第47卷第4期 中国科学技术大学学报 Vol.47,No.4
2017年4月 JOURNAL OF UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA Apr. 2017

文章编号:0253-2778(2017)4-0087-08

BDAP——一个基于Spark的数据挖掘工具平台

卜 尧[1,2],吴 斌[1,2],陈玉峰[3],白德盟[3]

(1. 北京邮电大学智能通信软件与多媒体北京重点实验室,北京 100876;
2. 北京邮电大学计算机学院 北京 100876; 3. 国网山东省电力公司电力科学研究院 济南 250000)

摘要:大数据处理系统是大数据领域的一个热点,为此首先研究大数据分析平台的架构与功能,将大数据分析平台分为数据源、数据吸收层、数据存储层、平台层、安全与监控层、设备层和应用层.平台包含多个数据预处理和算法模块,平台架构为大数据分析了奠定基础.在功能上,该平台功能全面,可以自由组合各种操作,模块之间耦合度低,便于维护和拓展.在用户体验上,调参、建立流程、监控、数据挖掘过程都是可视的,融合工作流和调度流技术.在性能上,该平台相应算法的性能优于Hive和MLlib.最后,举例说明大数据挖掘平台的应用场景.可以对电网线路故障和气象数据进行预处理,从而对故障进行预测和分类,可以通过视频挖掘组件,对数据分类.

关键词:大数据分析平台;Hadoop;Storm;Spark;批处理;数据挖掘

中图分类号:TP 391 **文献标识码**:A doi:10.3969/j.issn.0253-2778.2017.04.010

引用格式:卜尧,吴斌,陈玉峰,等. BDAP——一个基于Spark的数据挖掘工具平台[J]. 中国科学技术大学学报,2017,47(4):323-330.

BU Yao, WU Bin, CHEN Yufeng, et al. BDAP: A data mining platform based on Spark[J]. Journal of University of Science and Technology of China, 2017,47(4):323-330.

BDAP: A data mining platform based on Spark

BU Yao[1,2], WU Bin[1,2], CHEN Yufeng[3], BAI Demeng[3]

(1. *Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing* 100876, *China*;
2. *School of Computer Science, Beijing University of Posts and Telecommunications, Beijing* 100876, *China*;
3. *State Grid Shandong Electric Power Research Institute, Jinan* 250000, *China*)

Abstract: Large data processing system has become a hot spot research issue in the field of large data. First of all, we study the data analysis platform architecture and the function. The data analysis platform contains data source layer, data absorption layer, a data storage layer, data platform layer, security and monitoring layer, equipment layer and application layer. Platform multiple data preprocessing and algorithm modules. Platform architecture provides a foundation for the large data analysis. From a functional perspective, the platform features a comprehensive, combination of the various steps of the operation can be freely operated. The coupling degree between the modules is low, which is convenient for maintenance and development. From the user's point of view, the adjustment of parameters, the establishment of the process, monitoring, and data mining process are all visual, workflow and scheduling

收稿日期:2016-08-28;**修回日期**:2017-12-08

基金项目:国家高技术研究发展计划(863)(2015AA050204),国网科技项目(520626150032),北京市教育委员会共建项目建设计划资助.

作者简介:卜尧,女,1993年生,硕士生,研究方向:机器学习. E-mail: buyao1993@bupt.edu.cn

通讯作者:吴斌,博士/教授. E-mail: wubin@bupt.edu.cn

stream technology are compromised. In performance, the performance of the BDAP algorithm is better than that of Hive and MLlib. Finally, an example illustrates the application scenarios of this data mining platform. After we analyze the circuit fault and meteorological data, we can predict and classify the fault. Also we can use video mining to get useful information classified.

Key words: big data analysis framework; Hadoop; Storm; Spark; batch processing; data mining

0 引言

随着移动互联网的兴起,人们可以利用网络足不出户做很多事情:转账、购物、打车、代驾、交水电费、充值等,产生的数据量大且繁杂. IT 巨头纷纷开始抢占大数据领域,如 Google、Facebook、Oracle、IBM、Microsoft、EMC、Intel 等公司都陆续推出大数据的解决方案和相关产品.各种大数据技术也是迅猛发展,如 Hadoop 与 Storm 的配合、tachyon 的引入、Spark、HBase、MongoDB 等,这些技术推动着大数据快速向前发展.

