[Intelij IDEA 远程调试spark standalone 集群](http://www.cnblogs.com/ooon/p/5490247.html)

环境： WIN7笔记本

　　　 spark 集群（4个虚拟机搭建的）

　　　 Intelij IDEA15

　　　 scala-2.10.4

　　　 java-1.7.0

版本问题：

个人选择的是hadoop2.6.0 spark1.5.0 scala2.10.4  jdk1.7.0

关于搭建集群环境，见个人的上一篇博客：[(一) Spark Standalone集群环境搭建](http://www.cnblogs.com/ooon/p/5460060.html)，接下来就是用Intelij IDEA来远程连接spark集群，这样就可以方便的在本机上进行调试。

首先需要注意windows可以设置hosts，在 C:\Windows\System32\drivers\etc 有个hosts，把以下映射地址填进去， 这样能省去不少事

172.21.75.102   spark1

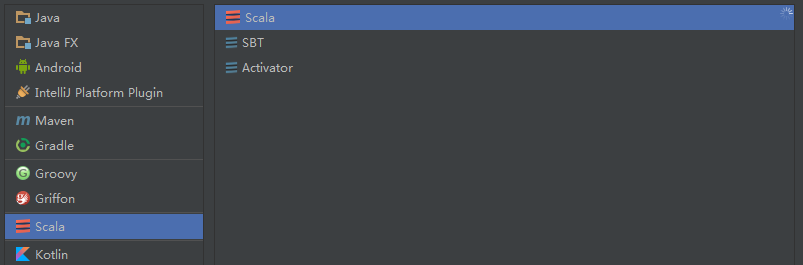
172.21.75.194   spark2

172.21.75.122   spark3

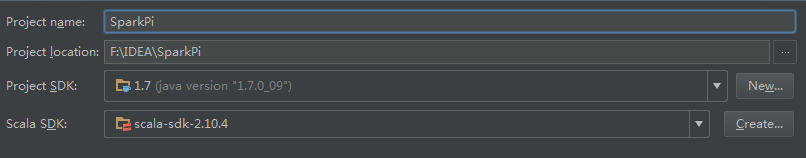
172.21.75.95   spark4

1）首先在个人WIN7本上搭好java，scala环境，并配置好环境变量，安装好Intelij IDEA，并安装好scala插件。

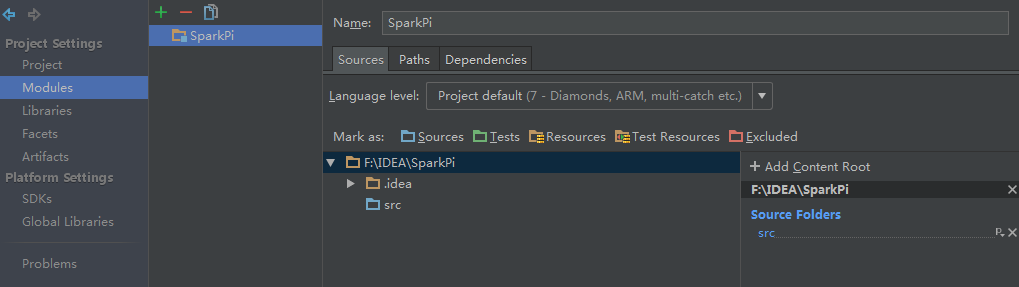
2）新建Scala项目，选择Scala：



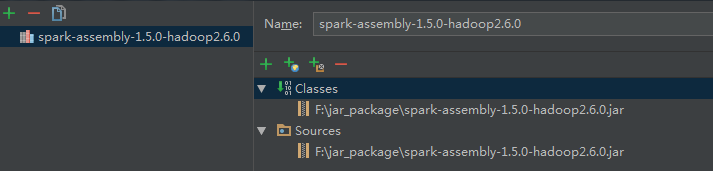
3）分别引入 java 与 Scala SDK，并对项目命名，这里一会我们运行SparkPi的程序，名字可以随意



