浙 江 大 学

硕士学位论文开题报告

**（专业学位）**

论文题目： GenBase on Spark—Spark性能评测系统的实现与优化

姓 名： 袁博

学 号： 21451031

专 业： 软件工程

院 别： 软件学院

导 师： 张泉方

二零一五 年 十一月十号

目录

[1.课题来源及类型 3](#_Toc435609145)

[2.课题的意义及国内外现状分析 4](#_Toc435609146)

[2.1大数据应用的研究概况 4](#_Toc435609147)

[2.2数据处理系统概况 5](#_Toc435609148)

[2.3数据处理系统评测的意义 8](#_Toc435609149)

[3. 课题的研究目标、研究内容和拟解决的关键问题 9](#_Toc435609150)

[3.1 课题研究目标 9](#_Toc435609151)

[3.2课题研究内容 9](#_Toc435609152)

[3.3 拟解决的关键问题 10](#_Toc435609153)

[4. 课题的研究方法、设计及试验方案，可行性分析 11](#_Toc435609154)

[4.1课题的研究方法 11](#_Toc435609155)

[4.2课题设计方案 11](#_Toc435609156)

[4.3课题可行性分析 12](#_Toc435609157)

[5.课题计划进度和预期成果 13](#_Toc435609158)

[5.1计划进度 13](#_Toc435609159)

[5.2预期成果 13](#_Toc435609160)

[5.2.1 Spark数据处理系统架构分析 13](#_Toc435609161)

[5.2.2 Spark性能评测系统建设 14](#_Toc435609162)

[5.2.3不同数据规模下性能测试 14](#_Toc435609163)

# 1.课题来源及类型

目前，用于分析数据，处理数据的数据处理系统有很多，包括基于HDFS存储的Hadoop和基于内存的流式处理系统Spark等等。在不同的应用场景，根据不同的性能要求需要在这些数据处理系统之间进行选择。这时，我们就需要一套评测系统数据处理能力的标准，并基于此标准构建数据处理系统的评测系统。我在intel实习期间的工作是关于Spark的性能测试和优化。因此，基于数据处理能力评测标准，本文着重研究如何构建Spark性能评测系统。

# 2.课题的意义及国内外现状分析

## 2.1大数据应用的研究概况

数据从计算机时代以前的纸张记账到现在的大型数据库记录都毫无疑问极大地影响着人类的社会生活。政府从统计数据和总结趋势中了解社会经济的发展，制定相关的政策法律；企业从市场调研和定期财报中调整发展战略或及时推出符合市场需求的产品；医疗研究机构通过分析往年疫情的发展状况为当前具有苗头的疾病扩散做好提前防疫的准备。随着移动互联网的发展和各类电子终端的普及，新时代数据产生的源头越来越多，种类也各种各样。股票市场每天产生成千上万条交易记录；一款流行的移动手机应用会有千万甚至上亿的装机量，每天产生的用户使用数据就将超过百亿条；可穿戴式设备的兴起将再一次拓展用户产生的信息量，为医学监测、用户行为分析带来新的挑战与机遇。近年来，大数据已经由技术圈逐渐进入主流市场并取得了比较显著的经济和社会效益。早在2012年3月，美国联邦政府就公布了一项耗资2亿美元的大数据计算研究计划。同年，中国政府批复了“十二五国家政务信息化建设工程规划”，利用百亿的投资建设涵盖社会各方面的数据资源库工程。大数据时代的前景是美好的。借助政府数据开放,美国的医疗服务业节省了3000亿美元, 制造业在产品开发、组装等环节节省了50%的成本。

由于技术的原生性与直接相关性，大数据技术与互联网产业的联系天生就比较密切。 目前国内外的很多大型互联网企业都有专门的大数据分析研发团队，比如英特尔 (Intel) 的大数据部 (SSG-BDT), 淘宝技术部, 腾讯数据平台部，百度大数据部。其中百度甚至将人工智能和机器学习领域的国际权威之一吴恩达(Andrew Ng)纳入麾下。淘宝作为中国最大的电商，利用大数据来进行个性推荐、打击商户盗刷信誉和满意度等作弊行为，以此净化交易环境，提高用户体验。

相比互联网企业，医疗和金融领域同样能够极大受益于大数据分析，但目前在这两个领域的技术开发，尤其是医疗方面，落后于互联网领域。很多大规模数据挖掘和分析的算法都能够针对不同的使用场景进行优化，但早已应用在互联网数据分析上的聚类算法还没有广泛应用于医疗和金融领域。在医疗和金融领域广泛使用的传统统计分析方法也没有针对大规模的数据作实现和优化。这两个领域的很多企业仍然在使用微软办公套件EXCEL或者单机的算法工具包比如R、Weka和Scikit-learn等并结合适当的采样算法进行数据统计分析或者机器学习。

