparkILoop.scala:64)  at org.apache.spark.repl.SparkILoopInit$class.postInitialization(SparkILoopInit.scala:108)  at org.apache.spark.repl.SparkILoop.postInitialization(SparkILoop.scala:64)  at org.apache.spark.repl.SparkILoop$$anonfun$org$apache$spark$repl$SparkILoop$$process$1.apply$mcZ$sp(SparkILoop.scala:991)  at org.apache.spark.repl.SparkILoop$$anonfun$org$apache$spark$repl$SparkILoop$$process$1.apply(SparkILoop.scala:945)  at org.apache.spark.repl.SparkILoop$$anonfun$org$apache$spark$repl$SparkILoop$$process$1.apply(SparkILoop.scala:945)  at scala.tools.nsc.util.ScalaClassLoader$.savingContextLoader(ScalaClassLoader.scala:135)  at org.apache.spark.repl.SparkILoop.org$apache$spark$repl$SparkILoop$$process(SparkILoop.scala:945)  at org.apache.spark.repl.SparkILoop.process(SparkILoop.scala:1064)  at org.apache.spark.repl.Main$.main(Main.scala:35)  at org.apache.spark.repl.Main.main(Main.scala)  at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)  at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)  at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)  at java.lang.reflect.Method.invoke(Method.java:606)  at org.apache.spark.deploy.SparkSubmit$.org$apache$spark$deploy$SparkSubmit$$runMain(SparkSubmit.scala:731)  at org.apache.spark.deploy.SparkSubmit$.doRunMain$1(SparkSubmit.scala:181)  at org.apache.spark.deploy.SparkSubmit$.submit(SparkSubmit.scala:206)  at org.apache.spark.deploy.SparkSubmit$.main(SparkSubmit.scala:121)  at org.apache.spark.deploy.SparkSubmit.main(SparkSubmit.scala)  java.lang.NullPointerException  at org.apache.spark.sql.SQLContext$.createListenerAndUI(SQLContext.scala:1375)  at org.apache.spark.sql.hive.HiveContext.<init>(HiveContext.scala:101)  at sun.reflect.NativeConstructorAccessorImpl.newInstance0(Native Method)  at sun.reflect.NativeConstructorAccessorImpl.newInstance(NativeConstructorAccessorImpl.java:57)  at sun.reflect.DelegatingConstructorAccessorImpl.newInstance(DelegatingConstructorAccessorImpl.java:45)  at java.lang.reflect.Constructor.newInstance(Constructor.java:526)  at org.apache.spark.repl.SparkILoop.createSQLContext(SparkILoop.scala:1033)  at $iwC$$iwC.<init>(<console>:15)  at $iwC.<init>(<console>:24)  at <init>(<console>:26)  at .<init>(<console>:30)  at .<clinit>(<console>)  at .<init>(<console>:7)  at .<clinit>(<console>)  at $print(<console>)  at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)  at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)  at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)  at java.lang.reflect.Method.invoke(Method.java:606)  at org.apache.spark.repl.SparkIMain$ReadEvalPrint.call(SparkIMain.scala:1045)  at org.apache.spark.repl.SparkIMain$Request.loadAndRun(SparkIMain.scala:1326)  at org.apache.spark.repl.SparkIMain.loadAndRunReq$1(SparkIMain.scala:821)  at org.apache.spark.repl.SparkIMain.interpret(SparkIMain.scala:852)  at org.apache.spark.repl.SparkIMain.interpret(SparkIMain.scala:800)  at org.apache.spark.repl.SparkILoop.reallyInterpret$1(SparkILoop.scala:857)  at org.apache.spark.repl.SparkILoop.interpretStartingWith(SparkILoop.scala:902)  at org.apache.spark.repl.SparkILoop.command(SparkILoop.scala:814)  at org.apache.spark.repl.SparkILoopInit$$anonfun$initializeSpark$1.apply(SparkILoopInit.scala:133)  at org.apache.spark.repl.SparkILoopInit$$anonfun$initializeSpark$1.apply(SparkILoopInit.scala:124)  at org.apache.spark.repl.SparkIMain.beQuietDuring(SparkIMain.scala:305)  at org.apache.spark.repl.SparkILoopInit$class.initializeSpark(SparkILoopInit.scala:124)  at org.apache.spark.repl.SparkILoop.initializeSpark(SparkILoop.scala:64)  at org.apache.spark.repl.SparkILoop$$anonfun$org$apache$spark$repl$SparkILoop$$process$1$$anonfun$apply$mcZ$sp$5.apply$mcV$sp(SparkILoop.scala:974)  at org.apache.spark.repl.SparkILoopInit$class.runThunks(SparkILoopInit.scala:160)  at org.apache.spark.repl.SparkILoop.runThunks(SparkILoop.scala:64)  at org.apache.spark.repl.SparkILoopInit$class.postInitialization(SparkILoopInit.scala:108)  at org.apache.spark.repl.SparkILoop.postInitialization(SparkILoop.scala:64)  at org.apache.spark.repl.SparkILoop$$anonfun$org$apache$spark$repl$SparkILoop$$process$1.apply$mcZ$sp(SparkILoop.scala:991)  at org.apache.spark.repl.SparkILoop$$anonfun$org$apache$spark$repl$SparkILoop$$process$1.apply(SparkILoop.scala:945)  at org.apache.spark.repl.SparkILoop$$anonfun$org$apache$spark$repl$SparkILoop$$process$1.apply(SparkILoop.scala:945)  at scala.tools.nsc.util.ScalaClassLoader$.savingContextLoader(ScalaClassLoader.scala:135)  at org.apache.spark.repl.SparkILoop.org$apache$spark$repl$SparkILoop$$process(SparkILoop.scala:945)  at org.apache.spark.repl.SparkILoop.process(SparkILoop.scala:1064)  at org.apache.spark.repl.Main$.main(Main.scala:35)  at org.apache.spark.repl.Main.main(Main.scala)  at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)  at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)  at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)  at java.lang.reflect.Method.invoke(Method.java:606)  at org.apache.spark.deploy.SparkSubmit$.org$apache$spark$deploy$SparkSubmit$$runMain(SparkSubmit.scala:731)  at org.apache.spark.deploy.SparkSubmit$.doRunMain$1(SparkSubmit.scala:181)  at org.apache.spark.deploy.SparkSubmit$.submit(SparkSubmit.scala:206)  at org.apache.spark.deploy.SparkSubmit$.main(SparkSubmit.scala:121)  at org.apache.spark.deploy.SparkSubmit.main(SparkSubmit.scala)  <console>:16: error: not found: value sqlContext  import sqlContext.implicits.\_  ^  <console>:16: error: not found: value sqlContext  import sqlContext.sql  ^  scala> sc  <console>:20: error: not found: value sc  sc  ^ |

To resolve this issue Change container memory setting in Cloudera (YARN>>Configuration) and restart stale services



