**一、下载安装如下工具**

1、下载安装

jdk1.8

C:\Program Files\Java\jdk1.8.0\_121

2、解压缩Hadoop

hadoop-2.6.5

2、下载安装

scala-2.11.8 windows版本，如安装在

C:\scala

3、下载解压缩

spark-2.1.0-bin-hadoop2.6

如解压缩在

C:\spark-2.1.0-bin-hadoop2.6

4、配置环境变量

Java\_Home,Scala\_Home,Spark\_Home，HADOOP\_HOME

5、1）、下载winutils.exe放到spark目录的bin目录下，并设置环境变量HADOOP\_HOME=%SPARK\_HOME%

2、出现throws java.lang.RuntimeException: The root scratch dir: /tmp/hive on HDFS should be writable.错误

运行：

**winutils.exe chmod 777 /tmp/hive**

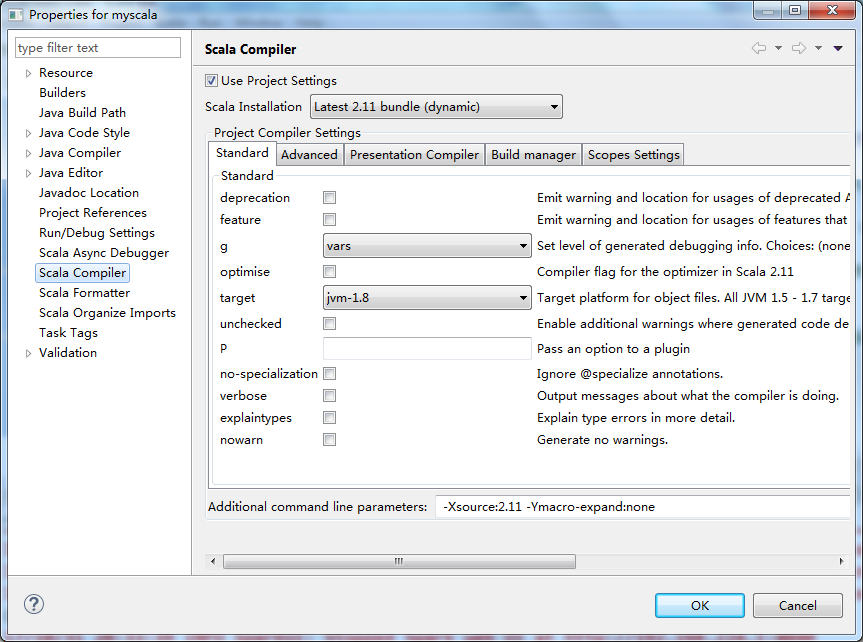
再使用以下命令启动就OK了

**spark-shell --master local[2]**

6、Eclipse for Scala 开发spark程序

解压缩scala-SDK-4.6.0-vfinal-2.12-win32.win32.x86\_64 创建Scala项目，导入spark jars目录中的jar文件。

右击项目属性



**package** com

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

**object** WordCount {

**def** main(args: Array[*String*]) {

/\*\*

\* 第1步；创建Spark的配置对象SparkConf，设置Spark程序运行时的配置信息

\* 例如 setAppName用来设置应用程序的名称，在程序运行的监控界面可以看到该名称，

\* setMaster设置程序运行在本地还是运行在集群中，运行在本地可是使用local参数，也可以使用local[K]/local[\*],

\* 可以去spark官网查看它们不同的意义。 如果要运行在集群中，以Standalone模式运行的话，需要使用spark://HOST:PORT

\* 的形式指定master的IP和端口号，默认是7077

\*/

**val** conf = **new** SparkConf().setAppName("WordCount").setMaster("local")

/\*\*

\* 第2步：创建SparkContext 对象

\* SparkContext是Spark程序所有功能的唯一入口

\* SparkContext核心作用： 初始化Spark应用程序运行所需要的核心组件，包括DAGScheduler、TaskScheduler、SchedulerBackend