网络数据量的急速增加,数据形式也越来越复杂,需要快速地从这些海量数据中高效率获得有价值并且直观的结果.本文构建了一个大数据分析平台(big data analysis platform,BDAP),在用户不用理解掌握编写算法的条件下,可以简单自如地使用该平台处理大量数据,进行预测和分析.BDAP 所用到的相关技术有:批处理技术、工作流引擎、实时处理技术、NoSQL 存储和前端技术.

该平台的特点和优点有:①可以将每一步数据处理操作连接成为一条工作流,保存后可以定时或随时调度,工作流和调度流的引入,使系统简单易用.②功能全面,涵盖多种数据挖掘预处理和算法模块,可以处理多种数据,如结构化数据,半结构化数据、非结构化数据.③模块之间相互独立,低耦合性,易维护易扩展.经性能测试得出:ETL 在性能上优于 Hive,算法在性能上优于 MLlib.调参、流程、监控、结果可视化,使之易于调控和修改.

BDAP 可用于企业故障及设备缺陷的分类和预测、用户及用户消费套餐的分类、制定、预测.从开源工具来看,早期有基于 MapReduce 的 Mahout 项目,现在有基于 Spark 的 MLlib,但它们更侧重算法的并行实现,一方面只是单一算法包,集成性易用性并不好;另一方面算法性能也有待提高.

现有的大数据分析平台各有侧重,例如,Datix[1]实现一个可扩展网络分析平台,该平台提出智能预分区存储方案,加快了过滤查询等预处理的时间.Kloud[2]平台实现自由表辅助索引设计,提高了自由表的访问效率.Big-Provision[3]提供了一个可以比较同一算法在不同框架下的准确性与性能的平台.相对较完善的企业项目有 PDMiner[4]和 BC-PDM[5].这些平台集成一些基本的数据挖掘预处理及算法,并不是很全面,并且模块与模块之间关联性强,不利于增加和修改.BDAP 功能完善,不仅涵盖多个数据预处理和数据挖掘算法,还包含对特定企业定制的个性化算法组件,在平台框架上也做了重大改进,将 OSGI 与 Tomcat 相结合,构建了低耦合易扩展的工作流引擎.此外还引入基于 Storm 的实时分析,运用 Storm 和 Kafka 开展实时分析,设计并实现了基于 Storm 的实时监控系统.

1 技术基础

1.1 Hadoop

Hadoop[6]框架最核心的设计是:HDFS(Hadoop distributed file system)和 MapReduce. HDFS 实现一个主从结构的分布式系统基础架构,有高容错性、高吞吐量的特点,适合超大数据集的应用程序[7].MapReduce[8]主要用于对存储在分布式文件系统上的海量数据进行分布式计算.此外,Hadoop 还拥有包括 HBase[9]、Pig、Hive、Sqoop 等多样化的工具,具有分布式数据库、分布式数据仓库、数据传输等功能.

1.2 YARN

Apache Hadoop YARN[10]是 Hadoop 重构的新的 MapReduce 框架.它是一个通用资源管理系统,可为上层应用提供统一的资源管理和调度,它的引入为集群在利用率、资源统一管理和数据共享等方面带来巨大好处.

1.3 Tachyon

Tachyon[11]是一种高性能、高容错、基于内存的开源分布式存储系统,兼容 Hadoop MapReduce 和 Apache Spark 等特征.Tachyon 能提供内存级速度的跨集群文件共享服务.

1.4　Spark

Apache Spark[12] 是开源计算框架，可作为 MapReduce 的一种替代方案，兼容分布式文件系统 HDFS，可以在一定程度上弥补 MapReduce 的不足. Spark 同时还含有大量开箱即用的机器学习库.

1.5　工作流引擎

工作流引擎[13]可以根据角色、分工和条件的不同决定信息传递路线、内容等级等核心解决方案.工作流引擎包括：流程的节点管理、流向管理、流程样例管理等重要功能.成熟的工作流引擎有：JBPM、Activiti、Shark 等.

1.6　实时处理

实时处理主要是为弥补批处理延时的缺点，能够实时、快速处理不断产生的流数据.目前较为成熟的实时处理框架是 Storm[14]，Storm 一般和 Kafka[15]等消息系统配合使用.