Increase below parameters

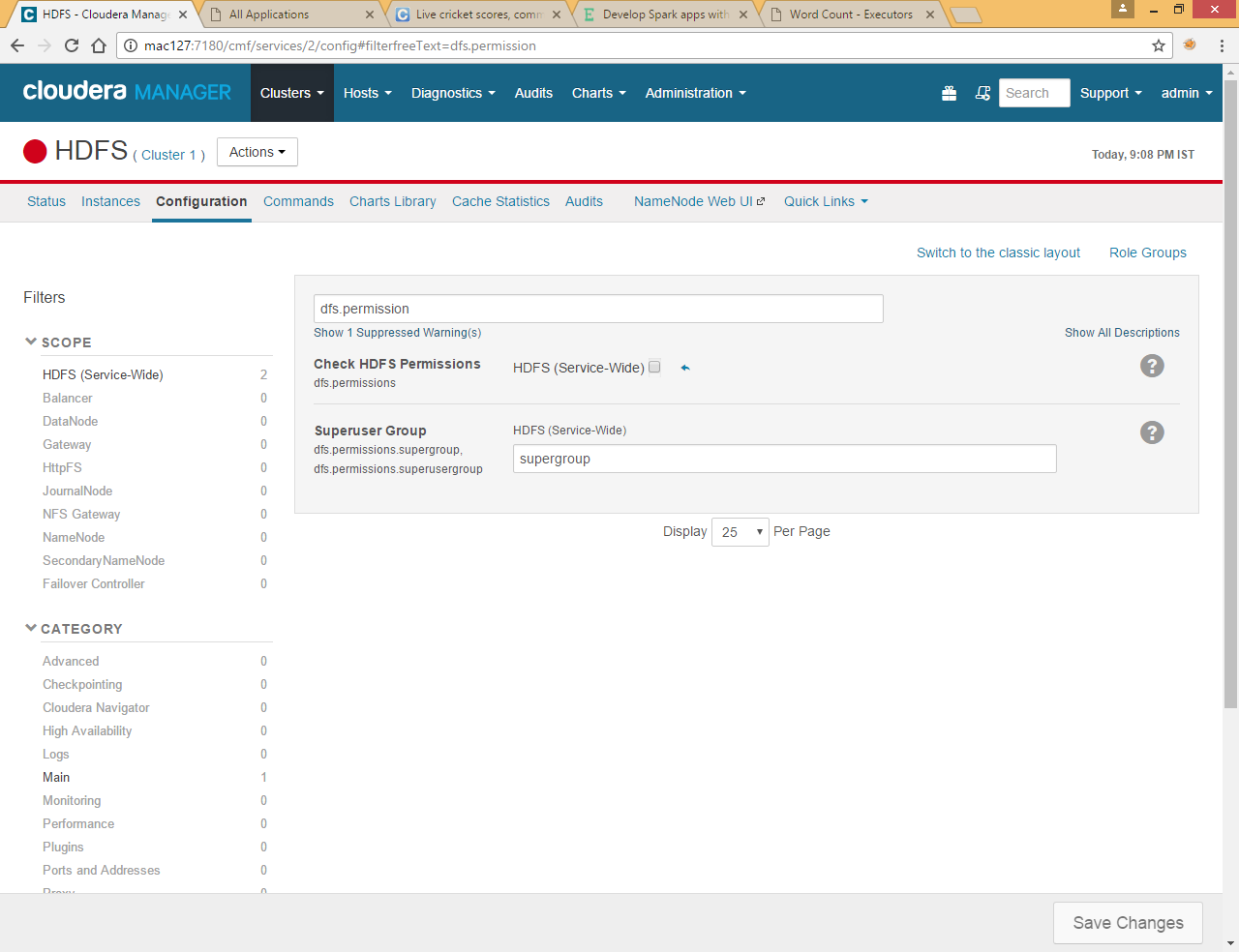
'yarn.scheduler.maximum-allocation-mb' = 2 GB and

'yarn.nodemanager.resource.memory-mb'. = 3 GB

Now create /user/test/input directory in hadoop cluster

|  |
| --- |
| [root@mac127 ~]# hdfs dfs -mkdir -p /user/test/input  mkdir: Permission denied: user=root, access=WRITE, inode="/user":hdfs:supergroup:drwxr-xr-x |

To resolve this issue go to HDFS >> configuration under Cloudera and uncheck dfs.permission to make it false



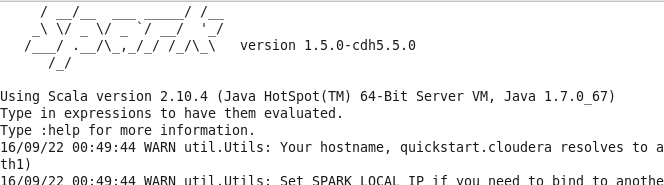
Now again create directory

Put any file in input directory from local file system that we use for word counting in further program

First check spark and scala version on which spark build upon

Once you start spark-shell you will get this info

Now check scala and java version in spark-shell



Or type sc.version in spark-shell

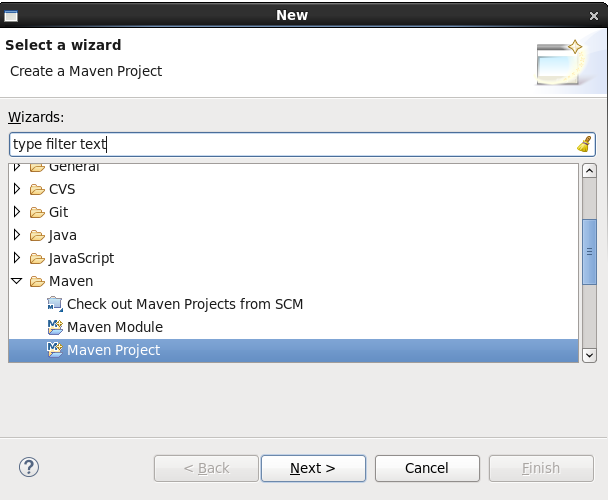


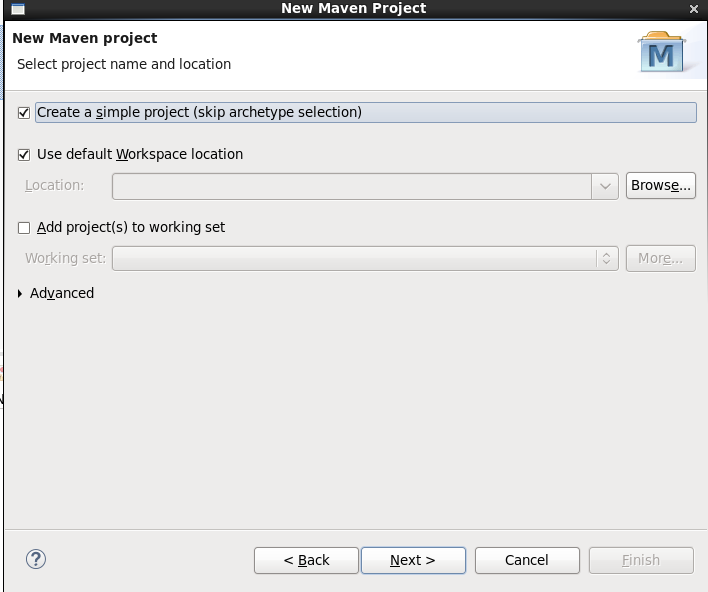
Note : You can find most of the jars under /usr/jars folder which are linked in other folder too like /usr/lib/spark/lib but we are going to use Maven to download the dependencies jars

