基于 Spark 可视化大数据挖掘平台

李 文[1,2] 程华良[4] 彭 耀[1,2] 温明杰[1,2] 肖威清[1,2] 张陈斌[1,3] 陈宗海[1,3]

(1. 中国科大-象形大数据商业智能联合实验室，安徽合肥，中国，230031；2. 安徽象形信息科技有限公司，安徽合肥，中国，230031)

(3. 中国科学技术大学自动化系，安徽合肥，中国，230027；4. 安徽中烟工业有限责任公司合肥卷烟厂，安徽合肥，中国，230027)

摘 要：在千亿级大数据环境下，特征挖掘、实时处理、即席分析、离线计算等场景对计算、存储的性能要求非常高。基于传统的关系型数据库、分布式 Hadoop 平台实现的数据挖掘平台，无法满足所有的计算场景的要求。鉴于此，本文介绍了基于内存迭代计算框架 Spark，实现大数据环境下的可视化大数据挖掘平台。该平台不仅充分利用了内存计算，提高了迭代速度，而且支持各种分布式计算、存储场景，具有很强的扩展性，解决了大数据环境下各种计算场景问题。

关键词：Spark；内存计算；大数据；数据挖掘

中图分类号：TP391

Visualized Data Mining Platform Based on the Spark

Li Wen[1,2] *Cheng Hua-liang*[4] *Peng Yao*[1,2] *Wen Ming-jie*[1,2] *Xiao Wei-qing*[1,2]
Zhang Chen-bin[1,3] *Chen Zong-hai*[1,3]

(1.USTC – ETHINK Big Data Business Intelligence Joint Laboratory, Anhui, Hefei, 230031)

(2.Anhui ETHINK information technology Co., LTD, Anhui, Hefei, 230031)

(3.Department of Automation, University of Science and Technology of China, Anhui, Hefei, 230027)

(4.Hefei Cigarette Factory of China Tobacco Anhui Industrial Co., LTD, Anhui, Hefei, 230031)

Abstract: In the scene of hundred billions of big data environment, feature mining, real-time processing, ad hoc querying and off-line calculation had high computing performance and storage requirements. It couldn't meet the requirements of all the computing environments based on the relational database or Hadoop distributed computing framework. This paper mainly introduced a visual data mining platform based on iterative computing framework Spark platform. The framework made full use of the memory, so as to improve iterative speed. In addition, it supported a variety of distributed computing and storage scene and had strong scalability for solving the big data computing problem.

Key words: Spark; Memory Computing; Big Data; Data Mining

1 引言

本文尝试使用基于内存计算的开源的集群计算系统 Spark 实现可视化大数据挖掘平台，以解决可视化挖掘对平台高性能要求的问题。该平台为最终用户提供了一个自助分析的平台，涵盖自助查询、即席分析、多维查询、仪表板、智能搜索等功能。业务建模人员直接通过可视化 web 前端完成大数据分析建模，而后端数据分析、数据挖掘逻辑模块，由开发人员基于 Spark[1-2]平台开发实现。

2 大数据挖掘现状

2.1 大数据基础平台

目前市场上主流的大数据基础平台，以 Hadoop[3-4] 生态系统为核心，Hadoop 似乎就是大数据的代名词。在 Hadoop 生态系统下，实现了各种开源应用框架，不仅降低了用户使用门槛，而且可以让各种角色用户均能实现

作者简介：李文 (1986-)，男，湖北人，硕士，研究方向为大数据挖掘；程华良(1964-)，男，安徽人，硕士，研究方向为政治经济；陈宗海 (1963-)，男，安徽人，教授，博士生导师，研究方向为复杂系统的建模与仿真与控制、机器人与智能系统。