2　平台架构及功能

大数据平台可划分为 7 个层次，如图 1 所示.

图 1　七层结构图

Fig.1　Seven layers of structure

2.1　数据源

数据源是大数据分析的输入，泛指所有现实世界中的原始数据.一般来说，可以将数据分为结构化数据、半结构化数据和非结构化数据；从研究领域来说，可以分为商业数据、网络数据和科研数据等；从数据来源来说，可以将数据分为三类：传感器数据、日志数据和爬虫数据.无论从哪个角度出发，数据源都是十分复杂和重要的，处理好数据源将为后续大数据的分析提供良好的基础.

2.2　数据吸收层

从数据源获取的数据是原始数据，这些数据种类多数量大，需要大量的存储空间，但并不是都是有用的数据，所以要进行数据吸收.数据吸收层的主要任务包括数据收集和数据预处理.

首先是数据采集.Flume 主要用于从现有的系统中收集数据，可以是远程或本地集群的数据.对于传统应用，数据一般存储于传统的关系型数据库(RDBMS)之中，例如，MySQL、Oracle 等，为了进行大数据分析，需要将这些数据导入大数据系统中. Sqoop 工具可以将数据从关系型数据库中导入到 HDFS 上，也可以将 HDFS 的数据导出到关系型数据库中.

数据预处理主要有：去重、去极值、压缩、抽样等.去重可以将重复的数据除去；通过去极值可以将数据中的异常值除去；通过压缩可以减少数据的传输量，减少网络传输；通过抽样可以按照所需比例抽取一定数据.经过预处理的数据直接存储于 HDFS 或 NoSQL 数据库中，用于后续处理和分析，也可以进行实时分析.

数据预处理在平台中由 ETL 模块实现，其中 ETL 包含：清洗类 ETL，转换类 ETL，集成类 ETL，计算类 ETL，集合类 ETL，抽样类 ETL，更新

类ETL.

2.3 数据存储层

数据存储层核心是HDFS,数据源的数据经过数据吸收后存储于HDFS之上,对于非结构化的数据可以直接存储;对于结构化的数据,如从数据库中导出的记录,则是以文本形式存储于HDFS之上.有以下三种数据访问方式:

(Ⅰ)可以通过REST接口直接访问HDFS上的数据和元数据,结果以JSON的形式返回客户端.

(Ⅱ)对于已经得到分析结果的、可以直接展现给用户的数据,可以存储于RDBMS或NoSQL数据库中,通过SQL语句方便地进行查询,也可以基于这些数据的结果进行应用的开发.

(Ⅲ)对于经常访问的数据可以将其缓存起来,再次访问时直接从内存中获取,提升访问的速度.

以上三种数据访问方式是系统中主要的数据访问方式,HDFS是系统的存储基础.

平台也提供数据交换功能,数据交换是指将传统的数据仓库中的数据直接导入分布式文件系统(HDFS)以及将处理结果从HDFS导出到数据仓库的过程.其主要目的是方便用户在数据仓库与分布式文件系统之间传输数据,使得数据仓库中的数据可以被BDAP中的算法所使用,并可以将处理结果导回数据仓库中存储.其处理的主要对象就是数据仓库中各类数据.

2.4 大数据平台层

平台层是大数据分析系统的核心,该层可兼容运行多个计算框架.

MapReduce和Spark都运行于YARN之上,Storm可以单独运行,也可以运行于YARN之上,YARN是一种新的Hadoop资源管理器,它是一个通用资源管理系统,可为上层应用提供统一的资源管理和调度.Tachyon[16]是分布式文件系统,位于各种计算框架和分布式文件存储HDFS之间,目标是将那些不需要落地到DFS的文件,落地到分布式内存文件系统中,达到共享内存,从而提高效率,同时可以减少冗余存储,GC时间等.

