**Spark Streaming的分析与应用**

**项目计划**

Version 1.2

小组成员：

陈少杰

姜鑫

蒲彦均

邹嘉欣

**版本变更记录**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 版本 | 变更时间 | 修改人 | 审核人 | 备注 |
| 1.0 | 2017/03/16 | 姜鑫 | 陈少杰 蒲彦均邹嘉欣 | 初稿 |
| 1.1 | 2017/03/21 | 邹嘉欣 | 陈少杰 蒲彦均  姜鑫 | 二稿 |
| 1.2 | 2017/04/25 | 陈少杰 | 姜鑫、邹嘉欣、蒲彦均 | 三稿 |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

目录

[1. 前言 4](#_Toc481009519)

[1.1目的 4](#_Toc481009520)

[1.2术语与缩略语 4](#_Toc481009521)

[2. 项目概述 5](#_Toc481009522)

[2.1简介 5](#_Toc481009523)

[2.2 Spark的历史 6](#_Toc481009524)

[3. 实施计划 7](#_Toc481009525)

[3.1工作内容分解 7](#_Toc481009526)

[3.2参加人员 7](#_Toc481009527)

[3.3完成项目最后期限 7](#_Toc481009528)

[3.4本计划的批准者与批准日期 7](#_Toc481009529)

[3.5项目分工 7](#_Toc481009530)

[3.6组织形式 8](#_Toc481009531)

[3.7交付的产品 8](#_Toc481009532)

[3.8进度安排 9](#_Toc481009533)

[3.9时间预算 9](#_Toc481009534)

[3.10遵循标准 9](#_Toc481009535)

[3.11硬件环境 9](#_Toc481009536)

[3.12软件环境 9](#_Toc481009537)

[3.13任务分配方案与工时计算 10](#_Toc481009538)

[参考资料 10](#_Toc481009539)

# 前言

### 1.1目的

为了便于协调组内成员进行后期的工作，对项目进行跟踪和监控，对任务的进度进行安排与调控，故对后期工作进行计划。

### 1.2术语与缩略语

|  |  |
| --- | --- |
| SparkConf | 包含了集群配置的一系列参数，包括master节点，app名称等。 |
| SparkContext | 所有Spark程序的入口，相当于main函数。可以完成变量广播，RDD的操作等。SparkContext的构建需要SparkConf。 |
| DStream | DStream是Spark的一个基本抽象概念，它代表了一个连续的数据流，数据流按照时间间隔进行切分，会产生一系列的RDD。所以，DStream也可以看作是RDD的序列。 |
| RDD | RDD(Resilient Distributed Datasets)弹性分布式数据集，是分布式内存的一个抽象概念，本质上是一个只读的分区记录集合。 |
| filter | Spark Streaming主要适用于流式的数据处理，而输入流中的数据并不是全部都需要处理，这时便可以使用filter进行数据过滤，减少集群计算压力。 |
| map/mapPartition/flatMap等 | map系列的函数是基于一个RDD中的元素来产生一个新的RDD的方式。其具体的用法，区别，性能等也是本项目研究的方向之一。 |

# 项目概述

### 2.1简介

Apache Spark是一个围绕速度、易用性和复杂分析构建的大数据处理框架。最初在2009年由加州大学伯克利分校的AMP-Lab开发，并于2010年成为Apache的开源项目之一。

Spark是一个类似于Hadoop的MapReduce的分布式计算框架，其核心是弹性分布式数据集，提供了比MapReduce更丰富的模型，可以在快速在内存中对数据集进行多次迭代，以支持复杂的数据挖掘算法和图形计算算法。

Spark Streaming是建立在Spark上的实时计算框架，通过它提供的丰富的API、基于内存的高速执行引擎，用户可以结合流式、批处理和交互式查询应用，扩展了Spark处理大规模流式数据的能力。图1展示了Spark与其配套的Spark SQL, Spark Streaming, MLib和GraphX等上层模块构成的生态系统。



图 1 Spark生态系统

Spark Streaming的基本原理是将输入数据流以时间片（秒级）为单位进行拆分，然后以类似批处理的方式处理每个时间片数据。

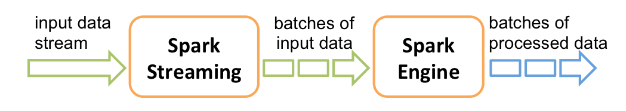


图 2 Spark Streaming工作基本原理

### 2.2 Spark的历史

Spark起初只是一个学术性研究项目，从创立到如今蜚声海外并且在大数据领域成为风口浪尖的热门项目只花了6年左右的时间，其具体发展大事记如下。

* 2009年Spark诞生于伯克利AMPLab。
* 项目在2010年早些时候开源，很多早期关于Spark系统的思想在不同论文中发表。
* 项目开源之后，在GitHub上成立了Spark开发社区并在2013年成为Apache孵化项目。
* 该项目在2014年2月成为Apache顶级项目。
* 2014年5月30日Spark 1.0.0版正式上线。
* 截止到2015年，Spark官方维护运营公司Databricks已经组织并举办了三年Spark Summit技术峰会。