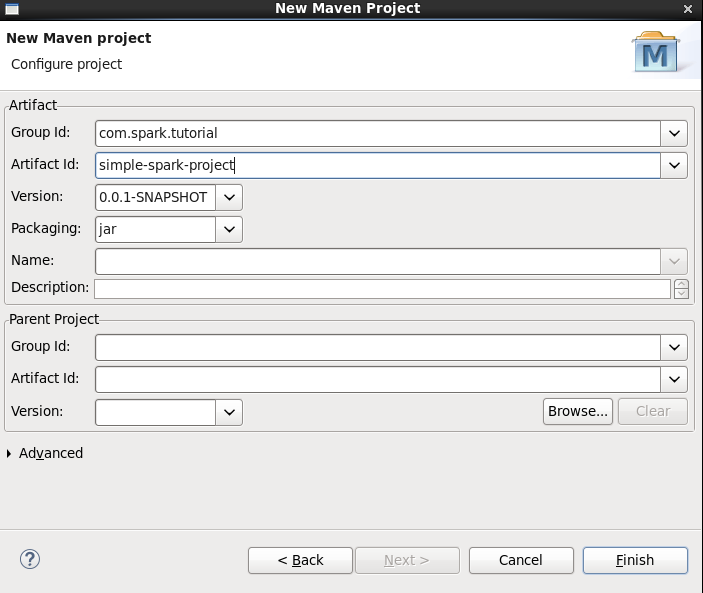
# Create Scala Project using Maven

Now we are going to create scala project in windows environment using maven

Select new Maven Project





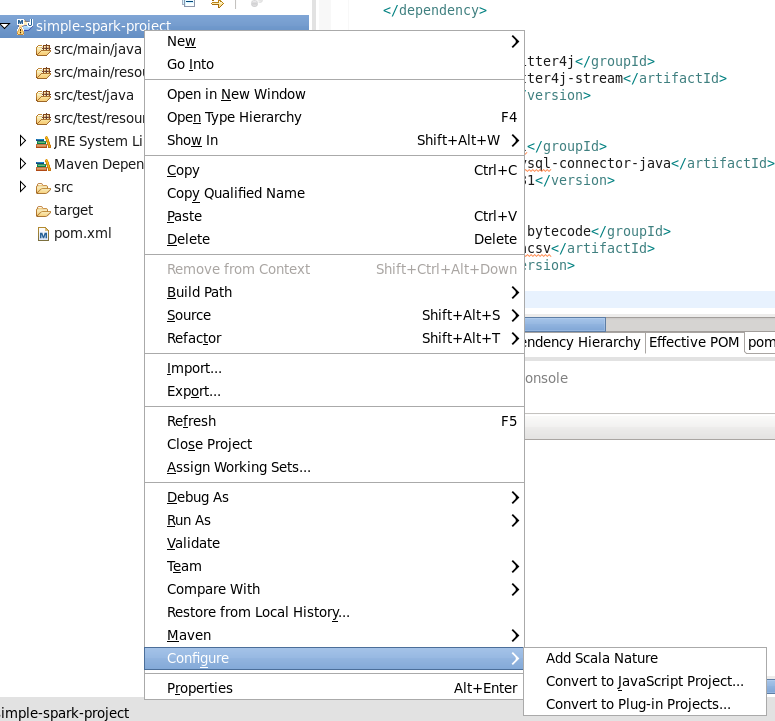


Specify project name and group id

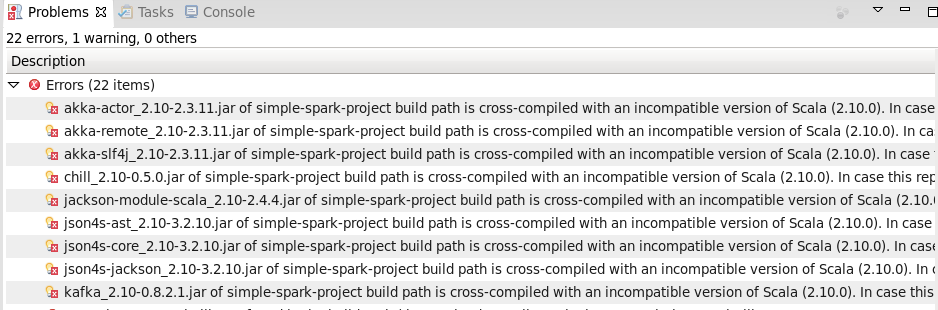
Now update the pom.xml as shown below and update project

|  |
| --- |
| <project xmlns="http://maven.apache.org/POM/4.0.0" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://maven.apache.org/POM/4.0.0 http://maven.apache.org/xsd/maven-4.0.0.xsd">  <modelVersion>4.0.0</modelVersion>  <groupId>com.spark.tutorial</groupId>  <artifactId>simple-spark-project</artifactId>  <version>0.0.1-SNAPSHOT</version>  <dependencies>  <dependency>  <groupId>org.apache.spark</groupId>  <artifactId>spark-core\_2.10</artifactId>  <version>1.5.0</version>  </dependency>  </dependencies>  </project> |

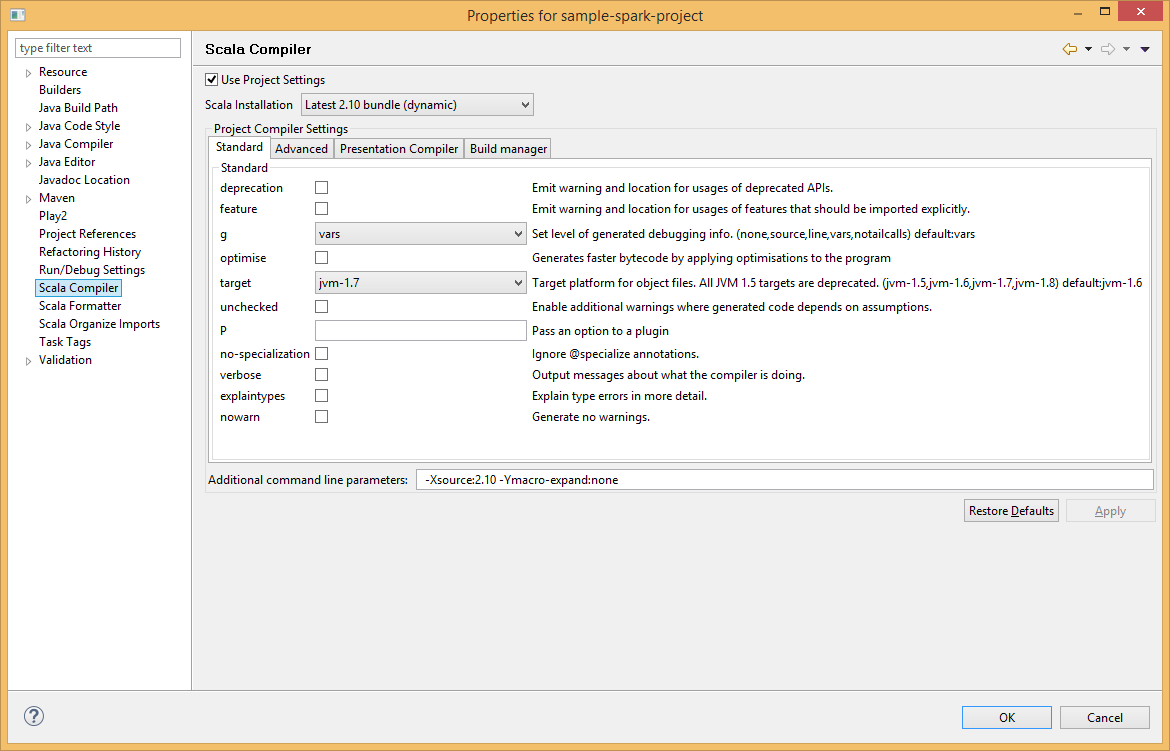
Right click on project and click on Configure>> Add scala nature



