|  |
| --- |
|  |
| 软件项目计划书 |
| 基于开源框架Spark的研究与应用 |
|  |
| **温元祯** |
| **2017/3/17** |

|  |
| --- |
|  |
|  |

**文档更改历史记录**

|  |  |  |  |
| --- | --- | --- | --- |
| **初始信息** | | | |
| **文件名称** | 项目计划书 | **编写人** | 温元祯 |
| **初始版本号** | V1.0 | **发布日期** | 2017/3/17 |
| **更改记录** | | | |
| **版本号** | **更改要点** | **修改人** | **修改日期** |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

目录

[1. 引言 4](#_Toc477534919)

[1.1 目的 4](#_Toc477534920)

[1.2 术语 4](#_Toc477534921)

[1.3 参考资料 4](#_Toc477534922)

[2. 项目概述 5](#_Toc477534923)

[2.1 项目简介 5](#_Toc477534924)

[2.2 产品目标与范围 5](#_Toc477534925)

[2.3 应交付成果 5](#_Toc477534926)

[2.4 项目开发环境 6](#_Toc477534927)

[2.4.1 软件环境 6](#_Toc477534928)

[2.4.2 硬件环境 6](#_Toc477534929)

[2.4.3 开发工具 6](#_Toc477534930)

[2.5 项目验收方式和依据 6](#_Toc477534931)

[2.5.1 验收方式 6](#_Toc477534932)

[2.5.2 验收依据 6](#_Toc477534933)

[3. 项目团队组织 7](#_Toc477534934)

[3.1 组织结构 7](#_Toc477534935)

[3.2 项目分工 7](#_Toc477534936)

[3.3 协作与沟通 7](#_Toc477534937)

[4. 实施计划 7](#_Toc477534938)

[4.1 工作流程 7](#_Toc477534939)

[4.2 总体进度计划 7](#_Toc477534940)

[4.3 项目控制计划 7](#_Toc477534941)

1. **引言**
   1. **目的**

本文档的编写目的是为了协调组内成员开展后期的工作，对项目的进度进行合理的安排和调控。

* 1. **术语**

关于本文档中出现的一些专业术语、缩略语如表1所示。

|  |  |
| --- | --- |
| 术语/缩略语 | 描述 |
| AMP | Algorithms Machines and People |
| Spark SQL | Spark SQL是一个即席查询系统，可以通过SQL表达式、HiveQL或者Scala DSL在Spark上执行查询。 |
| Spark Streaming | Spark Streaming是一个对实时数据流进行高通量、容错处理的流式处理系统，可以对多种数据源进行类似map、reduce等复杂操作，并将结果保存到外部文件系统、数据库或应用到实时仪表盘。 |
| MLLib | MLlib是Spark实现一些常见的机器学习算法和实用程序，包括分类，回归，聚类，协同过滤，降维，以及底层。 |
| GraphX | GraphX是基于Spark的图处理和图并行计算API。 |
| BDAS | The Berkeley Data Analytics Stack |
| RDD | Resilient Distributed Dataset(弹性分布式数据集), 一个可并行操作的有容错机制的数据集合。 |
| task | stage 下的一个任务执行单元。每一个 task 只是处理一个 partition 上的数据。 |
| Spark job | 包含很多task的并行计算，可以认为**是Spark RDD 里面的action**,每个action的计算会生成一个job。 |

表 1专业术语/缩略图描述表

* 1. **参考资料**

[1]《spark-programming-guide-zh-cn》

[2]《spark运维实战》

1. **项目概述**
   1. **项目简介**

Apache Spark是加州大学伯克利分校AMP实验室开发的通用内存并行计算框架。Spark以其先进的设计理念，迅速成为社区的热门项目，围绕着Spark推出了Spark SQL、Spark Streaming、MLLib和GraphX等组件，也就是BDAS（伯克利数据分析栈），这些组件逐渐形成大数据处理一站式的解决平台。

Spark使用内存构建弹性分布式数据集RDD,对数据进行运算和cache。其中，计算中间数据在内存中维护，存取速度是存盘的多个数量级。Task以线程的方式维护，对小数据集的读取能达到亚秒级的延迟。Spark引擎提供了在集群中所有主机上进行分布式内存数据处理的能力，图1显示了一个典型Spark job的处理流程。

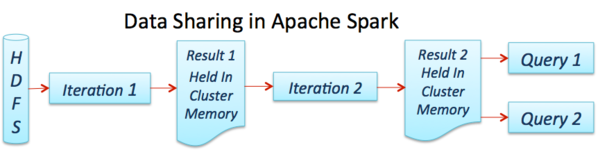


图 1. 典型Spark job的处理流程

* 1. **产品目标与范围**

对Spark做需求分析，编写需求说明书。对Spark做测试需求分析，编写需求报告，并进行模块测试。在Spark的基础上扩展应用。

* 1. **应交付成果**

项目应交付的产品如表2所示。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 编号 | 名称 | 交付时间 | 要求 | 备注 |
| No.1 | 项目计划书 | 2016/3/17 | 完整、明确 | 暂无 |
| No.2 | 项目进度表 | 2016/3/16 | 完整、详细 | 暂无 |
| No.3 | 项目工作日志 | 待定 | 暂无 | 暂无 |
| No.4 | 软件需求说明书 | 待定 | 暂无 | 暂无 |
| No.5 | 需求检查单 | 待定 | 暂无 | 暂无 |
| No.6 | 测试需求说明书 | 待定 | 暂无 | 暂无 |
| No.7 | 软件测试分析报告 | 待定 | 暂无 | 暂无 |
| No.8 | 软件测试评审报告 | 待定 | 暂无 | 暂无 |
| No.9 | 项目工作量统计报告 | 待定 | 暂无 | 暂无 |
| No.10 | 项目方法总结报告 | 待定 | 暂无 | 暂无 |
| No.11 | 项目个人总结 | 待定 | 暂无 | 暂无 |

表 2 应交付成果列表

* 1. **项目开发环境**
     1. 软件环境

Windows 10/CentOS7

* + 1. 硬件环境

4G内存 + 250G硬盘 +百兆网卡 + 百兆局域网

* + 1. 开发工具

JDK1.8 + Eclipse4.X + Maven4.X

* 1. **项目验收方式和依据**
     1. 验收方式

验收方式为每周上课汇报、项目扩展代码以及实验验收文档。

* + 1. 验收依据

验收依据“应交付成果”清单进行验收。

1. **项目团队组织**
   1. **组织结构**

项目团队包括一名组长，三名组员。组长和组员协调合作。组长对项目进度、工作计划进行合理安排并分配相应的任务。组长和组员需及时完成任务。

* 1. **项目分工**

项目具体分工详见《Spark项目规划书v1.0版》。

* 1. **协作与沟通**

小组成员利用微信、邮件等通讯工具及时交流，并一周举行1-2次面谈会议。小组通过github进行文档的提交和更新。

所有沟通协作都要进行记录。

1. **实施计划**
   1. **工作流程**

工作流程参照《Spark项目规划书v1.0版》。

* 1. **总体进度计划**

总体进度计划参照《Spark项目规划书v1.0版》。

* 1. **项目控制计划**

暂无。