\* 同时还会负责Spark程序往Master注册程序

\*

\* 通过传入SparkConf实例来定制Spark运行的具体参数和配置信息

\*/

**val** sc = **new** SparkContext(conf)

/\*\*

\* 第3步： 根据具体的数据来源(HDFS、 HBase、Local FS、DB、 S3等)通过SparkContext来创建RDD

\* RDD 的创建基本有三种方式： 根据外部的数据来源(例如HDFS)、根据Scala集合使用SparkContext的parallelize方法、

\* 由其他的RDD操作产生

\* 数据会被RDD划分成为一系列的Partitions，分配到每个Partition的数据属于一个Task的处理范畴

\*/

**val** lines = sc.textFile("D:/word.txt") // 读取本地文件

// val lines = sc.textFile("/library/wordcount/input")

// 读取HDFS文件，并切分成不同的Partition

val lines = sc.textFile("hdfs://master:9000/libarary/wordcount/input")

// 或者明确指明是从HDFS上获取数据

/\*\*

\* 第4步： 对初始的RDD进行Transformation级别的处理，例如 map、filter等高阶函数来进行具体的数据计算

\*/

**val** words = lines.flatMap(\_.split(" ")).filter(word => word != " ") // 拆分单词，并过滤掉空格，当然还可以继续进行过滤，如去掉标点符号

**val** pairs = words.map(word => (word, 1)) // 在单词拆分的基础上对每个单词实例计数为1, 也就是 word => (word, 1)

**val** wordscount = pairs.reduceByKey(\_ + \_) // 在每个单词实例计数为1的基础之上统计每个单词在文件中出现的总次数, 即key相同的value相加

// val wordscount = pairs.reduceByKey((v1, v2) => v1 + v2) // 等同于

wordscount.collect.foreach(println) // 打印结果，使用collect会将集群中的数据收集到当前运行drive的机器上，需要保证单台机器能放得下所有数据

sc.stop() // 释放资源

}

}

运行结果如下：

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

17/10/31 20:33:35 INFO SparkContext: Running Spark version 2.1.0

17/10/31 20:33:35 INFO SecurityManager: Changing view acls to: zhouping

17/10/31 20:33:35 INFO SecurityManager: Changing modify acls to: zhouping

17/10/31 20:33:35 INFO SecurityManager: Changing view acls groups to:

17/10/31 20:33:35 INFO SecurityManager: Changing modify acls groups to:

17/10/31 20:33:35 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(zhouping); groups with view permissions: Set(); users with modify permissions: Set(zhouping); groups with modify permissions: Set()

17/10/31 20:33:36 INFO Utils: Successfully started service 'sparkDriver' on port 5818.

17/10/31 20:33:36 INFO SparkEnv: Registering MapOutputTracker

17/10/31 20:33:36 INFO SparkEnv: Registering BlockManagerMaster

17/10/31 20:33:36 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information

17/10/31 20:33:36 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up

17/10/31 20:33:36 INFO DiskBlockManager: Created local directory at C:\Users\zhouping\AppData\Local\Temp\blockmgr-f435c28c-510a-47ff-9a91-fd5796a7bb07

17/10/31 20:33:36 INFO MemoryStore: MemoryStore started with capacity 897.6 MB

17/10/31 20:33:36 INFO SparkEnv: Registering OutputCommitCoordinator

17/10/31 20:33:37 INFO Utils: Successfully started service 'SparkUI' on port 4040.

17/10/31 20:33:37 INFO SparkUI: Bound SparkUI to 0.0.0.0, and started at http://192.168.224.1:4040

17/10/31 20:33:37 INFO Executor: Starting executor ID driver on host localhost

17/10/31 20:33:37 INFO Utils: Successfully started service 'org.apache.spark.network.netty.NettyBlockTransferService' on port 5827.