## 2.2数据处理系统概况

（1） MPI

MPI( Message Passing Interface) 是1994年5月发布的一套消息传递接口。它是一个消息传递函数库的标准说明，集多个消息传递系统的优点于一身，是目前国际上最流行的并行编程环境之一。在基于MPI的编程模型中，计算是由进程间的消息通信所组成。程序员通过调用MPI库程序来达到并行的目的。MPI提供C语言和Fortran语言程序接口。

MPI程序属于进程级的并行粒度，具有可移植性和易用性，异步通信功能完备，接口定义正式而精确。但是其编程模型比较复杂，需要将任务映射到分布式进程集合中去计算，且负载分配不局限于任务划分和数据划分。这样的强灵活性造成了强复杂性。同时，MPI程序经常需要解决通信延迟和负载不平衡问题，程序调试起来比较耗时。 MPI程序的容错性较低，一个进程的出错就会导致整个程序无法运行。

（2） MapReduce

MapReduce是谷歌(Google)公司在2004年公开的一种大规模并行计算编程模型。MapReduce的灵感来源于函数式语言 (如Lisp) 中的内置函数map和reduce。这种编程模型主要可以理解为两个阶段。首先映射 (map) 阶段把一堆杂乱的数据按照某种特征进行归纳，期间可能发生洗牌 (shuffle) 动作, 以便把具有相同特征的、存储在不同服务器中的数据转移到一起。归纳的结果以键值对的形式输出给下一个归约 (Reduce) 阶段。在Reduce阶段程序进一步处理得到总结以后的分析结果。整个流程的框图如图1.1所示。MapReduce对程序员隐藏了并行式程序的底层调度细节，极大方便了大规模数据处理和算法开发。

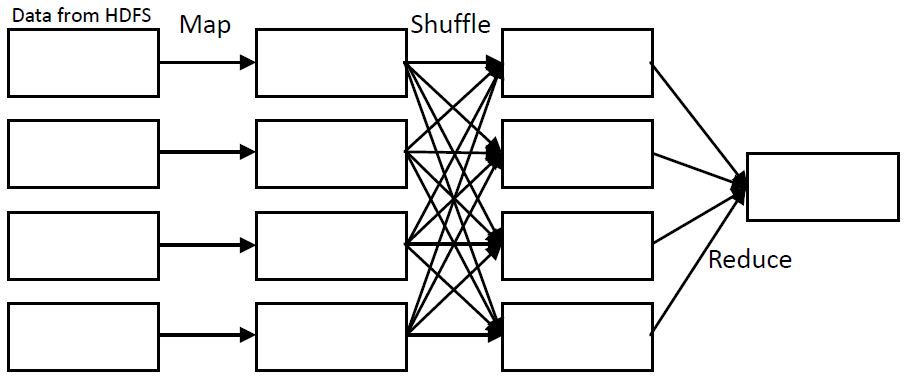


图 1-1 MapReduce思想示意图

MapReduce为了降低复杂性，比MPI引入了更多规则，同时增强了容错性。配合分布式文件存储系统，MapReduce希望尽量计算本地存储的数据，仅当节点无法处理本地数据时再采用就近原则寻找其他可用的计算节点。因此，它的计算引擎需要配套一个分布式文件存储系统。在容错性上，MapReduce使用了多种有效的错误检测和恢复机制。集群不会因为单节点的失效而影响计算质量。

目前最流行的实现了MapReduce思想的开源计算平台是Apache Hadoop，其配套的分布式存储系统是HDFS (Hadoop Distributed File System)。

（3）Spark

Spark 是2010年由加州大学伯克利分校的AmpLab实验室研发的一种新型并行内存计算框架。2014年已经发展成为Apache开源社区最活跃的项目，被越来越多地应用在工业和学术领域。Spark的计算逻辑范式和MapReduce一脉相承。但它主要通过将基本计算操作都在内存中完成来解决Hadoop架构中每一次映射和归约过程都必须读写磁盘，导致计算速度很慢的问题。此外，Spark提出了一种叫做弹性分布式数据集 (RDD : Resilient Distributed Datasets) 的数据抽象，使得内存计算的底层实现细节对用户隐藏， 并以有向无环图 (DAG) 的形式表示RDD的各个世代 (Lineage), 以达到CPU负载和容错性之间的平衡。基于函数式编程语言Scala，Spark的编程范式显得格外简洁和优雅。在实际应用中，Spark被证实能够相比Hadoop取得10 – 100 倍的速度提升，对需要多次迭代的应用场景尤为有效。