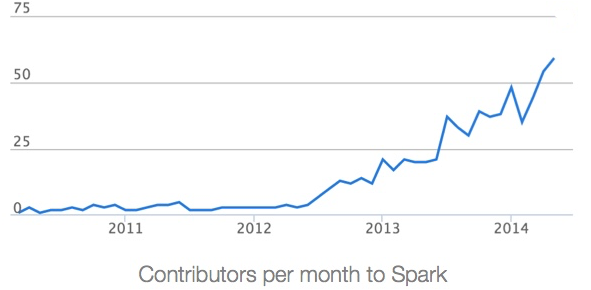


图 3 Spark热度变化趋势

# 实施计划

### 3.1工作内容分解

首先，了解Spark以及Spark Streaming的相关基础概念及其工作流程，并进行集群环境搭建。

然后，了解其不同的数据处理方式的使用方法和应用场景，分析其功能与实际需求的对应关系。

最后，针对项目探究过程中发现的问题与兴趣点，进行针对性的深入研究，并针对Spark Streaming进行上层应用的开发。

### 3.2参加人员

陈少杰 姜鑫 蒲彦均 邹嘉欣

### 3.3完成项目最后期限

校历第十七周

### 3.4本计划的批准者与批准日期

刘超老师与任建老师

校历第三周

### 3.5项目分工

初期工作：

首先进行环境的搭建，项目组成员分别搭建一个3-5个节点的spark集群（可通过虚拟机搭建），spark版本选择1.6及以上。

姜鑫负责搜集hadoop+spark相关介绍以及集群安装指导资料，并协助其他成员完成环境搭建。

陈少杰完成项目的计划制定工作，包括在Microsoft project上进行计划的安排，以及会议内容的记录。

蒲彦均负责github的使用介绍以及相关管理工作，并指导其他成员使用GitHub。

邹嘉欣负责PPT的制作与答辩相关工作。

中期工作：

项目组成员均需学习Spark Streaming的概念，使用，以及特性等，在对Spark和Spark Streaming有了初步的认识并能够完成简单程序的编写的基础上，分别着重对某一块的内容进行深入的研究学习，分享学习资料，笔记，并组织例会进行交流探讨。

后期工作：

在对Spark Streaming有了较为深入的了解后，基于其进行上层应用的开发。姜鑫与陈少杰负责后台开发，蒲彦均进行前台的开发，邹嘉欣负责测试工作。计划基于机器学习、自然语言处理等技术实现指定主题短文本的自动识别功能，本软件产品是一项独立的软件，用户根据软件提供的调用命令实现各种指定功能，达到在十台机器的spark集群上每分钟5万条处理速度的性能指标。

### 3.6组织形式

a)每周一次会议，讨论当前工作进展以及各自的问题。

b)建微信群，保持实时的交流。

c)学习资料与代码的共享通过github实现。

### 3.7交付的产品

|  |  |  |
| --- | --- | --- |
| **序号** | **产品名称** | **交付时间** |
| 1 | 软件项目计划书 | 第三周 |
| 2 | 需求规格说明书 | 第四周 |
| 3 | 项目计划mpp文件 | 每周更新 |
| 4 | 每周会议记录 | 每周更新 |
| 5 | 需求检查单 | 第八周 |
| 6 | 软件评审报告 | 待定 |
| 7 | 软件产品改进与展示 | 第十一周 |
| 8 | 软件测试分析报告 | 第十二周 |
| 9 | 软件测试评审报告 | 待定 |
| 10 | 软件进度分析报告 | 待定 |
| 11 | 工作量分析报告 | 每周更新 |
| 12 | 配置管理 | 每周更新 |

### 3.8进度安排

项目进度总体上按照老师的进度安排进行，具体的进度计划参见项目计划mpp文件。

### 3.9时间预算

小组成员每周开会两次约2人时，额外的用于本项目的时间约为15小时。项目从第二周开始到第十七周共有16周的时间。总工作量为960人时。

### 3.10遵循标准

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 9386-2008 计算机测试文档编制

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 8567-2006 计算机软件文档编制规范

### 3.11硬件环境

小组每名成员在实验室都有自己的工作电脑，使用实验室的硬件设备即可进行代码的阅读、开发、测试、ppt的制作演示等等。

### 3.12软件环境

|  |  |
| --- | --- |
| 操作系统 | Windows7及以上 |
| 开发环境 | eclipse |
| 项目管理 | Github |
| 虚拟机系统 | centos 6.5及以上 |
| Hadoop | 2.2及以上 |
| Spark | 1.6及以上 |

### 3.13任务分配方案与工时计算

每周周初小组召开例会，根据老师的总体进度安排，确定本周具体任务。根据组员实现对系统了解程度等因素，对任务进行分工，大致要做到小组成员的工作量大致均等。任务分工上有些侧重点：陈少杰负责总体管控，包括具体计划执行，会议记录，mpp文件操作等工作；邹嘉欣侧重于文档编写、修改；姜鑫、蒲彦均侧重于项目工作的设计与实现部分。

在实验开始阶段对项目整体任务，以周为单位建立一级任务，并对每个一级任务进行粗略的细化。在每周周初开会确定本周任务分工后，对本周对应的一级任务进行细化。

在工时计算方面，在细化mpp任务时，对每部分工时进行估算。在每周结束时，大家各自写一周工作日志，大家各自对小组其他成员的工时进行评估，以保证公式计算的准确性。

# 参考资料

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

[2] <http://spark.apache.org/docs/latest/streaming-programming-guide.html>