You will get below errors after adding Add Scala Nature

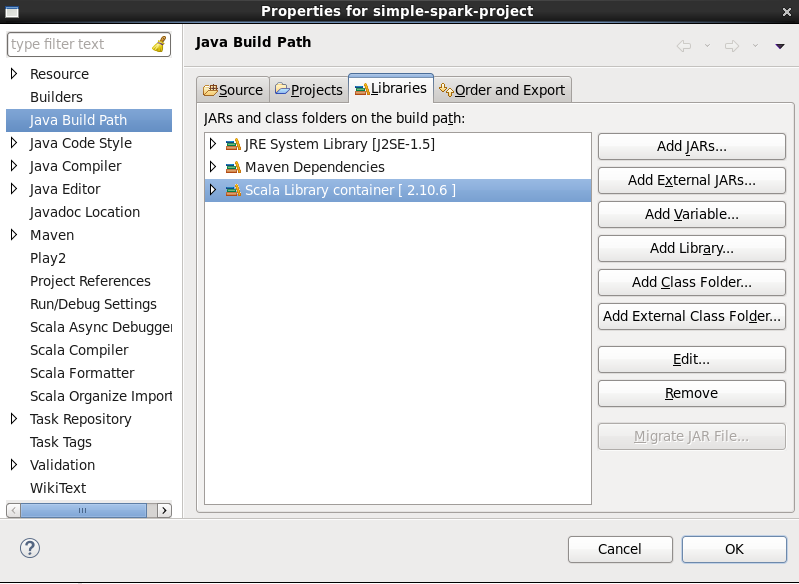


Right click on project >>Properties>>Scala Compiler

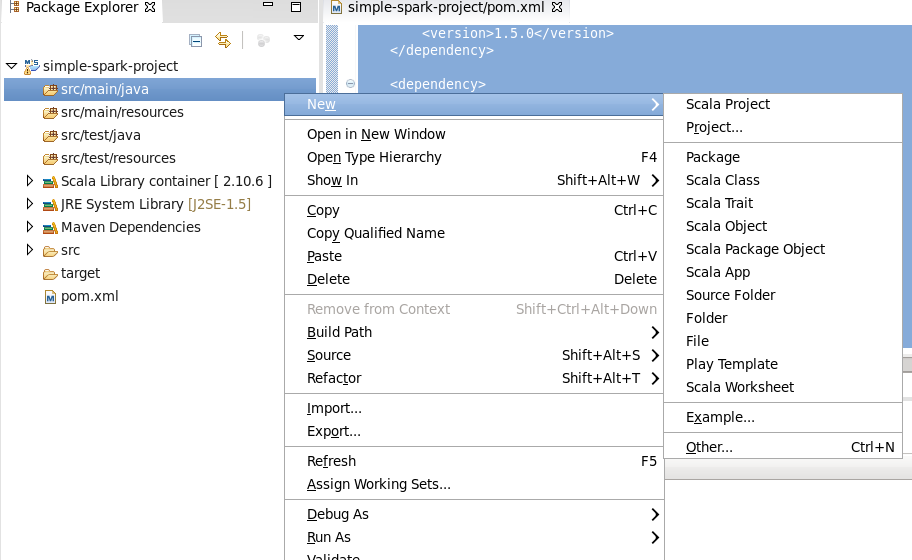


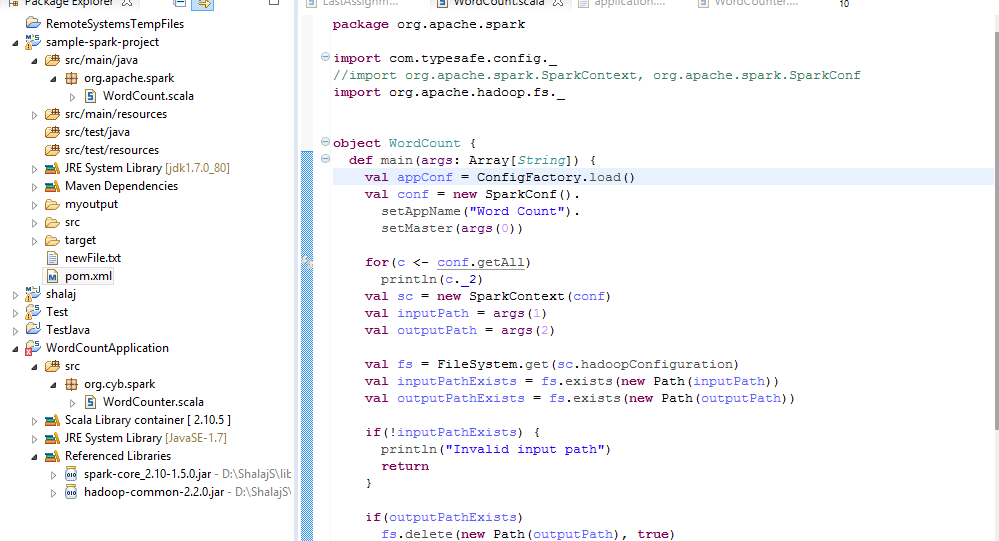
And choose 2.10 version of scala as our spark version build on 2.10 version

Remove the scala libarary from project as these are coming from spark configured under maven



Now create new Scala object



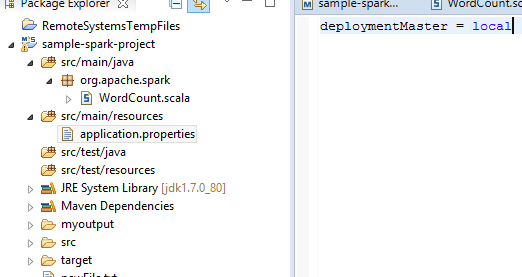


|  |
| --- |
| **package** org.apache.spark  **import** com.typesafe.config.\_  //import org.apache.spark.SparkContext, org.apache.spark.SparkConf  **import** org.apache.hadoop.fs.\_  **object** WordCount {  **def** main(args: Array[*String*]) {  **val** appConf = ConfigFactory.load()  **val** conf = **new** SparkConf().  setAppName("Word Count").  setMaster(args(0))    **for**(c <- conf.getAll)  println(c.\_2)  **val** sc = **new** SparkContext(conf)  **val** inputPath = args(1)  **val** outputPath = args(2)    **val** fs = FileSystem.get(sc.hadoopConfiguration)  **val** inputPathExists = fs.exists(**new** Path(inputPath))  **val** outputPathExists = fs.exists(**new** Path(outputPath))    **if**(!inputPathExists) {  println("Invalid input path")  **return**  }    **if**(outputPathExists)  fs.delete(**new** Path(outputPath), **true**)    **val** wc = sc.textFile(inputPath).  flatMap(rec => rec.split(" ")).  map(rec => (rec, 1)).  reduceByKey((acc, value) => acc + value)    wc.saveAsTextFile(outputPath)  }  } |

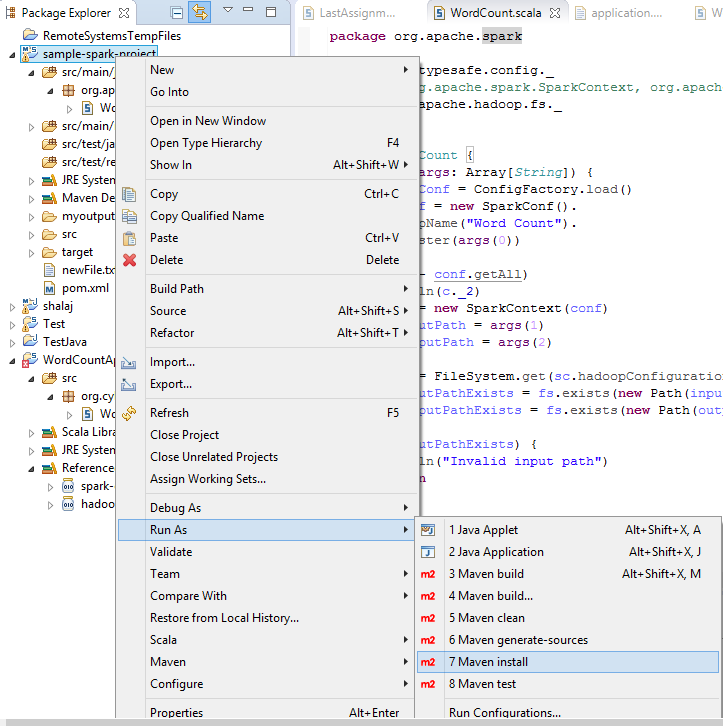
This program required three parameters developer-mode (yarn-client or local), input file and output path

Here I am not using configuration factory but we can use it by putting application.properties file under /src/main/resources folder in eclipse and define key value parameter in that file and get the value of that parameter using getString method like below

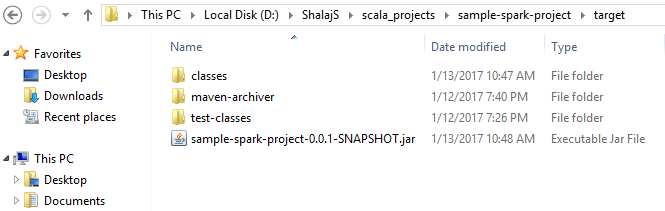
appConf.getString("deploymentMaster")



Now right click on project and select maven install to create jar file



Now go to target folder under project and find out jar file



Copy this jar file into your Hadoop cluster

Here I have created spark-test folder under /root and copy this jar file over there

Now go to spark-test folder

And run below command

|  |
| --- |
| spark-submit --master yarn --executor-memory 512m --total-executor-cores 1 --class org.apache.spark.WordCount sample-spark-project-0.0.1-SNAPSHOT.jar yarn-client /user/test/input /user/test/output |

Note:-To run the same program on local mode you need to pass local in place of yarn-client and remove --master yarn as by default it is local or you can also use --master local