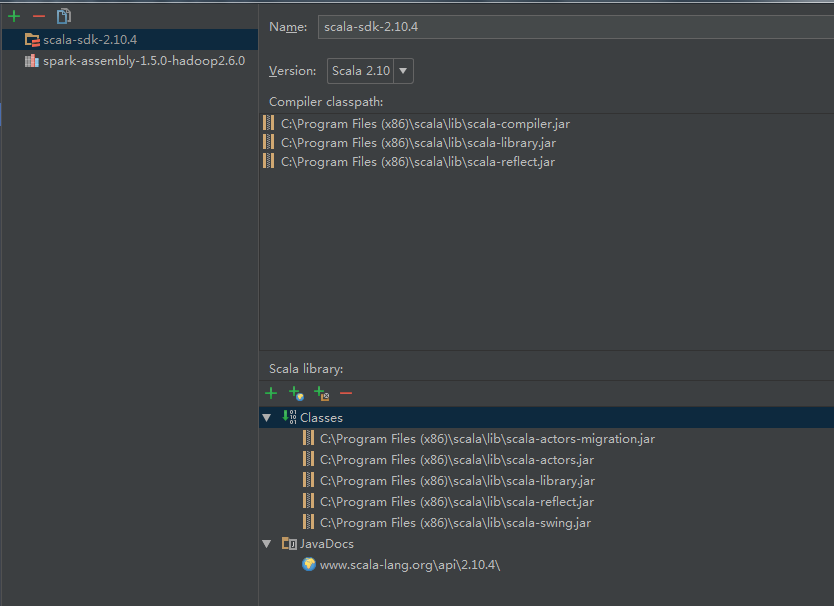
4）进入主界面，双击src，或者File->Project Structer,进入程序配置界面



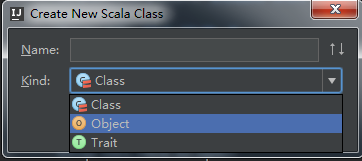
5）点击library里“+”，点击java，添加spark-1.5.0-hadoop-2.6.0的jar包



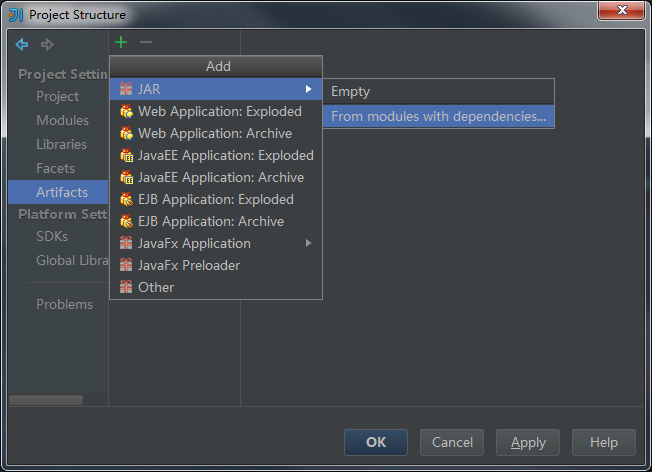
6）点击library里“+”，点击Scala SDK 添加Scala SDK



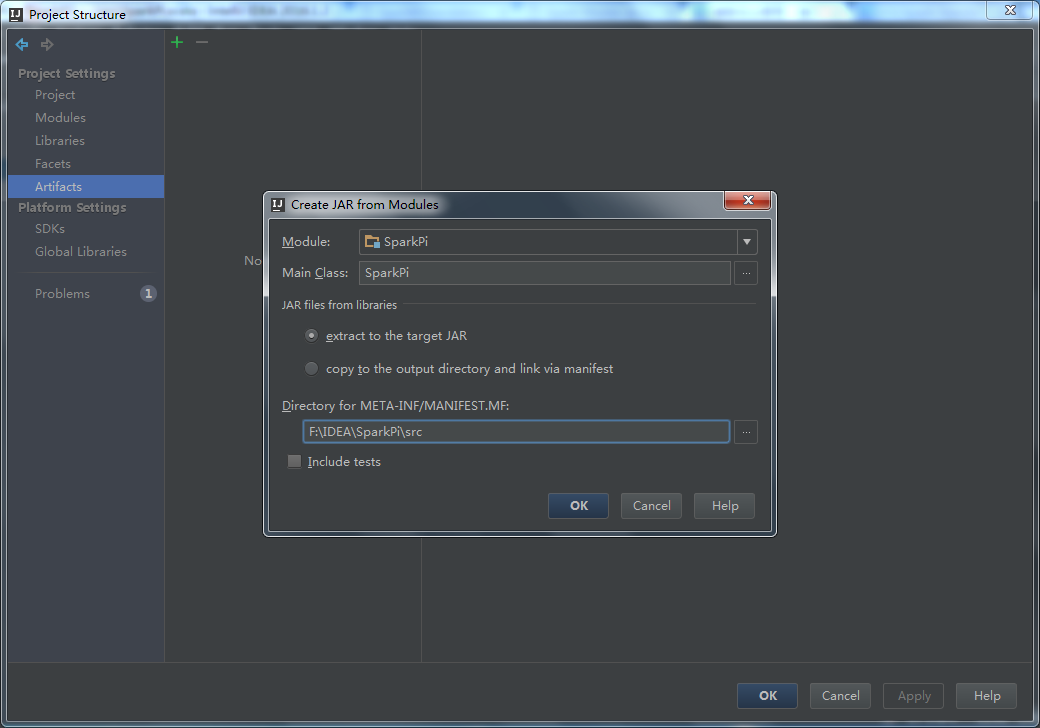
7）以上步骤点击OK退出，在src新建 SparkPi.scala 的scala object文件