MapReduce可以完成系统旧的算法功能,同时基于Hadoop的数据仓库工具Hive也是基于MapReduce的,可以将结构化数据文件映射成数据库的表,并提供SQL查询功能.Spark迭代计算相对于Hadoop有很大的优势,同时Spark的数据存储基于内存,相对于Hadoop基于磁盘的存储也有很大的提升,在一些算法性能方面,Spark的性能比Hadoop好,BDAP中聚类与分类的算法都是基于Spark实现的.Storm主要用于实时分析,通过实现相应的Spout(数据源)和Bolt(数据操作),并将其组合成Topology(拓扑图),可完成实时应用的构建.通过以上基本组件和框架,实现大数据平台层的基础构建.

2.5 监控层

对已启动的挖掘任务的进度进行监控,将计算服务器返回给Web服务器的监控结果以进程条的形式显示在用户界面上,同时允许用户对系统的操作步骤进行监控,结果以文本形式存放于日志文件中并显示在Web页面上,任务监控包括内容如表1所示.

表1 任务监控内容

Tab.1 Monitoring task content

监控项目名称	监控任务内容	展现方式
数据预处理	已完成预处理数据的百分比	显示进度条
数据计算	数据计算运行时间	显示进度条

以往的监控实现主要采用轮询方式,即前台以固定的时间间隔不断向后台发送请求来获取工作流中各个任务的数据处理进度等相关信息,进而显示在前台监控界面中,这种实现方法效率比较低,浪费资源,而且更新不及时.改进后的Tomcat Web容器可实时监听流程执行情况,一旦有状态更新就向Redis pub/sub更新情况,如图2所示.

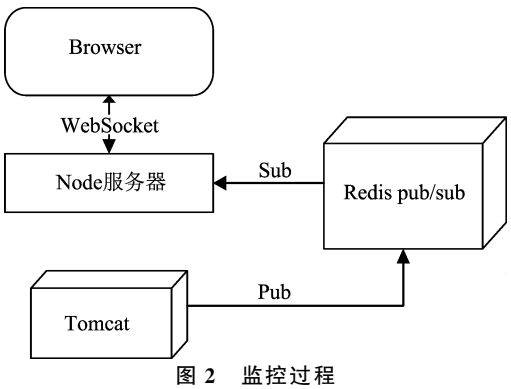

图2 监控过程

Fig.2 Monitoring process

2.6 设备层

设备层是决定平台节点的组织,资源分配的基础.由于大数据系统节点较多,集群节点组织一般采用的是树形结构,多个节点通过交换机连接起来组成小型局域网络,多个小型局域网再连接起来组成大型局域网络.

2.7 应用层

应用层主要包含基础平台和基于应用平台的拓展.基础平台是构建于 Tomcat 之上的 Web 服务,开放工作流接口、调度流接口、HDFS 操作接口等,基于此可以拓展各种功能.基础平台可以作为系统开放平台,通过调用开放平台的接口来实现子系统.在用户界面呈现出来的功能有:社会网络分析、视频挖掘、微博分析、推荐系统、Web 报表、文本分类、调度流.

2.7.1 社会网络分析

社会网络分析模块可以从社会网络角度出发,对输入的文件进行节点与属性的解析.节点与节点之间的联系由所选择的属性决定,构成节点之间的边,形成网络.该功能通过分析这些社会关系,把从偶然相识的泛泛之交到紧密结合的家庭关系的个人或组织串连起来,社会关系包括朋友关系、同学关系、生意伙伴关系、种族信仰关系等.

2.7.2 视频挖掘

平台集成基于 Spark 的大规模视频数据挖掘框架,基于开源软件的三层框架,研究视频数据特征模型的相关并行算法.

视频挖掘的过程包括:特征抽取→视频词汇生成→bag of feature 特征模型生成.bag of words(bag of features)模型将文档表示为一些特征词汇的组合,忽略特征词之间的顺序关系,统计文本中特征词出现的频率,然后进行相关计算和匹配.最后利用生成的特征词汇,将图像中的大量特征向量通过加权的形式表示成 bag of features 向量.

2.7.3 微博分析

微博分析模块针对企业社交网络,从整体、社团、个人三个层次对用户的微博信息进行分析.具体提供以下几个方面的分析:

(Ⅰ)用户交往关系图的生成.利用用户的粉丝关注关系和转发评论,构建用户交往关系图.

(Ⅱ)用户地点信息统计.根据微博签到信息,统计出用户经常出现的地点,从中发现一些规律.

(Ⅲ)热点话题发现.从用户近期发布微博中发现用户关心的热点事件.