17/10/31 20:33:37 INFO NettyBlockTransferService: Server created on 192.168.224.1:5827

17/10/31 20:33:37 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy

17/10/31 20:33:37 INFO BlockManagerMaster: Registering BlockManager BlockManagerId(driver, 192.168.224.1, 5827, None)

17/10/31 20:33:37 INFO BlockManagerMasterEndpoint: Registering block manager 192.168.224.1:5827 with 897.6 MB RAM, BlockManagerId(driver, 192.168.224.1, 5827, None)

17/10/31 20:33:37 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver, 192.168.224.1, 5827, None)

17/10/31 20:33:37 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, 192.168.224.1, 5827, None)

17/10/31 20:33:38 INFO MemoryStore: Block broadcast\_0 stored as values in memory (estimated size 208.6 KB, free 897.4 MB)

17/10/31 20:33:38 INFO MemoryStore: Block broadcast\_0\_piece0 stored as bytes in memory (estimated size 20.1 KB, free 897.4 MB)

17/10/31 20:33:38 INFO BlockManagerInfo: Added broadcast\_0\_piece0 in memory on 192.168.224.1:5827 (size: 20.1 KB, free: 897.6 MB)

17/10/31 20:33:38 INFO SparkContext: Created broadcast 0 from textFile at WordCount.scala:37

17/10/31 20:33:38 INFO FileInputFormat: Total input paths to process : 1

17/10/31 20:33:38 INFO SparkContext: Starting job: collect at WordCount.scala:51

17/10/31 20:33:38 INFO DAGScheduler: Registering RDD 4 (map at WordCount.scala:46)

17/10/31 20:33:38 INFO DAGScheduler: Got job 0 (collect at WordCount.scala:51) with 1 output partitions

17/10/31 20:33:38 INFO DAGScheduler: Final stage: ResultStage 1 (collect at WordCount.scala:51)

17/10/31 20:33:38 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 0)

17/10/31 20:33:38 INFO DAGScheduler: Missing parents: List(ShuffleMapStage 0)

17/10/31 20:33:38 INFO DAGScheduler: Submitting ShuffleMapStage 0 (MapPartitionsRDD[4] at map at WordCount.scala:46), which has no missing parents

17/10/31 20:33:38 INFO MemoryStore: Block broadcast\_1 stored as values in memory (estimated size 4.9 KB, free 897.4 MB)

17/10/31 20:33:38 INFO MemoryStore: Block broadcast\_1\_piece0 stored as bytes in memory (estimated size 2.8 KB, free 897.4 MB)

17/10/31 20:33:38 INFO BlockManagerInfo: Added broadcast\_1\_piece0 in memory on 192.168.224.1:5827 (size: 2.8 KB, free: 897.6 MB)

17/10/31 20:33:38 INFO SparkContext: Created broadcast 1 from broadcast at DAGScheduler.scala:996

17/10/31 20:33:38 INFO DAGScheduler: Submitting 1 missing tasks from ShuffleMapStage 0 (MapPartitionsRDD[4] at map at WordCount.scala:46)

17/10/31 20:33:38 INFO TaskSchedulerImpl: Adding task set 0.0 with 1 tasks

17/10/31 20:33:38 INFO TaskSetManager: Starting task 0.0 in stage 0.0 (TID 0, localhost, executor driver, partition 0, PROCESS\_LOCAL, 5954 bytes)

17/10/31 20:33:38 INFO Executor: Running task 0.0 in stage 0.0 (TID 0)

17/10/31 20:33:39 INFO HadoopRDD: Input split: file:/D:/word.txt:0+51

17/10/31 20:33:39 INFO deprecation: mapred.tip.id is deprecated. Instead, use mapreduce.task.id

17/10/31 20:33:39 INFO deprecation: mapred.task.id is deprecated. Instead, use mapreduce.task.attempt.id

17/10/31 20:33:39 INFO deprecation: mapred.task.is.map is deprecated. Instead, use mapreduce.task.ismap