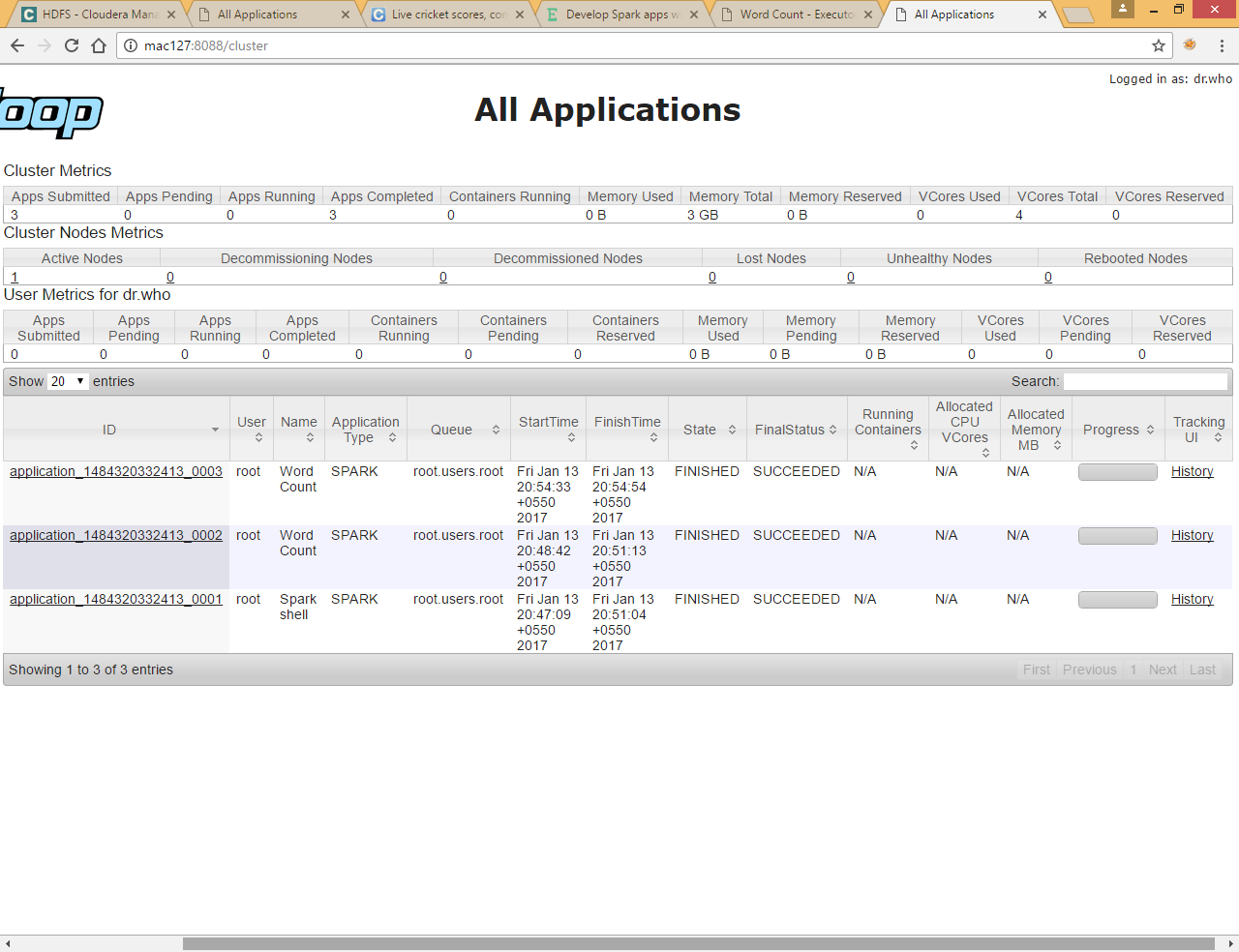
|  |
| --- |
| [root@mac127 ~]# cd spark-test  [root@mac127 spark-test]# spark-submit --master yarn --executor-memory 512m --total-executor-cores 1 --class org.apache.spark.WordCount sample-spark-project-0.0.1-SNAPSHOT.jar yarn-client /user/test/input /user/test/output  true  512m  local:/opt/cloudera/parcels/CDH-5.9.0-1.cdh5.9.0.p0.23/lib/spark/lib/spark-assembly.jar  true  hdfs://mac127:8020/user/spark/applicationHistory  yarn-client  /opt/cloudera/parcels/CDH-5.9.0-1.cdh5.9.0.p0.23/lib/hadoop/lib/native  1  7337  /opt/cloudera/parcels/CDH-5.9.0-1.cdh5.9.0.p0.23/lib/hadoop/lib/native  true  Word Count  60  /opt/cloudera/parcels  org.apache.spark.serializer.KryoSerializer  http://mac127:18088  false  file:/root/spark-test/sample-spark-project-0.0.1-SNAPSHOT.jar  {{HADOOP\_COMMON\_HOME}}/../../..  0  client  true  /opt/cloudera/parcels/CDH-5.9.0-1.cdh5.9.0.p0.23/lib/hadoop/lib/native  17/01/13 20:54:27 INFO spark.SparkContext: Running Spark version 1.6.0  17/01/13 20:54:28 INFO spark.SecurityManager: Changing view acls to: root  17/01/13 20:54:28 INFO spark.SecurityManager: Changing modify acls to: root  17/01/13 20:54:28 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(root); users with modify permissions: Set(root)  17/01/13 20:54:29 INFO util.Utils: Successfully started service 'sparkDriver' on port 42544.  17/01/13 20:54:29 INFO slf4j.Slf4jLogger: Slf4jLogger started  17/01/13 20:54:29 INFO Remoting: Starting remoting  17/01/13 20:54:30 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriverActorSystem@172.27.155.127:58067]  17/01/13 20:54:30 INFO Remoting: Remoting now listens on addresses: [akka.tcp://sparkDriverActorSystem@172.27.155.127:58067]  17/01/13 20:54:30 INFO util.Utils: Successfully started service 'sparkDriverActorSystem' on port 58067.  17/01/13 20:54:30 INFO spark.SparkEnv: Registering MapOutputTracker  17/01/13 20:54:30 INFO spark.SparkEnv: Registering BlockManagerMaster  17/01/13 20:54:30 INFO storage.DiskBlockManager: Created local directory at /tmp/blockmgr-461a3e98-f22e-40e0-b552-59974d56166a  17/01/13 20:54:30 INFO storage.MemoryStore: MemoryStore started with capacity 530.3 MB  17/01/13 20:54:30 INFO spark.SparkEnv: Registering OutputCommitCoordinator  17/01/13 20:54:30 INFO util.Utils: Successfully started service 'SparkUI' on port 4040.  17/01/13 20:54:30 INFO ui.SparkUI: Started SparkUI at http://172.27.155.127:4040  17/01/13 20:54:30 INFO spark.SparkContext: Added JAR file:/root/spark-test/sample-spark-project-0.0.1-SNAPSHOT.jar at spark://172.27.155.127:42544/jars/sample-spark-project-0.0.1-SNAPSHOT.jar with timestamp 1484321070855  17/01/13 20:54:31 INFO client.RMProxy: Connecting to ResourceManager at mac127/172.27.155.127:8032  17/01/13 20:54:31 INFO yarn.Client: Requesting a new application from cluster with 1 NodeManagers  17/01/13 20:54:31 INFO yarn.Client: Verifying our application has not requested more than the maximum memory capability of the cluster (2048 MB per container)  17/01/13 20:54:31 INFO yarn.Client: Will allocate AM container, with 896 MB memory including 384 MB overhead  17/01/13 20:54:31 INFO yarn.Client: Setting up container launch context for our AM  17/01/13 20:54:31 INFO yarn.Client: Setting up the launch environment for our AM container  17/01/13 20:54:31 INFO yarn.Client: Preparing resources for our AM container  17/01/13 20:54:32 INFO yarn.Client: Uploading resource file:/tmp/spark-aaf0f98c-1a59-4c2e-a208-ea8b49480012/\_\_spark\_conf\_\_5522928782895023891.zip -> hdfs://mac127:8020/user/root/.sparkStaging/application\_1484320332413\_0003/\_\_spark\_conf\_\_5522928782895023891.zip  17/01/13 20:54:33 INFO spark.SecurityManager: Changing view acls to: root  17/01/13 20:54:33 INFO spark.SecurityManager: Changing modify acls to: root  17/01/13 20:54:33 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(root); users with modify permissions: Set(root)  17/01/13 20:54:33 INFO yarn.Client: Submitting application 3 to ResourceManager  17/01/13 20:54:33 INFO impl.YarnClientImpl: Submitted application application\_1484320332413\_0003  17/01/13 20:54:34 INFO yarn.Client: Application report for application\_1484320332413\_0003 (state: ACCEPTED)  17/01/13 20:54:34 INFO yarn.Client:  client token: N/A  diagnostics: N/A  ApplicationMaster host: N/A  ApplicationMaster RPC port: -1  queue: root.users.root  start time: 1484321073040  final status: UNDEFINED  tracking URL: http://mac127:8088/proxy/application\_1484320332413\_0003/  user: root  17/01/13 20:54:35 INFO yarn.Client: Application report for application\_1484320332413\_0003 (state: ACCEPTED)  17/01/13 20:54:36 INFO yarn.Client: Application report for application\_1484320332413\_0003 (state: ACCEPTED)  17/01/13 20:54:37 INFO yarn.Client: Application report for application\_1484320332413\_0003 (state: ACCEPTED)  17/01/13 20:54:37 INFO cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: ApplicationMaster registered as NettyRpcEndpointRef(null)  17/01/13 20:54:37 INFO cluster.YarnClientSchedulerBackend: Add WebUI Filter. org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter, Map(PROXY\_HOSTS -> mac127, PROXY\_URI\_BASES -> http://mac127:8088/proxy/application\_1484320332413\_0003), /proxy/application\_1484320332413\_0003  17/01/13 20:54:37 INFO ui.JettyUtils: Adding filter: org.apache.hadoop.yarn.server.webproxy.amfilter.AmIpFilter  17/01/13 20:54:38 INFO yarn.Client: Application report for application\_1484320332413\_0003 (state: RUNNING)  17/01/13 20:54:38 INFO yarn.Client:  client token: N/A  diagnostics: N/A  ApplicationMaster host: 172.27.155.127  ApplicationMaster RPC port: 0  queue: root.users.root  start time: 1484321073040  final status: UNDEFINED  tracking URL: http://mac127:8088/proxy/application\_1484320332413\_0003/  user: root  17/01/13 20:54:38 INFO cluster.YarnClientSchedulerBackend: Application application\_1484320332413\_0003 has started running.  17/01/13 20:54:38 INFO util.Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 37592.  17/01/13 20:54:38 INFO netty.NettyBlockTransferService: Server created on 37592  17/01/13 20:54:38 INFO storage.BlockManager: external shuffle service port = 7337  17/01/13 20:54:38 INFO storage.BlockManagerMaster: Trying to register BlockManager  17/01/13 20:54:38 INFO storage.BlockManagerMasterEndpoint: Registering block manager 172.27.155.127:37592 with 530.3 MB RAM, BlockManagerId(driver, 172.27.155.127, 37592)  17/01/13 20:54:38 INFO storage.BlockManagerMaster: Registered BlockManager  17/01/13 20:54:38 INFO scheduler.EventLoggingListener: Logging events to hdfs://mac127:8020/user/spark/applicationHistory/application\_1484320332413\_0003  17/01/13 20:54:38 INFO cluster.YarnClientSchedulerBackend: SchedulerBackend is ready for scheduling beginning after reached minRegisteredResourcesRatio: 0.8  17/01/13 20:54:39 INFO storage.MemoryStore: Block broadcast\_0 stored as values in memory (estimated size 200.8 KB, free 200.8 KB)  17/01/13 20:54:39 INFO storage.MemoryStore: Block broadcast\_0\_piece0 stored as bytes in memory (estimated size 23.8 KB, free 224.6 KB)  17/01/13 20:54:39 INFO storage.BlockManagerInfo: Added broadcast\_0\_piece0 in memory on 172.27.155.127:37592 (size: 23.8 KB, free: 530.3 MB)  17/01/13 20:54:39 INFO spark.SparkContext: Created broadcast 0 from main at NativeMethodAccessorImpl.java:-2  17/01/13 20:54:40 INFO mapred.FileInputFormat: Total input paths to process : 1  17/01/13 20:54:40 INFO Configuration.deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id  17/01/13 20:54:40 INFO Configuration.deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id  17/01/13 20:54:40 INFO Configuration.deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap  17/01/13 20:54:40 INFO Configuration.deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition  17/01/13 20:54:40 INFO Configuration.deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id  17/01/13 20:54:40 INFO output.FileOutputCommitter: File Output Committer Algorithm version is 1  17/01/13 20:54:40 INFO spark.SparkContext: Starting job: main at NativeMethodAccessorImpl.java:-2  17/01/13 20:54:40 INFO scheduler.DAGScheduler: Registering RDD 3 (main at NativeMethodAccessorImpl.java:-2)  17/01/13 20:54:40 INFO scheduler.DAGScheduler: Got job 0 (main at NativeMethodAccessorImpl.java:-2) with 2 output