除了高效的数据处理框架外，Spark项目还致力于构建一个完整的大数据生态环境。除了计算核心外，项目还包括Spark Streaming，Spark SQL，MLlib，Graphx等组件，分别对应处理流式数据、调用SQL语句、机器学习、大规模图算法的应用。Spark还支持除原生的Scala外的Java，Python，R等多种程序设计语言，极大地扩展了Spark的应用场景，便利了不同背景的程序员进行Spark应用开发。

## 2.3 Spark性能评测的意义

近年来，大数据已经由技术圈逐渐进入主流市场并取得了比较显著的经济和社会效益。应用场景对于实时性的要求也越来越高。移动互联网行业中，一个更为高效的数据处理系统能够帮助企业优先于竞争对手给出用户需要的数据，从而在竞争中占据先机。政府在面对应急情况时，需要获得最新最全面的信息来做出正确的决策。 Spark凭借其内存计算的出色架构在实时性方面占有优势，越来越多的应用场景将会用到Spark作为其解决方案。但是，内存计算依赖于内存的容量的大小，面对大规模的数据容易产生性能的瓶颈。在不同规模下对于Spark的性能进行评测能够使我们了解Spark的性能瓶颈，通过优化可以知道如何最大限度地利用Spark的计算能力。这样，当我们遇到复杂的应用场景时，可以判断Spark是否适用于该场景，并在使用Spark时能够更好地面对不同的数据规模和不同的需求。

# 3. 课题的研究目标、研究内容和拟解决的关键问题

## 3.1 课题研究目标

数据处理系统对于数据处理的过程分成数据预处理、数据分析、数据处理三部分。本文研究如何评测Spark性能，从数据预处理、数据分析、数据处理三部分着手，搭建完整的处理流程，通过整个流程的执行结果评测Spark的性能。

研究目标是：结合基因组数据分析的流程，提出基于Spark的数据分析流程Trumpet，按照数据预处理、数据分析、数据处理三部分完成对于基因组数据的分析。并围绕此流程构建Spark性能评测系统，按照不同的数据规模评测Spark性能。

## 3.2课题研究内容

本文针对Spark数据处理系统的架构和特点，从数据预处理、数据分析、数据处理三方面对Spark的性能进行评测。根据基因组数据分析的流程，提出基于Spark的数据分析流程Trumpet，按照数据预处理、数据分析、数据处理三部分完成对于基因组数据的分析。接着围绕此流程构建Spark性能评测系统，并按照不同的数据规模评测Spark性能。

本文主要研究内容包括：

首先，分析Spark数据处理系统的架构和特点，理解弹性分布式数据集 (RDD : Resilient Distributed Datasets) 的数据抽象，以及内存计算的过程。

其次，根据基因组数据分析的流程，提出基于Spark的数据分析流程Trumpet，按照数据预处理、数据分析、数据处理三部分完成对于基因组数据的分析。

最后，围绕此流程构建Spark性能评测系统，并按照不同的数据规模对Spark的性能进行评测，并在此基础上对于现有的评测系统进行优化。

## 3.3 拟解决的关键问题

对于Spark数据处理系统的架构和特点的分析，我们主要从弹性分布式数据集 (RDD : Resilient Distributed Datasets) 的数据抽象，以及内存计算的过程两方面入手。这两方面是Spark内存计算的核心，也是Spark性能提升的关键之处。

对于Spark的评测系统在设计上需要达到两方面的要求，首先需要具备大数据建模与高仿真的数据生成的能力。其次，数据处理系统评测基准还需要具备能够满足多场景需求的综合负载生成能力。

