|  |
| --- |
| 北京航空航天大学 |
| 软件开发计划书 |
| Apache Spark |
|  |
| SY1506404 孟翰 SY1506409 苏若 SY1506425 李璇 SY1506406 孙敏芳 |
| **2016/03/16** |

**软件开发计划书**

# 1 范围

## 1.1 标识

Spark版本号： spark1.6.1

Spark源码实现语言：Scala

可编程语言：Scala.+Scala shell、Java、Python+Scala shell

环境：jdk1.8

模 块 名 称： 1. Spark环境搭建与调试

2. RDD容错性机制、分析Spark的优势与不足

3. Spark并行计算原理

4. Spark基于RDD编程模型的操作

需求报告版本：V1.0

## 1.2 项目概述

1.2.1 项目初衷

随着技术的进步以及人类对大数据处理性能的要求愈发苛刻，MapReduce逐渐存在的使用局限性愈发突出。而异军突起的Apache Spark,由于其具有的高性能、高灵活性、与Hadoop生态系统完美融合等三方面的特点，使得愈发成为大数据处理的宠儿，已成为大数据范畴最活跃的开源码项目之一。当然，Spark本身也存在一些缺陷，主要包括以下几个方面：  
 1、稳定性方面仍需加强。

2、不能处理大数据。  
 3、不能支持复杂的SQL统计。

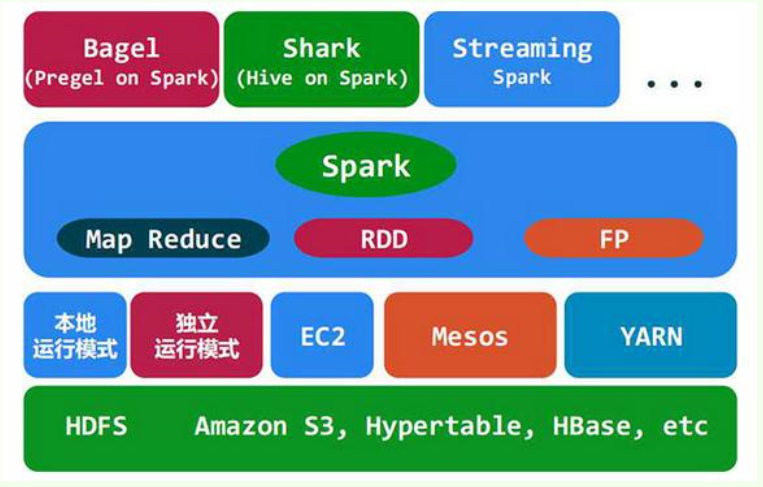
4、在可管理性方面，SparkYARN的结合不完善。

……

由于Spark集时代性趋势和开源性于一身，研究其不足与改进方案较有相对现实的意义和影响。

1.2.2 系统概述

Spark是UC Berkeley AMP lab所开源的类Hadoop MapReduce的通用的并行计算框架，Spark基于map reduce算法实现的分布式计算，拥有Hadoop MapReduce所具有的优点；但不同于MapReduce的是JobIMG_256中间输出和结果可以保存在内存中，从而不再需要读写HDFS，因此Spark能更好地适用于数据挖掘与机器学习等需要迭代的map reduce的算法。其架构如下图所示：