partitions  17/01/13 20:54:40 INFO scheduler.DAGScheduler: Final stage: ResultStage 1 (main at NativeMethodAccessorImpl.java:-2)  17/01/13 20:54:40 INFO scheduler.DAGScheduler: Parents of final stage: List(ShuffleMapStage 0)  17/01/13 20:54:40 INFO scheduler.DAGScheduler: Missing parents: List(ShuffleMapStage 0)  17/01/13 20:54:40 INFO scheduler.DAGScheduler: Submitting ShuffleMapStage 0 (MapPartitionsRDD[3] at main at NativeMethodAccessorImpl.java:-2), which has no missing parents  17/01/13 20:54:40 INFO storage.MemoryStore: Block broadcast\_1 stored as values in memory (estimated size 4.1 KB, free 228.7 KB)  17/01/13 20:54:40 INFO storage.MemoryStore: Block broadcast\_1\_piece0 stored as bytes in memory (estimated size 2.3 KB, free 231.0 KB)  17/01/13 20:54:40 INFO storage.BlockManagerInfo: Added broadcast\_1\_piece0 in memory on 172.27.155.127:37592 (size: 2.3 KB, free: 530.3 MB)  17/01/13 20:54:40 INFO spark.SparkContext: Created broadcast 1 from broadcast at DAGScheduler.scala:1006  17/01/13 20:54:40 INFO scheduler.DAGScheduler: Submitting 2 missing tasks from ShuffleMapStage 0 (MapPartitionsRDD[3] at main at NativeMethodAccessorImpl.java:-2)  17/01/13 20:54:40 INFO cluster.YarnScheduler: Adding task set 0.0 with 2 tasks  17/01/13 20:54:41 INFO spark.ExecutorAllocationManager: Requesting 1 new executor because tasks are backlogged (new desired total will be 1)  17/01/13 20:54:42 INFO spark.ExecutorAllocationManager: Requesting 1 new executor because tasks are backlogged (new desired total will be 2)  17/01/13 20:54:48 INFO cluster.YarnClientSchedulerBackend: Registered executor NettyRpcEndpointRef(null) (mac127:40864) with ID 1  17/01/13 20:54:48 INFO spark.ExecutorAllocationManager: New executor 1 has registered (new total is 1)  17/01/13 20:54:48 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, mac127, executor 1, partition 0,NODE\_LOCAL, 2219 bytes)  17/01/13 20:54:48 INFO storage.BlockManagerMasterEndpoint: Registering block manager mac127:43513 with 265.4 MB RAM, BlockManagerId(1, mac127, 43513)  17/01/13 20:54:49 INFO storage.BlockManagerInfo: Added broadcast\_1\_piece0 in memory on mac127:43513 (size: 2.3 KB, free: 265.4 MB)  17/01/13 20:54:49 INFO cluster.YarnClientSchedulerBackend: Registered executor NettyRpcEndpointRef(null) (mac127:40866) with ID 2  17/01/13 20:54:49 INFO scheduler.TaskSetManager: Starting task 1.0 in stage 0.0 (TID 1, mac127, executor 2, partition 1,NODE\_LOCAL, 2219 bytes)  17/01/13 20:54:49 INFO spark.ExecutorAllocationManager: New executor 2 has registered (new total is 2)  17/01/13 20:54:49 INFO storage.BlockManagerMasterEndpoint: Registering block manager mac127:45422 with 265.4 MB RAM, BlockManagerId(2, mac127, 45422)  17/01/13 20:54:50 INFO storage.BlockManagerInfo: Added broadcast\_0\_piece0 in memory on mac127:43513 (size: 23.8 KB, free: 265.4 MB)  17/01/13 20:54:50 INFO storage.BlockManagerInfo: Added broadcast\_1\_piece0 in memory on mac127:45422 (size: 2.3 KB, free: 265.4 MB)  17/01/13 20:54:51 INFO storage.BlockManagerInfo: Added broadcast\_0\_piece0 in memory on mac127:45422 (size: 23.8 KB, free: 265.4 MB)  17/01/13 20:54:52 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 3918 ms on mac127 (executor 1) (1/2)  17/01/13 20:54:53 INFO scheduler.TaskSetManager: Finished task 1.0 in stage 0.0 (TID 1) in 3322 ms on mac127 (executor 2) (2/2)  17/01/13 20:54:53 INFO scheduler.DAGScheduler: ShuffleMapStage 0 (main at NativeMethodAccessorImpl.java:-2) finished in 12.570 s  17/01/13 20:54:53 INFO scheduler.DAGScheduler: looking for newly runnable stages  17/01/13 20:54:53 INFO cluster.YarnScheduler: Removed TaskSet 0.0, whose tasks have all completed, from pool  17/01/13 20:54:53 INFO scheduler.DAGScheduler: running: Set()  17/01/13 20:54:53 INFO scheduler.DAGScheduler: waiting: Set(ResultStage 1)  17/01/13 20:54:53 INFO scheduler.DAGScheduler: failed: Set()  17/01/13 20:54:53 INFO scheduler.DAGScheduler: Submitting ResultStage 1 (MapPartitionsRDD[5] at main at NativeMethodAccessorImpl.java:-2), which has no missing parents  17/01/13 20:54:53 INFO storage.MemoryStore: Block broadcast\_2 stored as values in memory (estimated size 74.4 KB, free 305.4 KB)  17/01/13 20:54:53 INFO storage.MemoryStore: Block broadcast\_2\_piece0 stored as bytes in memory (estimated size 26.7 KB, free 332.1 KB)  17/01/13 20:54:53 INFO storage.BlockManagerInfo: Added broadcast\_2\_piece0 in memory on 172.27.155.127:37592 (size: 26.7 KB, free: 530.2 MB)  17/01/13 20:54:53 INFO spark.SparkContext: Created broadcast 2 from broadcast at DAGScheduler.scala:1006  17/01/13 20:54:53 INFO scheduler.DAGScheduler: Submitting 2 missing tasks from ResultStage 1 (MapPartitionsRDD[5] at main at NativeMethodAccessorImpl.java:-2)  17/01/13 20:54:53 INFO cluster.YarnScheduler: Adding task set 1.0 with 2 tasks  17/01/13 20:54:53 INFO scheduler.TaskSetManager: Starting task 0.0 in stage 1.0 (TID 2, mac127, executor 2, partition 0,NODE\_LOCAL, 1977 bytes)  17/01/13 20:54:53 INFO scheduler.TaskSetManager: Starting task 1.0 in stage 1.0 (TID 3, mac127, executor 1, partition 1,NODE\_LOCAL, 1977 bytes)  17/01/13 20:54:53 INFO storage.BlockManagerInfo: Added broadcast\_2\_piece0 in memory on mac127:43513 (size: 26.7 KB, free: 265.4 MB)  17/01/13 20:54:53 INFO storage.BlockManagerInfo: Added broadcast\_2\_piece0 in memory on mac127:45422 (size: 26.7 KB, free: 265.4 MB)  17/01/13 20:54:53 INFO spark.MapOutputTrackerMasterEndpoint: Asked to send map output locations for shuffle 0 to mac127:40864  17/01/13 20:54:53 INFO spark.MapOutputTrackerMasterEndpoint: Asked to send map output locations for shuffle 0 to mac127:40866  17/01/13 20:54:53 INFO spark.MapOutputTrackerMaster: Size of output statuses for shuffle 0 is 149 bytes  17/01/13 20:54:53 INFO spark.MapOutputTrackerMaster: Size of output statuses for shuffle 0 is 149 bytes  17/01/13 20:54:53 INFO scheduler.TaskSetManager: Finished task 0.0 in stage 1.0 (TID 2) in 639 ms on mac127 (executor 2) (1/2)  17/01/13 20:54:53 INFO scheduler.TaskSetManager: Finished task 1.0 in stage 1.0 (TID 3) in 696 ms on mac127 (executor 1) (2/2)  17/01/13 20:54:53 INFO cluster.YarnScheduler: Removed TaskSet 1.0, whose tasks have all completed, from pool  17/01/13 20:54:53 INFO scheduler.DAGScheduler: ResultStage 1 (main at NativeMethodAccessorImpl.java:-2) finished in 0.698 s  17/01/13 20:54:53 INFO scheduler.DAGScheduler: Job 0 finished: main at NativeMethodAccessorImpl.java:-2, took 13.581121 s  17/01/13 20:54:54 INFO spark.SparkContext: Invoking stop() from shutdown hook  17/01/13 20:54:54 WARN thread.QueuedThreadPool: 4 threads could not be stopped  17/01/13 20:54:54 INFO ui.SparkUI: Stopped Spark web UI at http://172.27.155.127:4040  17/01/13 20:54:54 INFO storage.BlockManagerInfo: Removed broadcast\_2\_piece0 on 172.27.155.127:37592 in memory (size: 26.7 KB, free: 530.3 MB)  17/01/13 20:54:54 INFO storage.BlockManagerInfo: Removed broadcast\_2\_piece0 on mac127:43513 in memory (size: 26.7 KB, free: 265.4 MB)  17/01/13 20:54:54 INFO storage.BlockManagerInfo: Removed broadcast\_2\_piece0 on mac127:45422 in memory (size: 26.7 KB, free: 265.4 MB)  17/01/13 20:54:54 INFO cluster.YarnClientSchedulerBackend: Shutting down all executors  17/01/13 20:54:54 INFO cluster.YarnClientSchedulerBackend: Interrupting monitor thread  17/01/13 20:54:54 INFO cluster.YarnClientSchedulerBackend: Asking each executor to shut down  17/01/13 20:54:54 INFO cluster.YarnClientSchedulerBackend: Stopped  17/01/13 20:54:54 INFO spark.MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!  17/01/13 20:54:54 INFO storage.MemoryStore: MemoryStore cleared  17/01/13 20:54:54 INFO storage.BlockManager: BlockManager stopped  17/01/13 20:54:54 INFO storage.BlockManagerMaster: BlockManagerMaster stopped  17/01/13 20:54:54 INFO scheduler.OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!  17/01/13 20:54:54 INFO remote.RemoteActorRefProvider$RemotingTerminator: Shutting down remote daemon.  17/01/13 20:54:54 INFO remote.RemoteActorRefProvider$RemotingTerminator: Remote daemon shut down; proceeding with flushing remote transports.  17/01/13 20:54:54 INFO spark.SparkContext: Successfully stopped SparkContext  17/01/13 20:54:54 INFO util.ShutdownHookManager: Shutdown hook called  17/01/13 20:54:54 INFO util.ShutdownHookManager: Deleting directory /tmp/spark-aaf0f98c-1a59-4c2e-a208-ea8b49480012 |