(Ⅳ)关键词提取.从用户近期微博中提取出关键词,从中可以发现一些用户特征.

(Ⅴ)用户信息挖掘.统计用户的性别、地域等基本信息.对用户个人分析,同时对用户发微博的时段分布、用户近期密友等进行统计.

系统提供以下三种数据导入的方式:①由用户定义群体的公司、学校、地域等属性,系统首先获取群体用户的微博 UID,再根据 UID 利用微博 API 爬取数据,并存入数据库中.②由第三方提供微博数据,系统解析并将其插入数据库.③系统其他模块执行时,在需要的情况实时爬取数据并插入数据库.

2.7.4 推荐系统

推荐系统通过挖掘用户、物品信息,进而向用户推荐,达到提升用户体验以及实现企业自动化智能营销的目的,高质量的推荐能够充分挖掘用户的潜在喜好.推荐算法模块是推荐系统的核心.

该平台实现混合推荐系统,包含数据获取模块、技术支撑层、推荐引擎模块以及推荐系统的展示层.推荐引擎作为整个混合推荐系统中的核心,提供多种推荐实现方式,包括非个性化推荐、个性化推荐、介于个性化与非个性化之间的群组推荐等.非个性化推荐提供基于统计以及物品相关性计算的推荐结果.个性化推荐结合用户自身的历史行为,根据用户个人偏好对每个用户产生独有的推荐.群组推荐对于群组内部用户,其推荐结果是一样的,对于群组之间则是根据群组自身历史行为不同而产生不同的推荐结果.

2.7.5 Web 报表

Web 报表提供可视化的数据显示,供用户分析数据和作出决策.该模块对结构化数据提供丰富的可视化显示和操作,主要分为两大部分功能:报表分析和图表分析.HDFS 文件系统上的数据可用报表或图表直观显示出来,并且可以修改图表和报表以达到方便操作和美观的目的.

2.7.6 文本分类

文本挖掘是从大量的非结构化文本数据中抽取模式和知识的过程.该平台数据挖掘模块集成 SVM 和朴素贝叶斯文本分类算法.

3 平台特色

3.1 功能全面

BDAP 大数据分析平台可处理的数据多样,可以处理结构化、半结构化和非结构化文本数据.相比于已有的大数据分析平台,其实现的数据预处理、分类算法、聚类算法、关联规则,可以建立复杂的数据挖掘任务,如图 3 所示.自动分类器和自动聚类器可以分别在多个分类、聚类算法之间选择最优算法,如

图 4 所示.BDAP 与机器学习库 MLlib 相比,集成的算法更全面.BDAP 基本涵盖 MLlib 的功能,如表 2 所示.

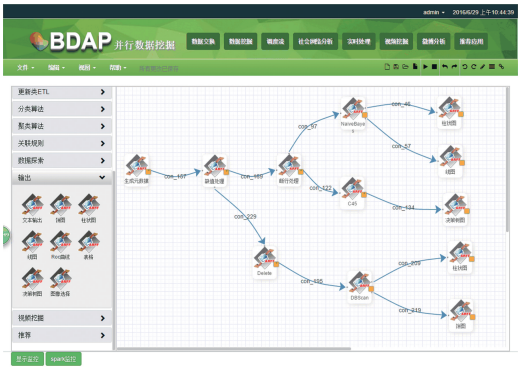

图 3 多挖掘任务的工作流

Fig.3 Workflow for multiple mining tasks

图 4 自动分类器设置面板

Fig.4 Automatic classifier settings panel

表 2 MLib 和 BDAP 的算法对比

Tab.2 Comparison between MLib and BDAP in algorithm

	MLlib	BDAP
分类算法	文本贝叶斯、SVM、线性回归、决策树、逻辑回归、神经网络	文本贝叶斯、朴素贝叶斯(适用于一般分类)、SVM、线性回归、逻辑回归、C45、CART、CHAIID、神经网络、自动分类器
聚类算法	*K*-means、LDA	*K*-means、LDA、DBScan、Brich、Clara、自动聚类器
关联规则	FpGrowth	Apriori、FpGrowth、时序关联规则、基于信息熵的关联规则

ETL 数据预处理算法:数据预处理算法在数据挖掘中起着非常重要的作用,其输出通常是数据挖掘算法的输入.由于数据量的剧增,串行数据预处理过程需要消耗大量的时间.