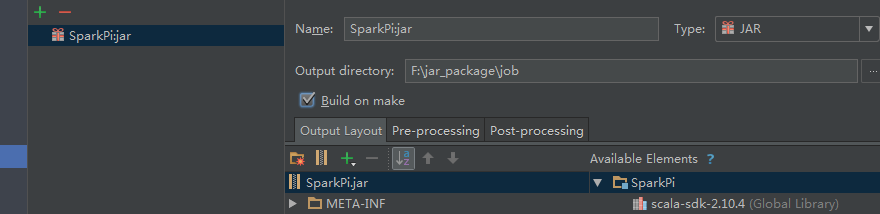
8）写代码之前，先进行一个jar包设置



9) 这里的路径一定要设置好，为jar包的输出路径，一会要写到程序里，使得spark集群的查找



10）选中这里的Build on make，程序就会编译后自动打包



11）注意以上的路径，这个路径就是提交给spark的jar包

.setJars(List("F:\\jar\_package\\job\\SparkPi.jar"))

12）复制如下代码到SparkPi.scala

import scala.math.random

import org.apache.spark.{SparkConf, SparkContext}

/\*\*

\* Created by Administrator on 2016/5/13.

\*/

//alt+Enter自动引入缺失的包

object SparkPi {

def main(args: Array[String]) {

val conf = new SparkConf().setAppName("Spark Pi").setMaster("spark://172.21.75.102:7077")

.setJars(List("F:\\jar\_package\\job\\SparkPi.jar"))

val spark = new SparkContext(conf)

val slices = if (args.length > 0) args(0).toInt else 2

val n = 100000 \* slices

val count = spark.parallelize(1 to n, slices).map { i =>

val x = random \* 2 - 1

val y = random \* 2 - 1

if (x \* x + y \* y < 1) 1 else 0

}.reduce(\_ + \_)