同时，Spark性能评测系统需要考虑到不同数据规模对于Spark的性能影响，我们需要保证该系统在大规模的数据输入情况下能够稳定地运行。

# 4. 课题的研究方法、设计及试验方案，可行性分析

## 4.1课题的研究方法

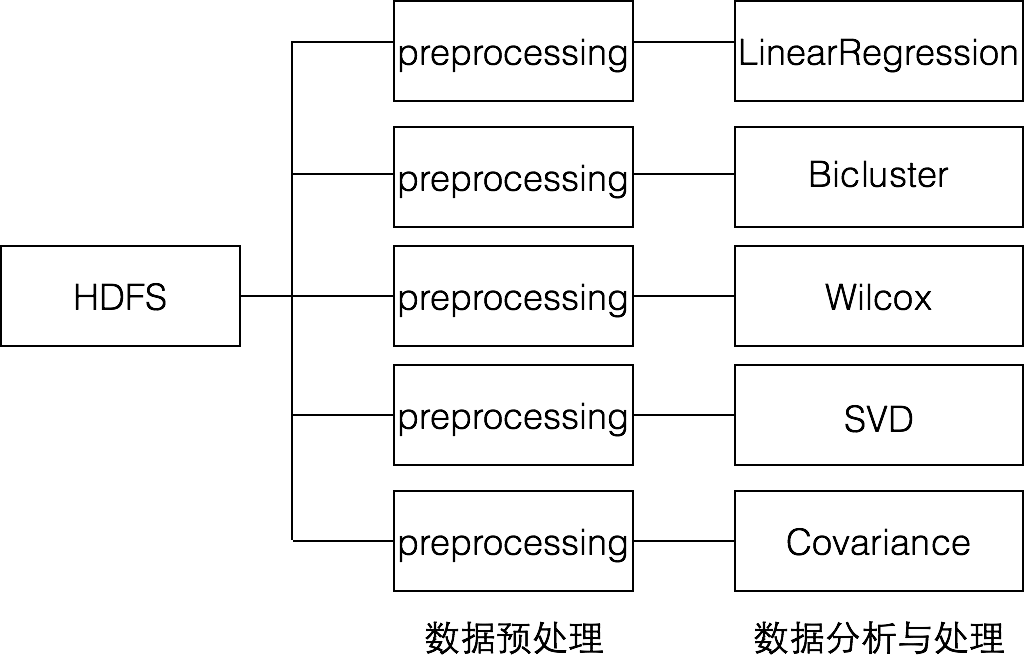
大数据分析平台的性能指标分为以下七个方面：

1. 数据导入（条每秒）
2. 数据导出（条每秒）
3. 处理速度（MB/S）
4. 磁盘读写（MB/S）
5. 存储容量（TB）
6. 响应时间（秒）
7. 查询速度（毫秒）

本文主要采用数据导入、处理速度、查询速度三大性能指标作为Spark性能评测的方法。评测时分成数据预处理、数据分析、数据处理三部分进行评测。其中数据预处理对应数据导入和查询速度两部分指标，数据分析对应处理速度指标，数据处理对应处理速度和查询速度两部分指标。通过不同的数据规模下，该系统在Spark平台上三部分的运行时间可对Spark的性能进行评测。

## 4.2课题设计方案

首先对于Spark数据处理系统的架构和特点的进行分析，从弹性分布式数据集 (RDD : Resilient Distributed Datasets) 的数据抽象，以及内存计算的过程两方面了解Spark处理数据的流程。然后，根据基因组数据分析的流程，提出基于Spark的基因数据分析流程Trumpet，按照数据预处理、数据分析、数据处理三部分完成对于基因组数据的分析。最后，围绕此流程构建Spark性能评测系统，并按照不同的数据规模对Spark的性能进行评测，并在此基础上对于现有的评测系统进行优化。



**图4.1 Trumpet系统流程图**

## 4.3课题可行性分析

基于以上分析结果，我们认为本文的研究方法和设计方案切实可行，可操作性高，研究结果也具有很高的实用性，课题具有一定的研究价值。

# 5.课题计划进度和预期成果

## 5.1计划进度

|  |  |  |
| --- | --- | --- |
| **开始时间** | **结束时间** | **主要工作内容** |
| 2015年10月 | 2015年11月 | 查阅文献资料，编写开题报告 |
| 2015年11月 | 2015年12月 | Spark数据处理系统架构分析 |
| 2015年12月 | 2015年1月 | 构建Trumpet流程 |
| 2015年1月 | 2015年2月 | 构建评测系统，在不同规模下对Spark性能进行测试 |
| 2015年2月 | 2015年3月 | 撰写论文正文 |

## 5.2预期成果

总结全文，本文预期取得的主要研究成果如下：

### 5.2.1 Spark数据处理系统架构分析

对于Spark数据处理系统的架构和特点的分析进行，从弹性分布式数据集 (RDD : Resilient Distributed Datasets) 的数据抽象，以及内存计算的过程两方面了解Spark处理数据的流程。

### 5.2.2 Spark性能评测系统建设

参考基因组数据分析的流程，提出了基于Spark的基因组数据分析流程Trumpet并围绕此系统构建Spark性能评测系统。并分成数据预处理、数据分析、数据处理三大部分进行评测。其中数据预处理对应数据导入和查询速度两部分指标，数据分析对应处理速度指标，数据处理对应处理速度和查询速度两部分指标。通过不同的数据规模下，该系统在Spark平台上三部分的运行时间对Spark的性能进行评测。

### 5.2.3不同数据规模下性能测试

分成不同的数据规模在Spark上运行性能评测系统，得到不同数据规模下Spark的性能测试结果。并在此过程中不断优化评测系统，使该系统能够承载足够大的数据规模。