流式实时分布式并行计算框架Spark的研究与应用

软件项目计划书

**编写人：于思民**

**2016年3月16日星期三**

目 录

[1. 引言 1](#_Toc445920420)

[1.1 背景 1](#_Toc445920421)

[1.2 定义 2](#_Toc445920422)

[1.3 参考资料 2](#_Toc445920423)

[1.4 标准、条约和约定 2](#_Toc445920424)

[2. 项目概述 3](#_Toc445920425)

[2.1 项目目标 3](#_Toc445920426)

[2.2 产品目标与范围 3](#_Toc445920427)

[2.3 假设与约定 3](#_Toc445920428)

[2.4 应交付成果 3](#_Toc445920429)

[2.5 项目开发环境 4](#_Toc445920430)

[2.6 项目验收方式和依据 4](#_Toc445920431)

[3. 项目团队组织 4](#_Toc445920432)

[3.1 组织结构 4](#_Toc445920433)

[3.2 人员分工 5](#_Toc445920434)

[3.3 协作与沟通 5](#_Toc445920435)

[4. 实施计划 5](#_Toc445920436)

[4.1 工作流程 5](#_Toc445920437)

[4.2 总体进度计划 5](#_Toc445920438)

[4.3 项目控制计划 6](#_Toc445920439)

[5. 关键问题 6](#_Toc445920440)

[6. 专题计划要点 6](#_Toc445920441)

# 引言

## 背景

近几年来，随着计算机和信息技术的迅猛发展和普及应用，行业应用系统的规模迅速扩大，行业应用所产生的数据呈爆炸性增长。动辄达到数百TB甚至数十至数百 PB规模的行业/企业大数据已远远超出了现有传统的计算技术和信息系统的处理能力，因此，寻求有效的大数据处理技术、方法和手段已经成为现实世界的迫切需 求。百度目前的总数据量已超过1000PB，每天需要处理的网页数据达到10PB～100PB;淘宝累计的交易数据量高达100PB;Twitter每天 发布超过2亿条消息，新浪微博每天发帖量达到8000万条;中国移动一个省的电话通联记录数据每月可达0.5PB～1PB;一个省会城市公安局道路车辆监 控数据三年可达200亿条、总量120TB。据世界权威IT信息咨询分析公司IDC研究报告预测：全世界数据量未来10年将从2009年的0.8ZB增长 到2020年的35ZB(1ZB=1000EB=1000000PB)，10年将增长44倍，年均增长40%。

由于大数据处理需求的迫切性和重要性，近年来大数据技术已经在全球学术界、工业界和各国政府得到高度关注和重视，全球掀起了一个可与20世纪90年代的信 息高速公路相提并论的研究热潮。美国和欧洲一些发达国家政府都从国家科技战略层面提出了一系列的大数据技术研发计划，以推动政府机构、重大行业、学术界和 工业界对大数据技术的探索研究和应用。

当前主要是通过分布式计算框架来处理海量数据。当前主流的分布式计算框架有MapReduce,Storm,Spark,Flink等。这些分布式框架的理念不一而同但又各有特色和专注领域。此次我们的关注点是Apache Spark这一款流式实时分布式计算框架的研究与应用，运用软件工程的方法对Spark进行分析，并在此基础上，将Spark扩展应用。

## 定义

关于Spark项目中的一些专业术语、缩略语如下表1所示。

|  |  |
| --- | --- |
| 术语/缩略语 | 描述 |
| Spark-shell | Spark-shell是一个应用程序，因为spark-shell在启动的时候创建了SparkContext对象。 |
| Application | Application是创建了SparkContext实例对象的Spark用户，包含了Driver程序。 |
| Job | 用户程序在Spark中的运行实体。 |
| Driver Program | 运行main函数并且新建SparkContext实例的程序。 |
| Cluster Manager | 集群资源管理的外部服务，在Spark上现在主要有Standalone、Yarn、Mesos等三种集群资源管理器，Spark自带的 Standalone模式能够满足绝大部分纯粹的Spark计算环境中对集群资源管理的需求，基本上只有在集群中运行多套计算框架的时候才建议考虑 Yarn和Mesos。 |
| Worker Node | 集群中可以运行应用程序代码的工作节点，相当于Hadoop的slave节点。 |
| Executor | 在一个Worker Node上为应用启动的工作进程，在进程中负责任务的运行，并且负责将数据存放在内存或磁盘上，必须注意的是，每个应用在一个Worker Node上只会有一个Executor，在Executor内部通过多线程的方式并发处理应用的任务。 |
| Task | 被Driver送到executor上的工作单元，通常情况下一个task会处理一个split的数据，每个split一般就是一个Block块的大小。 |
| Stage | 一个Job会被拆分成很多任务，每一组任务被成为Stage，这个MapReduce的map和reduce任务很像，划分Stage的依据在 于：Stage开始一般是由于读取外部数据或者Shuffle数据、一个Stage的结束一般是由于发生Shuffle（例如reduceByKey操 作）或者整个Job结束时例如要把数据放到hdfs等存储系统上。 |