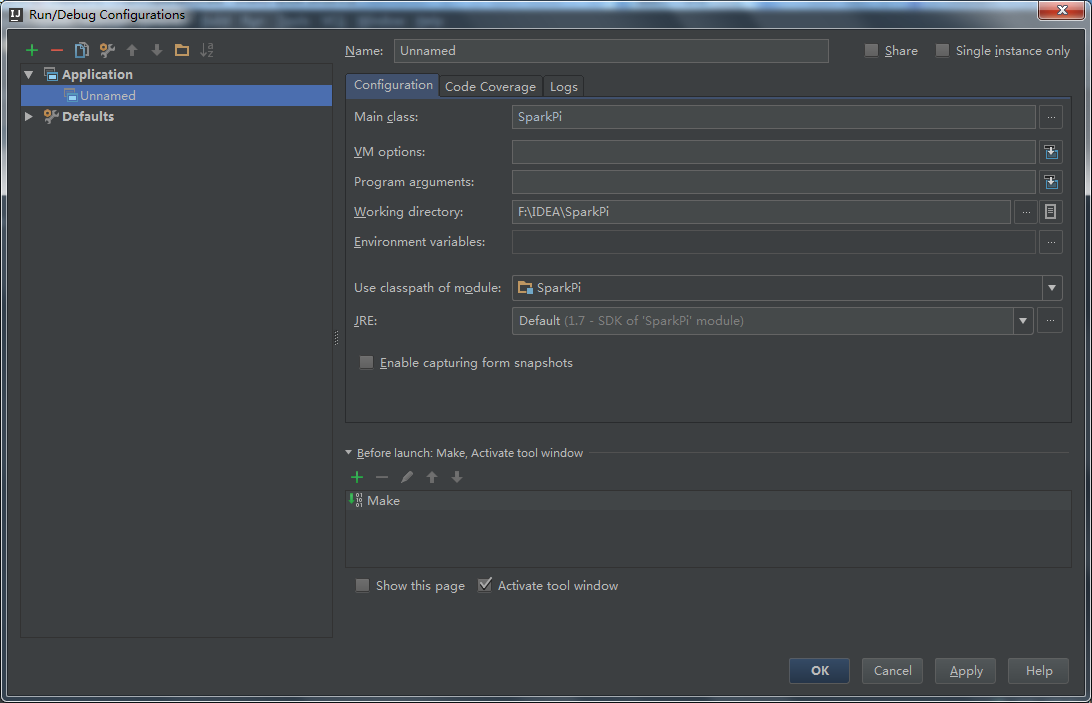
println("Pi is roughly " + 4.0 \* count / n)

spark.stop()

}

}

13）现在大功告成，设置Run 的Edit Configuration，点击+，Application，设置MainClass，点击OK！



14）点击Run即可运行程序了，程序会在刚才的路径生成对应的jar，然后会启动spark集群，去运行该jar文件，以下为执行结果：

"C:\Program Files\Java\jdk1.7.0\_09\bin\java" -Didea.launcher.port=7534 "-Didea.launcher.bin.path=D:\IntelliJ IDEA Community Edition 2016.1.2\bin" -Dfile.encoding=UTF-8 -classpath "C:\Program Files\Java\jdk1.7.0\_09\jre\lib\charsets.jar;C:\Program Files\Java\jdk1.7.0\_09\jre\lib\deploy.jar;C:\Program Files\Java\jdk1.7.0\_09\jre\lib\ext\access-bridge-64.jar;C:\Program Files\Java\jdk1.7.0\_09\jre\lib\ext\dnsns.jar;C:\Program Files\Java\jdk1.7.0\_09\jre\lib\ext\jaccess.jar;C:\Program Files\Java\jdk1.7.0\_09\jre\lib\ext\localedata.jar;C:\Program Files\Java\jdk1.7.0\_09\jre\lib\ext\sunec.jar;C:\Program Files\Java\jdk1.7.0\_09\jre\lib\ext\sunjce\_provider.jar;C:\Program Files\Java\jdk1.7.0\_09\jre\lib\ext\sunmscapi.jar;C:\Program Files\Java\jdk1.7.0\_09\jre\lib\ext\zipfs.jar;C:\Program Files\Java\jdk1.7.0\_09\jre\lib\javaws.jar;C:\Program Files\Java\jdk1.7.0\_09\jre\lib\jce.jar;C:\Program Files\Java\jdk1.7.0\_09\jre\lib\jfr.jar;C:\Program Files\Java\jdk1.7.0\_09\jre\lib\jfxrt.jar;C:\Program Files\Java\jdk1.7.0\_09\jre\lib\jsse.jar;C:\Program Files\Java\jdk1.7.0\_09\jre\lib\management-agent.jar;C:\Program Files\Java\jdk1.7.0\_09\jre\lib\plugin.jar;C:\Program Files\Java\jdk1.7.0\_09\jre\lib\resources.jar;C:\Program Files\Java\jdk1.7.0\_09\jre\lib\rt.jar;F:\IDEA\SparkPi\out\production\SparkPi;C:\Program Files (x86)\scala\lib\scala-actors-migration.jar;C:\Program Files (x86)\scala\lib\scala-actors.jar;C:\Program Files (x86)\scala\lib\scala-library.jar;C:\Program Files (x86)\scala\lib\scala-reflect.jar;C:\Program Files (x86)\scala\lib\scala-swing.jar;F:\jar\_package\spark-assembly-1.5.0-hadoop2.6.0.jar;D:\IntelliJ IDEA Community Edition 2016.1.2\lib\idea\_rt.jar" com.intellij.rt.execution.application.AppMain SparkPi

Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties

16/05/13 17:47:43 INFO SparkContext: Running Spark version 1.5.0

16/05/13 17:47:53 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

16/05/13 17:47:55 INFO SecurityManager: Changing view acls to: Administrator

16/05/13 17:47:55 INFO SecurityManager: Changing modify acls to: Administrator

16/05/13 17:47:55 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(Administrator); users with modify permissions: Set(Administrator)

16/05/13 17:47:58 INFO Slf4jLogger: Slf4jLogger started

16/05/13 17:47:58 INFO Remoting: Starting remoting

16/05/13 17:48:00 INFO Remoting: Remoting started; listening on addresses :[akka.tcp://sparkDriver@172.21.75.63:62339]

16/05/13 17:48:00 INFO Utils: Successfully started service 'sparkDriver' on port 62339.

16/05/13 17:48:00 INFO SparkEnv: Registering MapOutputTracker

16/05/13 17:48:00 INFO SparkEnv: Registering BlockManagerMaster

16/05/13 17:48:00 INFO DiskBlockManager: Created local directory at C:\Users\Administrator\AppData\Local\Temp\blockmgr-0046600a-5752-4cd5-89d6-cde41f7011d1

16/05/13 17:48:01 INFO MemoryStore: MemoryStore started with capacity 484.8 MB

16/05/13 17:48:01 INFO HttpFileServer: HTTP File server directory is C:\Users\Administrator\AppData\Local\Temp\spark-4d4d665e-45ad-4ea9-b664-c95eeeb5f8b5\httpd-756f1b24-34a1-48a2-969c-6cc7a5d4cb57

16/05/13 17:48:01 INFO HttpServer: Starting HTTP Server

16/05/13 17:48:01 INFO Utils: Successfully started service 'HTTP file server' on port 62340.

16/05/13 17:48:01 INFO SparkEnv: Registering OutputCommitCoordinator

16/05/13 17:48:02 INFO Utils: Successfully started service 'SparkUI' on port 4040.

16/05/13 17:48:02 INFO SparkUI: Started SparkUI at http://172.21.75.63:4040

16/05/13 17:48:03 INFO SparkContext: Added JAR F:\jar\_package\job\SparkPi.jar at http://172.21.75.63:62340/jars/SparkPi.jar with timestamp 1463132883308

16/05/13 17:48:04 WARN MetricsSystem: Using default name DAGScheduler for source because spark.app.id is not set.

16/05/13 17:48:04 INFO AppClient$ClientEndpoint: Connecting to master spark://172.21.75.102:7077...

16/05/13 17:48:06 INFO SparkDeploySchedulerBackend: Connected to Spark cluster with app ID app-20160513024433-0002

16/05/13 17:48:06 INFO AppClient$ClientEndpoint: Executor added: app-20160513024433-0002/0 on worker-20160513012923-172.21.75.102-44267 (172.21.75.102:44267) with 1 cores

16/05/13 17:48:06 INFO SparkDeploySchedulerBackend: Granted executor ID app-20160513024433-0002/0 on hostPort 172.21.75.102:44267 with 1 cores, 1024.0 MB RAM

16/05/13 17:48:06 INFO AppClient$ClientEndpoint: Executor added: app-20160513024433-0002/1 on worker-20160513012924-172.21.75.95-54009 (172.21.75.95:54009) with 1 cores

16/05/13 17:48:06 INFO SparkDeploySchedulerBackend: Granted executor ID app-20160513024433-0002/1 on hostPort 172.21.75.95:54009 with 1 cores, 1024.0 MB RAM

16/05/13 17:48:06 INFO AppClient$ClientEndpoint: Executor added: app-20160513024433-0002/2 on worker-20160513012924-172.21.75.194-35992 (172.21.75.194:35992) with 1 cores