Now check the output directory

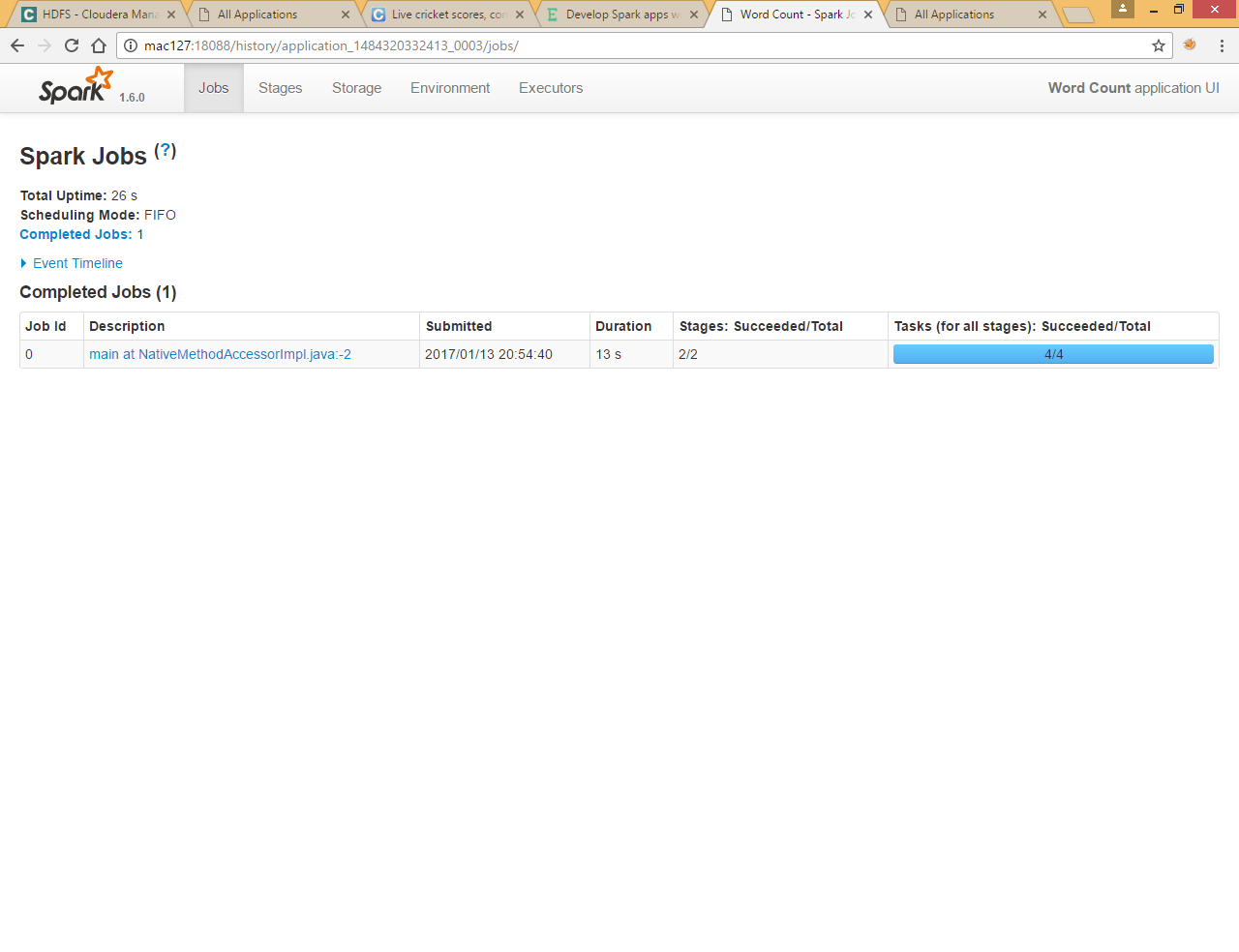
|  |
| --- |
| [root@mac127 spark-test]# hdfs dfs -ls /user/test/output  Found 3 items  -rw-r--r-- 3 root supergroup 0 2017-01-13 20:54 /user/test/output/\_SUCCESS  -rw-r--r-- 3 root supergroup 3605 2017-01-13 20:54 /user/test/output/part-00000  -rw-r--r-- 3 root supergroup 3634 2017-01-13 20:54 /user/test/output/part-00001 |
| [root@mac127 spark-test]# hdfs dfs -cat /user/test/output/part-00000  (transitions,1)  (created,1)  (underestimate,1)  (structures.,1)  (its,4)  (past,1)  (previously,1)  (ELECTROMAGNETIC,1)  (writing,1)  (PARAGRAPH,1)  (have,10)  (OUR,1)  (include,1)  (creatures,1)  (order,1)  ((with,1)  (several,2)  (we,9)  (This,2)  (been,2)  (dangerous,1)  (any,1)  (contrast,1)  (make,4)  (kind,1)  (subdivisions,1)  (constructed,1)  (little,1)  (illustrates,1)  (parts,1)  …. |

