学习笔记

# Task

## 1、任务

**环境部署**

Windows安装<https://blog.csdn.net/SummerHmh/article/details/89518567>

Linux参考1<https://mp.weixin.qq.com/s/gIf_QGQL26MqMzZKk-YTLg>

Linux参考2<https://mp.weixin.qq.com/s/vufnqfG1vYjT9Ec8kvjcAg>

**内容：运行原理，RDD设计，DAG，安装与使用**

1. Spark的设计与运行原理（大概了解）

[1.1 Spark简介](http://dblab.xmu.edu.cn/blog/1710-2/) <http://dblab.xmu.edu.cn/blog/1710-2/>

[1.2 Spark运行架构](http://dblab.xmu.edu.cn/blog/1711-2/) <http://dblab.xmu.edu.cn/blog/1711-2/>

[1.3 RDD的设计与运行原理](http://dblab.xmu.edu.cn/blog/1681-2/) <http://dblab.xmu.edu.cn/blog/1681-2/>

[1.4 Spark的部署模式](1.4%20Spark的部署模式) <http://dblab.xmu.edu.cn/blog/1681-2/>

第2章 Spark的安装与使用（主要内容）

[2.1 Spark的安装和使用](2.1%20Spark的安装和使用) <http://dblab.xmu.edu.cn/blog/1689-2/>（如果想在window上安装，[参考](https://blog.csdn.net/SummerHmh/article/details/89518567" \t "_blank)，之后可以用pyspark或者jupyter上进行学习）

* 1. 第一个Spark应用程序：WordCount

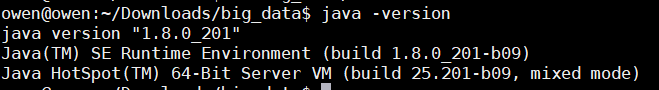
**问题扩展：**

spark 和 mapreduce有哪些区别，请用具体的例子说明？

rdd的本质是什么？

## 2、环境部署记录

检查安装java



## 3、学习记录

### Wordcout

text\_file = sc.textFile("hdfs://...")

counts = text\_file.flatMap(**lambda** line: line.split(" ")) \

.map(**lambda** word: (word, 1)) \

.reduceByKey(**lambda** a, b: a + b)

counts.saveAsTextFile("hdfs://...")

## 4、问题

### spark 和 mapreduce有哪些区别，请用具体的例子说明？

1. 首先了解一下Mapreduce，它最本质的两个过程就是Map和Reduce，Map的应用在于我们需要数据一对一的元素的映射转换，比如说进行截取，进行过滤，或者任何的转换操作，这些一对一的元素转换就称作是Map；Reduce主要就是元素的聚合，就是多个元素对一个元素的聚合，比如求Sum等，这就是Reduce。
2. 目前的MapReduce框架都是把中间结果写入到HDFS中，带来了大量的数据复制、磁盘IO和序列化开销。
3. Spark在借鉴Hadoop MapReduce优点的同时，很好地解决了MapReduce所面临的问题。相比于MapReduce，Spark主要具有如下优点：

Spark的计算模式也属于MapReduce，但不局限于Map和Reduce操作，还提供了多种数据集操作类型，编程模型比MapReduce更灵活；

Spark提供了内存计算，中间结果直接放到内存中，带来了更高的迭代运算效率；

Spark基于DAG的任务调度执行机制，要优于MapReduce的迭代执行机制。

### rdd的本质是什么？

RDD：是弹性分布式数据集（Resilient Distributed Dataset）的简称，是分布式内存的一个抽象概念，提供了一种高度受限的共享内存模型；