16/05/13 17:48:06 INFO SparkDeploySchedulerBackend: Granted executor ID app-20160513024433-0002/2 on hostPort 172.21.75.194:35992 with 1 cores, 1024.0 MB RAM

16/05/13 17:48:06 INFO AppClient$ClientEndpoint: Executor added: app-20160513024433-0002/3 on worker-20160513012923-172.21.75.122-39901 (172.21.75.122:39901) with 1 cores

16/05/13 17:48:06 INFO SparkDeploySchedulerBackend: Granted executor ID app-20160513024433-0002/3 on hostPort 172.21.75.122:39901 with 1 cores, 1024.0 MB RAM

16/05/13 17:48:06 INFO AppClient$ClientEndpoint: Executor updated: app-20160513024433-0002/1 is now LOADING

16/05/13 17:48:06 INFO AppClient$ClientEndpoint: Executor updated: app-20160513024433-0002/0 is now LOADING

16/05/13 17:48:06 INFO AppClient$ClientEndpoint: Executor updated: app-20160513024433-0002/2 is now LOADING

16/05/13 17:48:06 INFO AppClient$ClientEndpoint: Executor updated: app-20160513024433-0002/3 is now LOADING

16/05/13 17:48:06 INFO AppClient$ClientEndpoint: Executor updated: app-20160513024433-0002/0 is now RUNNING

16/05/13 17:48:06 INFO AppClient$ClientEndpoint: Executor updated: app-20160513024433-0002/1 is now RUNNING

16/05/13 17:48:06 INFO AppClient$ClientEndpoint: Executor updated: app-20160513024433-0002/2 is now RUNNING

16/05/13 17:48:06 INFO AppClient$ClientEndpoint: Executor updated: app-20160513024433-0002/3 is now RUNNING

16/05/13 17:48:07 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 62360.

16/05/13 17:48:07 INFO NettyBlockTransferService: Server created on 62360

16/05/13 17:48:07 INFO BlockManagerMaster: Trying to register BlockManager

16/05/13 17:48:07 INFO BlockManagerMasterEndpoint: Registering block manager 172.21.75.63:62360 with 484.8 MB RAM, BlockManagerId(driver, 172.21.75.63, 62360)

16/05/13 17:48:07 INFO BlockManagerMaster: Registered BlockManager

16/05/13 17:48:08 INFO SparkDeploySchedulerBackend: SchedulerBackend is ready for scheduling beginning after reached minRegisteredResourcesRatio: 0.0

16/05/13 17:48:09 INFO SparkDeploySchedulerBackend: Registered executor: AkkaRpcEndpointRef(Actor[akka.tcp://sparkExecutor@172.21.75.194:57560/user/Executor#-786956451]) with ID 2

16/05/13 17:48:10 INFO BlockManagerMasterEndpoint: Registering block manager 172.21.75.194:48333 with 530.3 MB RAM, BlockManagerId(2, 172.21.75.194, 48333)

16/05/13 17:48:10 INFO SparkDeploySchedulerBackend: Registered executor: AkkaRpcEndpointRef(Actor[akka.tcp://sparkExecutor@172.21.75.102:60131/user/Executor#1889839276]) with ID 0

16/05/13 17:48:10 INFO BlockManagerMasterEndpoint: Registering block manager 172.21.75.102:33896 with 530.3 MB RAM, BlockManagerId(0, 172.21.75.102, 33896)

16/05/13 17:48:10 INFO SparkContext: Starting job: reduce at SparkPi.scala:19

16/05/13 17:48:10 INFO DAGScheduler: Got job 0 (reduce at SparkPi.scala:19) with 2 output partitions

16/05/13 17:48:10 INFO DAGScheduler: Final stage: ResultStage 0(reduce at SparkPi.scala:19)

16/05/13 17:48:10 INFO DAGScheduler: Parents of final stage: List()

16/05/13 17:48:10 INFO DAGScheduler: Missing parents: List()

16/05/13 17:48:11 INFO DAGScheduler: Submitting ResultStage 0 (MapPartitionsRDD[1] at map at SparkPi.scala:15), which has no missing parents

16/05/13 17:48:11 INFO SparkDeploySchedulerBackend: Registered executor: AkkaRpcEndpointRef(Actor[akka.tcp://sparkExecutor@172.21.75.95:42263/user/Executor#1076811589]) with ID 1

