Spark 安装与操作

北京瑞德云网科技有限公司

Spark安装与操作

- 实验目的

掌握Spark的基本安装过程以及Spark shell 的基本操作。

实验步骤

• step 1.下载安装Spark

tar -xzf spark-1.5.1-bin-hadoop2.tgz

step 2 执行 bin/spark-shell

Spark抽象的分布式集群空间叫做Resilient Distributed Dataset (RDD)，即弹性数据集。

• step 3 利用Hdfs 上的一个文本文件创建一个新的RDD：

val tt = sc.textFile(“hdfs://hadoop:9000/profile”)

step 4 RDD有两种类型的操作，分别是Action（返回values）和Transformations(返回一个新的RDD）

tt.first() // 返回RDD第一行的内容

tt.count() //RDD 中有多少行

step 5 Transformations相当于一个转换，会将一个RDD转换，并返回一个新的RDD:

textFile.filter(line => line.contains(“hello”)).count() // 有多少行含有hello

实验步骤

step 6 wordcount

```
val file = sc.textFile( “hdfs://hadoop:9000/user/hadoop/wordcount/in” )
var count = file.flatMap(line => line.split( “ ” )).map(word => (word,1)).reduceByKey(_+_)
count.collect()
count.saveAsTextFile( “hdfs://hadoop:9000/wordcount” )
```

The End.