并行数据挖掘子系统包括:并行关联规则算法、并行分类算法以及并行聚类算法.目前 BDAP 中已经开发很多经典的数据挖掘算法.

BDAP 可以建立复杂挖掘任务,对同一数据进行多种处理,用多种算法执行,以不同的方式展示结果.自动分类器和自动聚类器可以分别运行多个分类和聚类算法,并表现性能最好的结果.

3.2 模块耦及工作流调度

功能高度模块化,从导入、各项预处理操作、各算法及导出各种形式结果都集成为独立模块,前者的输出是后者的输入,极大地提高了平台的扩展性、降低了耦合度.同时可以对算法和组件进行个性化定制,便于维护和功能扩展.

工作流和调度流是 BDAP 大数据分析平台的特点之一.工作流由用户手动搭建,由一个个处理单元和有指向性的箭头连接而成.如图 5 所示,该工作流的操作为:导入元数据→数据类型检查→字段类型转换→where→SVM→Roc.

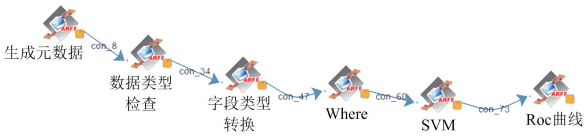

图 5 工作流

Fig.5 Workflow

工作流引擎是系统批处理的核心模块,主要通过 OSGI 容器来实现.将 OSGI 与 Tomcat 进行融合,通过 Tomcat 来调用服务,通过 OSGI 来实现服务管理. 基于 OSGI 的服务是一个可扩展、可伸缩、灵活的插件体系结构,各算法模块独立,具有良好的拓展性和较低的耦合度.较商业级的 JBPM、Shark 等工作流引擎更优,同时可以实现对大数据分析算法和组件的定制,维护也更简单.

流程调度主要是支持实现按照时间周期或者文件是否存在或者依赖关系定义多操作流调度任务,更改、删除已有调度任务等管理方式.调度流管理可以按设定时间或流程依赖关系(流程执行的先后顺序)自动运行数据流程或完整业务应用.调度流支持对调度流的新建,第三方应用的添加,制定调度计划,修改调度计划以及运行监控.

3.3 算法并行性能高

对 ETL 算法的组件与 Hive 以及 MLlib 分别作不同数据规模下的性能测试,测试结果均显示 BDAP 的性能更好.

Hive 是建立在 Hadoop 上的数据仓库基础构架，可以将结构化的数据文件映射为一张数据库表，并提供简单的 SQL 查询功能，称为 HQL，它允许熟悉 SQL 的用户查询数据.可以将 SQL 语句转换为 MapReduce 任务进行运行.其优点是学习成本低，可以通过类 SQL 语句快速实现简单的 MapReduce 统计，不必开发专门的 MapReduce 应用，十分适合数据仓库的统计分析.对于相对复杂的功能，Hive 需要多条语句才能实现，而 ETL 组件可以通过启动一个 MapReduce 程序来实现其功能，在灵活性和简洁度上都有很大的优势，因此复杂算法运行时间远小于 Hive 实现同样功能耗费的时间.

MLlib 是 Spark 对常用的机器学习算法的实现库，同时包括相关的测试和数据生成器.MLlib 目前支持四种常见的机器学习问题：二元分类，回归、聚类以及协同过滤，同时也包括一个底层的梯度下降优化基础算法.BDAP 中的算法，贝叶斯、逻辑回归等算法对 MLlib 的流程进行优化，提升运行速度；SVM 在算法上进行改进，提高算法的并行度，提升算法的运行效率.

3.4　可视化

本平台的可视化操作可提高用户体验、调参、流程、监控及结果构成一条完整的数据挖掘流程，用户在每一步都可以参与和修改.

（Ⅰ）调参可视化.点击某个组件时，会有一个用户交互界面读写该算法参数设置，这些参数包括训练数据、测试数据、输出结果以及模型文件在 HDFS 上的存储路径，包括 Reduce 任务个数的设置.