16/05/13 17:48:11 INFO BlockManagerMasterEndpoint: Registering block manager 172.21.75.95:50679 with 530.3 MB RAM, BlockManagerId(1, 172.21.75.95, 50679)

16/05/13 17:48:12 INFO SparkDeploySchedulerBackend: Registered executor: AkkaRpcEndpointRef(Actor[akka.tcp://sparkExecutor@172.21.75.122:36331/user/Executor#-893021210]) with ID 3

16/05/13 17:48:12 INFO MemoryStore: ensureFreeSpace(1832) called with curMem=0, maxMem=508369305

16/05/13 17:48:12 INFO MemoryStore: Block broadcast\_0 stored as values in memory (estimated size 1832.0 B, free 484.8 MB)

16/05/13 17:48:12 INFO MemoryStore: ensureFreeSpace(1189) called with curMem=1832, maxMem=508369305

16/05/13 17:48:12 INFO MemoryStore: Block broadcast\_0\_piece0 stored as bytes in memory (estimated size 1189.0 B, free 484.8 MB)

16/05/13 17:48:12 INFO BlockManagerInfo: Added broadcast\_0\_piece0 in memory on 172.21.75.63:62360 (size: 1189.0 B, free: 484.8 MB)

16/05/13 17:48:12 INFO SparkContext: Created broadcast 0 from broadcast at DAGScheduler.scala:861

16/05/13 17:48:12 INFO BlockManagerMasterEndpoint: Registering block manager 172.21.75.122:59662 with 530.3 MB RAM, BlockManagerId(3, 172.21.75.122, 59662)

16/05/13 17:48:12 INFO DAGScheduler: Submitting 2 missing tasks from ResultStage 0 (MapPartitionsRDD[1] at map at SparkPi.scala:15)

16/05/13 17:48:12 INFO TaskSchedulerImpl: Adding task set 0.0 with 2 tasks

16/05/13 17:48:13 INFO TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, 172.21.75.194, PROCESS\_LOCAL, 2137 bytes)

16/05/13 17:48:13 INFO TaskSetManager: Starting task 1.0 in stage 0.0 (TID 1, 172.21.75.102, PROCESS\_LOCAL, 2194 bytes)

16/05/13 17:49:21 INFO BlockManagerInfo: Added broadcast\_0\_piece0 in memory on 172.21.75.102:33896 (size: 1189.0 B, free: 530.3 MB)

16/05/13 17:49:22 INFO TaskSetManager: Finished task 1.0 in stage 0.0 (TID 1) in 68937 ms on 172.21.75.102 (1/2)

16/05/13 17:49:42 INFO BlockManagerInfo: Added broadcast\_0\_piece0 in memory on 172.21.75.194:48333 (size: 1189.0 B, free: 530.3 MB)

16/05/13 17:49:42 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 90038 ms on 172.21.75.194 (2/2)

16/05/13 17:49:42 INFO DAGScheduler: ResultStage 0 (reduce at SparkPi.scala:19) finished in 90.071 s

16/05/13 17:49:42 INFO TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool

16/05/13 17:49:42 INFO DAGScheduler: Job 0 finished: reduce at SparkPi.scala:19, took 92.205022 s

Pi is roughly 3.13816

16/05/13 17:49:42 INFO SparkUI: Stopped Spark web UI at http://172.21.75.63:4040

16/05/13 17:49:42 INFO DAGScheduler: Stopping DAGScheduler

16/05/13 17:49:42 INFO SparkDeploySchedulerBackend: Shutting down all executors

16/05/13 17:49:42 INFO SparkDeploySchedulerBackend: Asking each executor to shut down

16/05/13 17:49:43 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!

16/05/13 17:49:43 INFO MemoryStore: MemoryStore cleared

16/05/13 17:49:43 INFO BlockManager: BlockManager stopped

16/05/13 17:49:43 INFO BlockManagerMaster: BlockManagerMaster stopped

16/05/13 17:49:43 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!