大数据计算，图1绘制了Hadoop生态系统下主流的工具。

HBase[5]是一个针对结构化数据的可伸缩、高可靠、高性能、分布式和面向列的动态模式数据库。采用了增强的稀疏排序映射表数据模型，键由行关键字、列关键字和时间戳构成。HBase 中保存的数据可以使用 MapReduce 来处理，它将数据存储和并行计算完美地结合在一起，并能实现数据随机读写。Pig[6]定义了一种数据流语言——Pig Latin，将脚本转换为 MapReduce 任务在 Hadoop 上执行。Hive 定义了一种类似 SQL 的查询语言 HQL，将 SQL 转化为 MapReduce 任务在 Hadoop 上执行，通常用于离线分析。Zookeeper[7]是 Chubby 克隆版解决分布式环境下的数据管理问题，包括统一命名、状态同步、集群管理、配置同步等。Sqoop 主要用于在传统数据库和 Hadoop 之间传输数据。Mahout[8]实现了聚类、分类、推荐引擎（协同过滤）和频繁集挖掘等广泛使用的数据挖掘方法。Flume 是 Cloudera 开源的日志收集系统，具有分布式、高可靠、高容错、易于定制和扩展的特点。

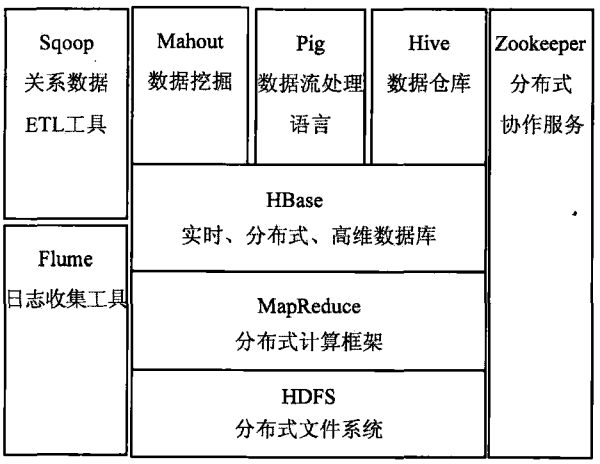

图1 Hadoop生态系统
Figure 1 Hadoop ecology

虽然 Hadoop 逐渐确立了其作为大数据生态系统基石的地位，但市场上依然有不少 Hadoop 的竞争和替代产品，如 Hydra、Spark。其中，基于 Hadoop 分布式文件系统的开源框架 Spark 成为关注的焦点，因为 Spark 弥补了 Hadoop 的短板，例如提高互动速度、更好的编程界面。实时计算和内存计算始终是大数据领域最热门的话题。

2.2 大数据分析工具

大数据分析工具是大数据市场最活跃的领域，从电子表格到时间线动画再到3D可视化，大数据创业公司们提供了各种各样的数据分析工具和界面。数据分析工具的选取，对数据分析工作效率、数据分析的建模结果、数据分析模型结果共享、模型实施等起到至关重要的作用。数据仓库建设工具收费的有 Oracle BIEE，免费的有 Pentaho。常规数据分析工具有 Excel、SPSS、R 语言。数据挖掘工具收费的有 SPSS moder，免费的有 Rapidminer 和 Weka。大数据挖掘建模工具有基于 Hadoop 的开源工具 Mahout。以上工具从各个层面解决了数据分析问题，但是在千亿级大数据环境下，无法完胜实时计算、即时交互查询、离线计算等各种复杂的场景。基于内存计算的开源框架 Spark，通过平台优化、采用内存计算方式解决了 Hadoop 的短板问题。

2.3 大数据应用

大数据应用的发展进程相对缓慢，但现阶段大数据确实已经进入应用层。金融和广告行业是大数据应用起步最早的行业，事实上在大数据概念出现之前就已经开始了。

聚类模型属于非监督式挖掘模型，以用户属性、行为、消费等特征数据为输入，将用户自动聚类为若干类，通常用来挖掘潜在目标客户群体，也可以用在大数据营销工具、CRM 工具和防欺诈解决方案上。

分类预测模型分析学习历史数据经验，预测分析未来数据发展方向。模型输出是离散数据或类别的称为分类模型，模型输出是数值类型数据的模型称为数值预测模型。分类模型根据训练数据集的类标号属性，学习现有分类数据的分类规则来构建分类器，最终被用于分类新数据。数值预测模型根据数据输入，对训练机数据进行模型拟合，最终建立连续性数值函数。分类预测模型的典型应用有欺诈检测、市场定位、性能预测、医疗诊断、价格预测等。

关联规则挖掘模型，用于发现隐藏在大型数据集中的令人感兴趣的联系。关联规则挖掘分为“频繁项集产生”和“关联规则产生”两个主要子任务。“频繁项集产生”的目的是发现满足最小支持度阈值的所有项集，这些项集称作频繁项集。“关联规则的产生”目标是从上一步发现的频繁项集中提取所有高置信度的规则，这些规则称为强规则。关联规则挖掘模型的典型应用有产品关联推荐和精准营销。