表 1 专业术语及缩略语

## 参考资料

[1] Apache Spark官方文档 <http://spark.apache.org>

## 标准、条约和约定

无。

# 项目概述

## 项目目标

从Spark开源项目出发，重新编写起需求规格说明书、测试需求规格说明书，并根据需求规格说明书与测试需求规格说明书对Spark进行模块化测试，完成本学期软件工程的8个实验，并在此基础上拓展应用。

## 产品目标与范围

编写Spark的需求规格说明书，为Spark的各个模块编写测试需求规格说明书、测试用例，并以此对Spark进行测试。

## 假设与约定

1. 假设

小组成员都能完成“《流式实时分布式并行计算框架Spark的研究与应用》项目进度表V1版”中的具体任务。

1. 约定

小组成员必须根据“《流式实时分布式并行计算框架Spark的研究与应用》项目进度表V1版”中分配的任务按时完成。

小组成员根据项目进度定期开头脑风暴会议，认真记录会议内容。

## 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 | 2016/3/16 | 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 | 软件需求规格说明书 | 2016/3/24 | 明确、详细 | 暂无 |
| No.5 | 需求检查单 | 2016/3/30 | 详细 | 暂无 |
| No.6 | 测试需求规格说明书 | 2016/4/6 | 完整、详细 | 暂无 |
| No.7 | 软件测试分析报告 | 2016/4/13 | 详尽 | 暂无 |
| No.8 | 软件测试评审报告 | 2016/4/20 | 充分 | 暂无 |
| No.9 | 项目工作量统计报告 | 2016/4/27 | 详细、充分 | 暂无 |
| No.10 | 项目方法总结报告 | 2016/5/4 | 完整、详细 | 暂无 |
| No.11 | 项目个人总结 | 2016/5/11 | 充分、详实 | 暂无 |

表 2 应交付成果列表

## 项目开发环境

此次项目软硬件开发环境如下：

1. 软件环境

Windows 7/Windows 10/CentOS7/Ubuntu 22

1. 硬件环境

4G内存 + 320G硬盘 + 百兆网卡 + 百兆局域网

1. 开发工具

Eclipse4.x + Maven 4.x + JDK1.8

## 项目验收方式和依据

1. 验收方式

项目验收方式采用中期课堂评审和终期交付验收。

1. 验收依据

验收依据“应交付成果”清单进行验收。

# 项目团队组织

## 组织结构

此次项目有两个角色：组员和组长。组长与组员沟通协调，安排具体事宜、分配相应的任务。而组员则需要根据组长的安排要求，及时完成任务。

## 人员分工

人员具体分工如下表3所示。

|  |  |  |
| --- | --- | --- |
| 姓名 | 角色 | 工作描述 |
| 阳艳红 | 组员 | 详见《项目进度表V1版》 |
| 于思民 | 组员、组长 | 详见《项目进度表V1版》 |
| 王铖成 | 组员 | 详见《项目进度表V1版》 |
| 武一杰 | 组员 | 详见《项目进度表V1版》 |

表 3 人员分工

## 协作与沟通

小组成员通过微信群、电话、邮件等方式进行及时沟通，在项目有重大决定时，可举行临时面谈会议。而且小组通过文档、Git工具协作完成相应的具体任务。

沟通和协作都要记录具体内容。

# 实施计划

## 工作流程

具体工作流程参照“《流式实时分布式并行计算框架Spark的研究与应用》项目进度表V1版”。

## 总体进度计划

具体工作流程参照“《流式实时分布式并行计算框架Spark的研究与应用》项目进度表V1版”。

## 项目控制计划

暂无。

# 关键问题

暂无。

# 专题计划要点

暂无。