Now open Resource manager web UI to see the logs

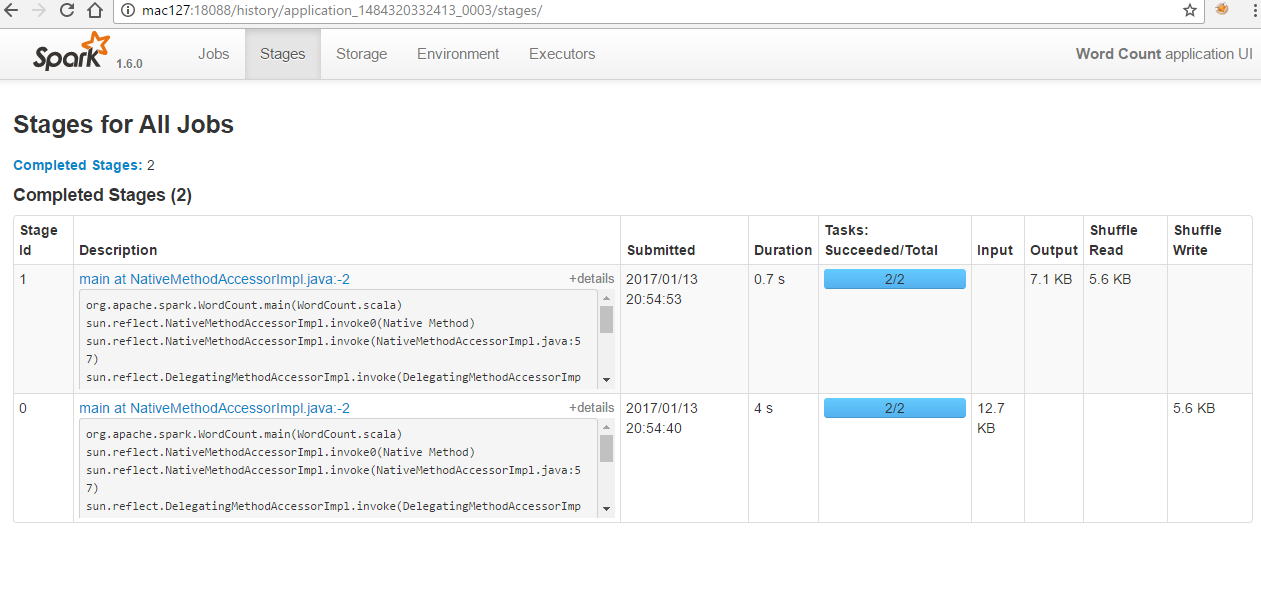
<http://mac127:8088/>



Click on History tab to see the Spark logs

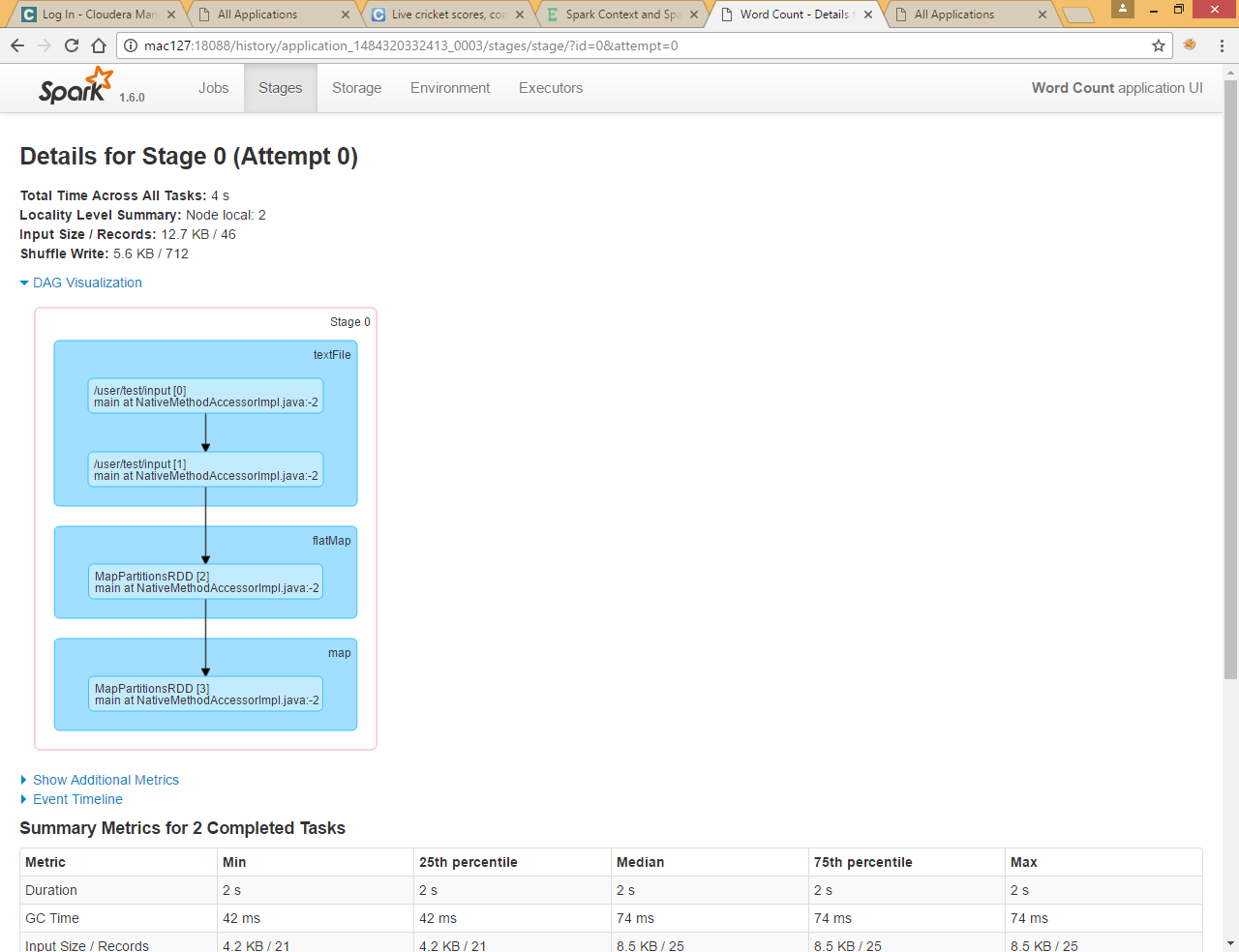


Here you can see different tab and you can get several information about spark job



Refer below link for more information

You can see further information regarding stage by clicking on stage links



<http://www.ask-dg.com/topic/spark-context-and-spark-configuration/>