|  |
| --- |
|  |
| 项目概要说明 |
| 基于Spark的分析与应用 |
|  |
| SY1606117 李岳檑  SY1606118 温元祯  SY1606413 谭伟良  PT1600283 王春柳 |

|  |
| --- |
|  |

项目概要说明

目录

[1. 项目选择理由 1](#_Toc477363755)

[2. 项目简介 1](#_Toc477363756)

[3. 项目特点 3](#_Toc477363757)

[3.1更高的性能 3](#_Toc477363758)

[3.2易用性好 3](#_Toc477363759)

[3.3通用性强 3](#_Toc477363760)

[3.4支持多种资源管理器 3](#_Toc477363761)

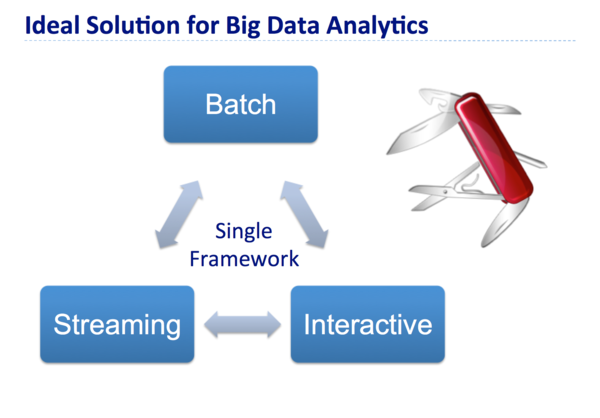
[4. 项目熟悉程度 3](#_Toc477363762)

1. 项目选择理由

时下，我们正处在一个“大数据”的时代，每时每刻，都有各种类型的数据被生产。此外，数据增幅的速度也在显著增加。从广义上看，这些数据包含交易数据、社交媒体内容（比如文本、图像和视频）以及传感器数据。而对这些海量数据进行分析和挖掘，其原因就是为了从中发现规律来更好的指导生活和生产实践。

雅虎2009年开源的Apache Hadoop通过使用商用服务器组成的集群大幅度地降低了海量数据处理的门槛。Hadoop主要提供了两个方面的功能：通过水平扩展商用主机，HDFS提供了一个廉价的方式对海量数据进行容错存储；MapReduce计算范例，提供了一个简单的编程模型来挖掘数据并获得洞见。

历经数年发展，Hadoop生态圈中的丰富工具已深受用户喜爱，然而仍然存在众多问题给使用带来了挑战。如每个用例都需要多个不同的技术堆栈来支撑，在不同使用场景下，大量的解决方案往往捉襟见肘；许多技术存在版本兼容性问题；无法在并行job中更快地共享数据等。而这些问题都在Spark中得到了解决。Apache Spark是一个轻量级的内存集群计算平台，通过不同的组件来支撑批、流和交互式用例。如下图。



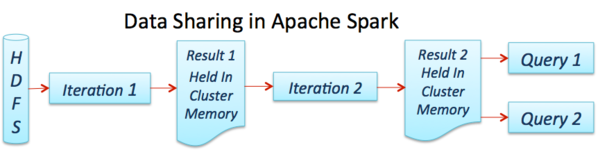
1. 项目简介

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在2013年6月进入Apache成为孵化项目，8个月后成为Apache顶级项目，速度之快足见其过人之处。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。按照这样的发展速度，Spark也希望能代替hadoop在大数据中的地位，成为大数据处理的主流标准。

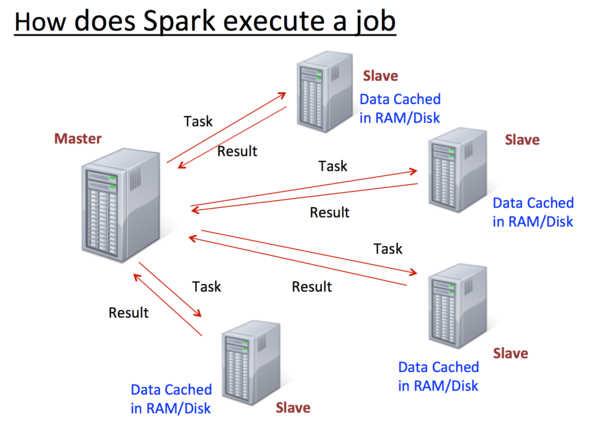
Spark使用Scala语言实现，Scala是一种面向对象、函数式编程语言，能够像操作本地集合一样轻松的操作分布式数据集。Scala把Erlan风格的基于actor的并发带进了JVM。开发者可以利用Scala的actor模型在JVM上设计具伸缩性的并发应用程序，它会自动获得多核心处理器带来的优势，而不必依照复杂的Java线程模型来编写程序。Scala具有运行速度快、易用性好、通用性强等优点。

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，下图显示了一个典型Spark job的处理流程。



下图显示了Apache Spark如何在集群中执行一个作业。



Master控制数据如何被分割，利用了数据本地性，并在Slaves上跟踪所有分布式计算。在某个Slave不可用时，其存储的数据会分配给其他可用的Slaves。

1. 项目特点

## 3.1更高的性能

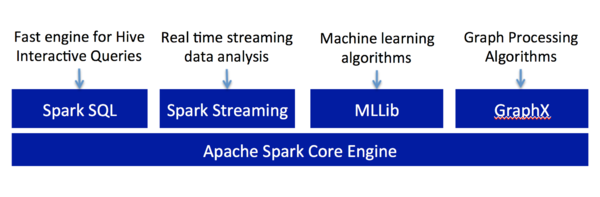
在Spark中，数据被加载到集群主机的分布式内存中。数据可以被快速的转换迭代，并缓存用以后续的频繁访问需求。在数据全部加载到内存的情况下，Spark可以比Hadoop快100倍，在内存不够存放所有数据的情况下快Hadoop 10倍。

## 3.2易用性好

Spark不仅支持Scale编写应用程序，而且支持Java和Python语言编写，建立了Java、Scala、Python、SQL（应对交互式查询）的标准API以方便使用，同时还含有大量开箱即用的机器学习库。

## 3.3通用性强

Spark生态圈包含了Spark Core、Spark SQL、Spark Streaming、MLLib和GraphX等组件。Spark Core提供内存计算框架，Spark Streaming实时处理应用，Spark SQL可即时查询，MLLib用于机器学习，GraphX做图像处理，如下图。 开发者可以在同一个应用程序中无缝组合使用这些库。



## 3.4支持多种资源管理器

Spark 支持 Hadoop YARN，Apache Mesos，及其自带的独立集群管理器，且与现有Hadoop v1 (SIMR) 和2.x (YARN) 生态兼容，因此机构可以进行无缝迁移。

1. 项目熟悉程度

项目成员均未接触过与大数据相关的技术，所有成员零基础开始学习和实践。