3 基于 Spark 可视化大数据挖掘

3.1 spark 介绍

Spark 是一个基于内存计算的开源的集群计算系统，目的是让数据分析更加快速。Spark 由加州伯克利大学 AMP 实验室(Algorithms, Machines, and People Lab)以 Matei 为主的团队所开发，支持分布式数据集上的迭代作业，它是对 Hadoop 的补充。Spark 启用了内存分布数据集，可以在 Hadoop 文件系统中并行运行，除了能够提供

交互式查询外，还可以优化迭代工作负载。Spark 平台通过 Mesos[9]管理集群，相关的项目有 Shark、Spark Streaming、BlinkDB、GraphX、MLib，如图 2 所示。

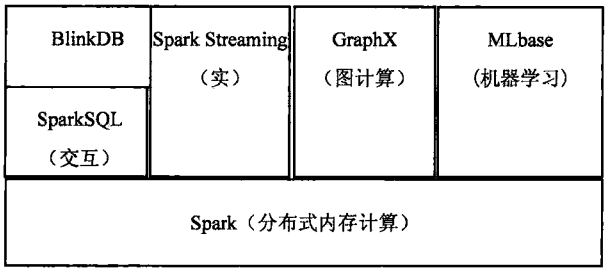

图 2 Spark 模块构成
Figure 2 Constitute of spark modules

Shark 实现交互式查询，在 Hive 的架构基础上，改写了“内存管理”、“执行计划”和“执行模块”三个模块，使 HQL 能够跑在 Spark 上。MLib 提供了一系列开箱即用的机器学习算法，涉及分类、回归分析、聚簇和推荐领域。GraphX 实现了图算法，结合了数据并行和“图并行（graph-parallel）”两种系统语义。GraphX 提供了可以与 Facebook 使用的著名的图处理系统 Giraph 相媲美甚或更好的性能。

3.2 架构图

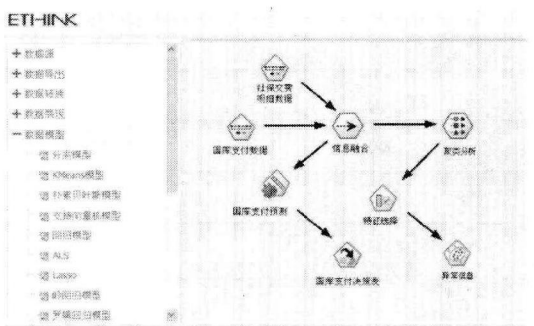

图 3 ETHINK 数据挖掘平台
Figure 3 ETHINK data mining platform

交互界面	可视化操作控件		
模型训练	图计算模型	数据挖掘算法	
任务调度	Spark引擎		
ETL引擎			
HDFS	Hive/Hbase	Oracle/Mysql	其它
业务数据、外部数据。。。			

图 4 ETHINK 数据挖掘平台架构
Figure 4 Constitute of the ETHINK data mining platform

本文所实现的可视化大数据挖掘平台 ETHINK，为最终用户提供了一个自助分析的平台，涵盖自助查询、即席分析、多维查询、仪表板、智能搜索等模块，为 IT 公司提供了一个强大的设计及开发平台。在传统 BI 平台功能之外，该平台能设计商业智能领域的 KPI、监控、绩效、地图分析、决策分析、仪表板、驾驶舱、统计挖掘等领域的功能。通过该平台，普通用户可以实现轻量级傻瓜式查询查看清单、浏览、统计查询清单、多维查询、分组及交叉查询（个性化自助、分享），IT 人员可以实现报表、仪表盘设计、开发共性需求，业务建模人员可以实现元数据、业务模型、指标定义和权限管理。

3.3 性能比较

Spark 官网(http://spark.apache.org/)公布数据，与 Apache Hadoop 相比，它在内存数据集上的性能提升了高达 100 倍，而在磁盘数据集上的性能则正常回落到 10 倍。通过将 Spark 下的各个项目与其他对应的成熟项目做比较，如图 5 所示，Spark 下的项目性能要明显优于其他项目。与实时计算框架 storm 相比，Spark 的速度是 Storm 的 2 倍，在交互查询上，基于磁盘的 Spark，查询速度是 Hive 的 5 倍，基于内存的 Spark 速度是 Hive 的 40 倍。