## 1.3 文档概述

本文档是Apache Spark的部分模块的开发计划书，主要依据相关标准和条约，对应交付的成果和任务分工等作了较为详细的安排，便于后续项目有条不紊的进行。

## 1.4 标准、条约和约定

本项目遵从以下标准：

GB/T 13702-1992 计算机软件分类与代码

GB/T 19003-2008 软件工程

GB/T 5538-1995 软件工程标准分类法

GB/T 9385-2008 计算机软件需求规格说明

GB/T 5532-2008 计算机软件测试规范

GB/T 18221-2000 信息技术程序设计语言

GB/T 8567-2006 计算机软件文档编制规范

# 2 项目概述

## 2.1 工作内容

Apache Spark是新兴的大数据处理引擎，主要特点是提供了一个集群的分布式内存抽象，以支持需要工作集的应用。

通过实验我们希望掌握关于Spark的以下机理与内容：

1） Spark的核心RDD（Resilient Distributed DataSet）的容错性（fault-tolerant）机制

2）Spark的并行计算（parallel）原理

3）Spark基于RDD编程模型的操作（map，flatMap，filter，union，sample，join等）。

4）分析Spark的优势与不足（与Hadoop的比较，对迭代式数据处理的优势）。

## 

## 2.3 假设与约束

使用此文档的人有相关的计算机背景知识，对项目管理也有一定的了解。

小组可调用的人员的上限为4人 ，本项目的规定的完成时间为第16周，最迟为第17周，预算资金无。

## 2.2 应交付成果。

### 2.2.1 文档

* 项目计划书

制定软件开发计划书作为项目管理的重要阶段，在项目中起承上启下的作用，因此在制定过程中要按照项目总目标、总计划进行详细计划。在开发计划制订过程中一般应遵循六个原则，即目的性、系统性、经济性、动态性、相关性和职能性。

* 需求规格说明书

软件需求说明书的编制是为了使用户和软件开发者双方对该软件的初始规定有一个共同的理解， 使之成为整个开发工作的基础。编制的原则如下：

* 从现实中分离功能，即描述要“做什么”而不是“怎样实现”；
* 要求使用面向处理的规格说明语言，从而得到“做什么”的规格说明；
* 如果被开发软件只是一个大系统中的一个元素，那么整个大系统也包括在规格说明的描述之中；
* 规格说明必须是可操作的；
* 规格说明必须容许不完备性并允许扩充；
* 规格说明必须局部化和松散耦合。
* 评审单

作为一种复查手段，在每一阶段的最后一步，应该对功能的正确性、完整性和清晰性，以及其它需求给予评价。而评审单正是评审人员评审的依据，更是他们关注的问题的总结，是评审阶段的产出物。

* 测试规格说明书

测试需求规格说明书由测试经理或测试人员撰写，是对产品实现的功能进行分析且细化为可以指导测试用例编写，同样，还需要对隐性需求进行分析，比如性能方面，兼容性，可扩展性，安全性，健壮性等等进行分析，以及测试计划说明，比如是否需要和外围进行联调，什么时间联调，测试执行轮次说明，准入准出条件等说明。

* 项目总结报告

综合所有的需求分析、测试分析、及软件更新与展示分析以及项目管理分析报告，完成最后和实验分析、总结报告，是对整个软件工程实践的总结和反思。

### 2.2.2 软件

Spark1.6.1，jdk1.8，sbt，scala，eclipse mars或linux。

# 3 实施计划

## 3.1 人员分工

每个人所负责的模块如下表，需要完成各自模块的需求分析和测试过程的相关文档、代码的编写工作。

|  |  |
| --- | --- |
| **成员** | **分工** |
| 孟翰 | Spark环境搭建与调试 |
| 苏若 | RDD容错性机制、分析Spark的优势与不足 |
| 孙敏芳 | Spark并行计算原理 |
| 李璇 | Spark基于RDD编程模型的操作 |

## 3.2 组织形式

小组会议：每周日晚21:00，交流各自负责部分的成果，互相督促检查；

网上协作：有问题随时在QQ群中讨论，分享相关资料和各自成果。

## 3.3 进度

项目计划进度表如下，详见附件中的“项目计划.mpp”：

# 4 参考资料

[1] <http://spark.apache.org/>

[2] <https://databricks.com/blog/2014/10/10/spark-petabyte-sort.html>

[3] 许鹏.《Apache Spark源码剖析》. [电子工业出版社](http://book.jd.com/publish/%E7%94%B5%E5%AD%90%E5%B7%A5%E4%B8%9A%E5%87%BA%E7%89%88%E7%A4%BE_1.html): 2015-03-01.