17/10/31 20:33:39 INFO deprecation: mapred.task.partition is deprecated. Instead, use mapreduce.task.partition

17/10/31 20:33:39 INFO deprecation: mapred.job.id is deprecated. Instead, use mapreduce.job.id

17/10/31 20:33:39 INFO Executor: Finished task 0.0 in stage 0.0 (TID 0). 1658 bytes result sent to driver

17/10/31 20:33:39 INFO TaskSetManager: Finished task 0.0 in stage 0.0 (TID 0) in 264 ms on localhost (executor driver) (1/1)

17/10/31 20:33:39 INFO TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool

17/10/31 20:33:39 INFO DAGScheduler: ShuffleMapStage 0 (map at WordCount.scala:46) finished in 0.296 s

17/10/31 20:33:39 INFO DAGScheduler: looking for newly runnable stages

17/10/31 20:33:39 INFO DAGScheduler: running: Set()

17/10/31 20:33:39 INFO DAGScheduler: waiting: Set(ResultStage 1)

17/10/31 20:33:39 INFO DAGScheduler: failed: Set()

17/10/31 20:33:39 INFO DAGScheduler: Submitting ResultStage 1 (ShuffledRDD[5] at reduceByKey at WordCount.scala:48), which has no missing parents

17/10/31 20:33:39 INFO MemoryStore: Block broadcast\_2 stored as values in memory (estimated size 3.2 KB, free 897.4 MB)

17/10/31 20:33:39 INFO MemoryStore: Block broadcast\_2\_piece0 stored as bytes in memory (estimated size 1967.0 B, free 897.4 MB)

17/10/31 20:33:39 INFO BlockManagerInfo: Added broadcast\_2\_piece0 in memory on 192.168.224.1:5827 (size: 1967.0 B, free: 897.6 MB)

17/10/31 20:33:39 INFO SparkContext: Created broadcast 2 from broadcast at DAGScheduler.scala:996

17/10/31 20:33:39 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 1 (ShuffledRDD[5] at reduceByKey at WordCount.scala:48)

17/10/31 20:33:39 INFO TaskSchedulerImpl: Adding task set 1.0 with 1 tasks

17/10/31 20:33:39 INFO TaskSetManager: Starting task 0.0 in stage 1.0 (TID 1, localhost, executor driver, partition 0, ANY, 5750 bytes)

17/10/31 20:33:39 INFO Executor: Running task 0.0 in stage 1.0 (TID 1)

17/10/31 20:33:39 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks

17/10/31 20:33:39 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 16 ms

17/10/31 20:33:39 INFO Executor: Finished task 0.0 in stage 1.0 (TID 1). 2035 bytes result sent to driver

17/10/31 20:33:39 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 62 ms on localhost (executor driver) (1/1)

17/10/31 20:33:39 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool

17/10/31 20:33:39 INFO DAGScheduler: ResultStage 1 (collect at WordCount.scala:51) finished in 0.062 s

17/10/31 20:33:39 INFO DAGScheduler: Job 0 finished: collect at WordCount.scala:51, took 0.820817 s

(c#,2)

17/10/31 20:33:39 INFO SparkUI: Stopped Spark web UI at http://192.168.224.1:4040

(vb,2)

(Hello,3)

(php,1)

(java,3)

17/10/31 20:33:39 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!

17/10/31 20:33:39 INFO MemoryStore: MemoryStore cleared

17/10/31 20:33:39 INFO BlockManager: BlockManager stopped

17/10/31 20:33:39 INFO BlockManagerMaster: BlockManagerMaster stopped

17/10/31 20:33:39 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!

17/10/31 20:33:39 INFO SparkContext: Successfully stopped SparkContext

17/10/31 20:33:39 INFO ShutdownHookManager: Shutdown hook called

17/10/31 20:33:39 INFO ShutdownHookManager: Deleting directory C:\Users\zhouping\AppData\Local\Temp\spark-c5a8f888-5d7a-4d98-8d7a-84bb593833b5