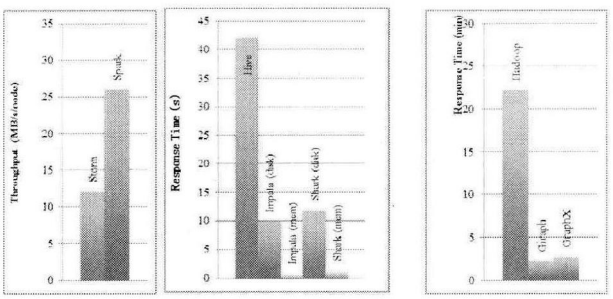

图 5 Spark 各模块性能对比
Figure 5 Performance comparison of spark module

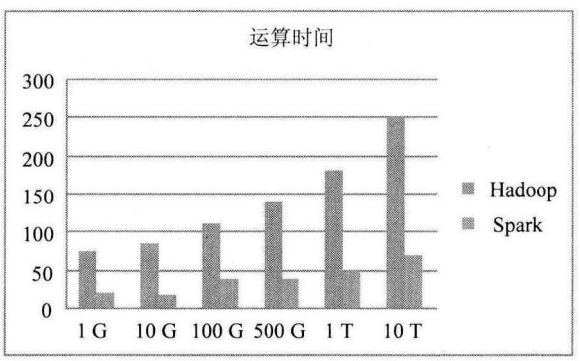

图 6 ETHINK 与 Hadoop 性能对比
Figure 6 Performance comparison of ETHINK and Hadoop

本文以基于 Spark 的可视化大数据挖掘平台 ETHINK 为基础，采用同样的算法，对 1 G、10 G、100 G、500 G、1 T、10 T 的网站行为数据完成常用数据统计计算，包括数据分组、求和、查询、比较等运算，结果显示如图 6，在数据量增大的同时，Spark 的的性能优势越来越明显，数据量达到 10T 的时候，Spark 速度是 Hadoop 速度的 6 倍，符合 Spark 官方公布的 10 倍性能数据。

4 问题及展望

平台稳定性：Spark 的最新版本是 1.0，虽然版本更新速度快，开源社区开发人员实力强，但是鉴于商业公司使用时间还比较短，平台稳定性还需要考量，因此在商业化过程中需要谨慎，且要更加注重测试。

内存迭代计算的引入，极大地提高了计算效率和速度，但是由于内存使用量较大，资源占用需求较高，对高并发的计算请求资源高效合理调度存在着较大问题。

已经实现的数据挖掘算法比较少，目前公布的数据挖掘项目 MLlib，实现的数据挖掘算法还比较有限，大量的数据挖掘算法还没有并行化，如文本数据挖掘，因此在具体应用中，需要针对具体问题，实现未在 Spark 上并行化的数据挖掘算法。

参 考 文 献

[1] Viktor meyer schon berg, Zhou Tao. Big data era [J]. Real estate Tribune. 2013,(04):21.
[2] Zaharia M, Chowdhury M, Franklin M J, et al. Spark: Cluster Computing with Working Sets[J]. *HotCloud 2010*, June 2010.
[3] White. Hadoop: The Definitive Guide ,2012.
[4] Ghemawat S, Gobioff H, Leung v. The Google file system. Proceedings of the nineteenth ACM symposium on Operating systems principles. 2003.
[5] George L. HBase:The Definitive Guide. 2011.
[6] Christopher Olston,Benjamin Reed,Utkarsh Srivastava.Pig latin: a not-so-foreign language for data processing. Proceedings of the 2008 ACM SIGMOD international conference on Management of data. 2008.
[7] Xue-hou Tan.A2-approximation algorithm for the zookeeper's problem. Information Processing Letters. 2006
[8] Owen S, Anil R, Dunning T, et al. Mahout in action.2011.
[9] Benjamin Hinman,Andrew Konwinski,Matei Zahariaetc. Mesos:A Platform for Fine-Grained Resource Sharing in the Data Center.http://www.eecs.berkeley.edu/Pubs/TechRpts/2010/EECS-2010f-87.html.2011.