16/05/13 17:49:43 INFO SparkContext: Successfully stopped SparkContext

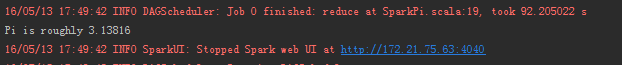
16/05/13 17:49:43 INFO RemoteActorRefProvider$RemotingTerminator: Shutting down remote daemon.

16/05/13 17:49:43 INFO ShutdownHookManager: Shutdown hook called

16/05/13 17:49:43 INFO RemoteActorRefProvider$RemotingTerminator: Remote daemon shut down; proceeding with flushing remote transports.

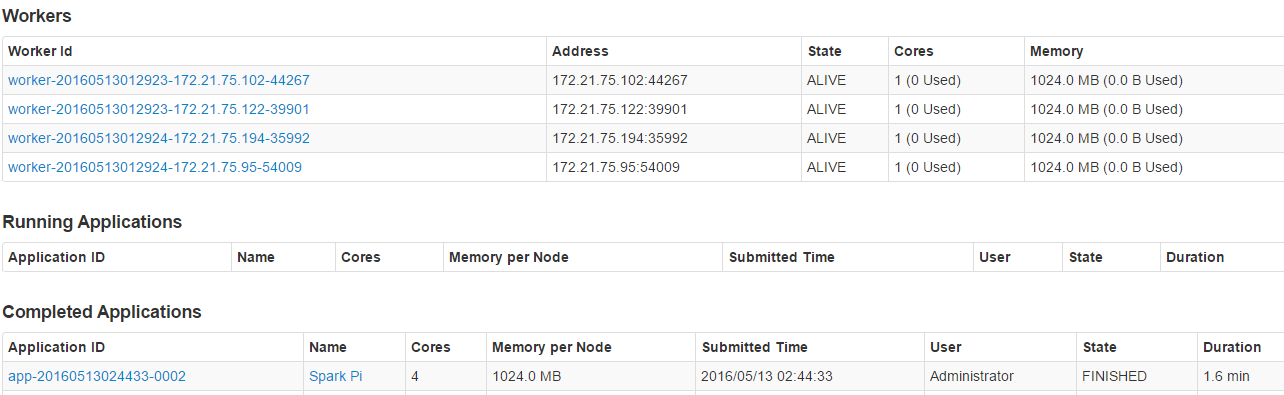
16/05/13 17:49:43 INFO ShutdownHookManager: Deleting directory C:\Users\Administrator\AppData\Local\Temp\spark-4d4d665e-45ad-4ea9-b664-c95eeeb5f8b5

Process finished with exit code 0



看着真是有点小激动！

15）去172.21.75.102:8080查看运行的痕迹



16）搭建调试环境过程中的错误

* null\bin\winutils.exe，这个错误很简单，是因为本win7压根就没装hadoop系统，解决办法是从集群上复制一份过来,放到F盘，并且配置好环境变量

HADOOP\_HOME=F:\hadoop-2.6.0

Path=%HADOOP\_HOME%\bin

   接下来下载对应的版本的winutils放到 F:\hadoop-2.6.0\bin 文件夹下，应该就解决了

* SparkUncaughtExceptionHandler: Uncaught exception in thread Thread

这个错误好坑，查了好久的资料，才解决，原来是搭建集群时候spark-env.sh设置的问题

将SPARK\_MASTER\_IP=spark1改成

SPARK\_MASTER\_IP=172.21.75.102即可解决，改了之后再网页里也能查出来



* Exception in thread "main" java.lang.IllegalArgumentException: java.net.UnknownHostException : spark1

以上是当需要操作HDFS时候，写上HDFS地址 hdfs://spark1:9000,会出现，后来发现原来windows也可以设置hosts，在 C:\Windows\System32\drivers\etc 有个hosts，把需要映射的地址填进去即可

172.21.75.102   spark1

* FAILED: RuntimeException org.apache.hadoop.security.AccessControlException: org.apache.hadoop.security.AccessControlException: Permission denied: user=dbs, access=WRITE, inode="/opt/hadoop-1.0.1":hadoop:supergroup:drwxr-xr-x

解决办法：

在 hdfs-site.xml 总添加参数：

<configuration>  
  <property>  
        <name>dfs.permissions</name>  
        <value>**false**</value>  
  </property>    
</configuration>

改完后记得重启HDFS