（Ⅱ）流程可视化.通过图像化 UI 为用户提供服务，支持用户通过点击图标、拖拽连接线、自行编辑参数配置，可以根据数据处理需求拖拽多个组件构成一条工作流，也可以在一个面板下建立多个工作流任务，工作流内部和工作流之间，都是并行的.

（Ⅲ）结果可视化.输出组件实现对结果的可视化，可以通过拖拽连接算法组件后，输出结果.目前支持的输出结果有：文本输出、饼图、柱状图、线图、Roc 曲线、表格、决策树图、图像选择.如图 6 所示.

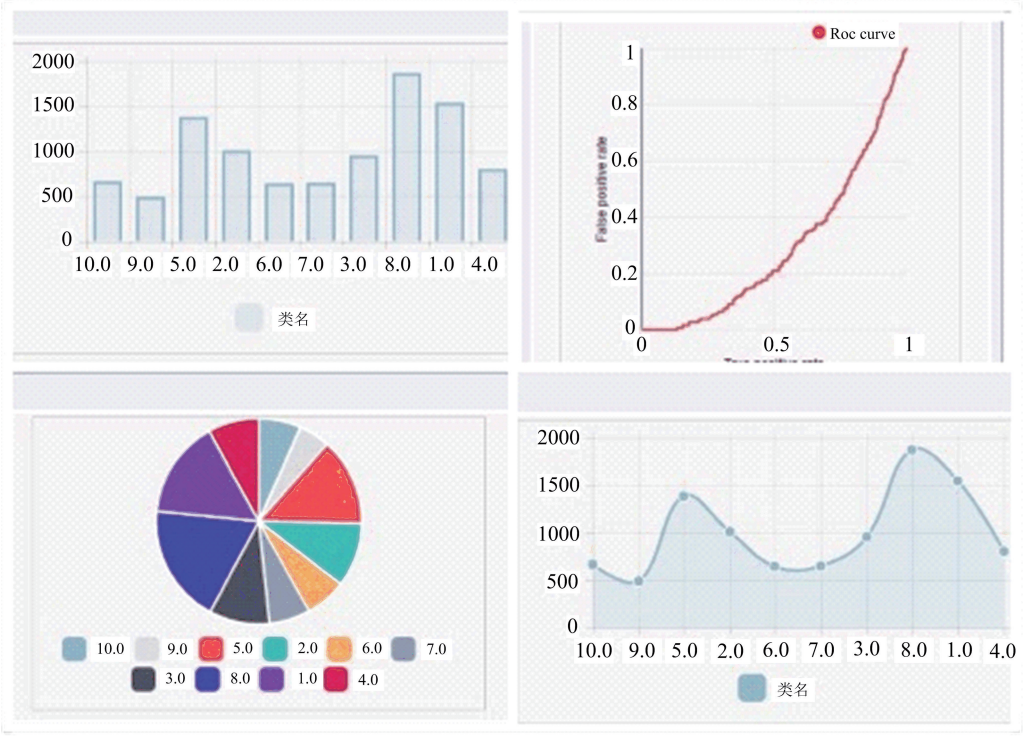

图 6　结果展示

Fig.6　Results show

（Ⅳ）监控可视化.当一条工作流开始运行后，会出现监控窗口，显示运行 job 的名称、开始时间、结束时间、持续时间、job 状态.左侧显示 Map 和 Reduce 的处理进度和时间，这样可以直观体现算法的性能，如图 7 所示.

4　性能分析

实验环境介绍：Spark 集群包括 4 台 slave 节

图 7 监控数据

Fig.7 Monitoring data

点,1 台 Master 节点.服务器是 Dell R720,CPU 是 Intel(R) Xeon(R) CPU E5-2620 v2,每台机器配有 2 个处理器,以太网卡是 Broadcom Corporation NetXtreme BCM5720 Gigabit Ethernet PCIe * 4,硬盘读写速度 199.00 MB/s,64G 运行内存.操作系统为 64 位 centos6.5,JDK 版本为 jdk1.7.并行计算框架为 Hadoop 2.6.0,Spark 1.5.1.

4.1 ETL 与 Hive

为了测试大规模数据,本文用程序随机生成 25 维,数据规模分别为 100M、1G、2G、5G 这 4 组数据.以 OperGenColumn 组件和属性交换组件为例,OperGenColumn 需要的 SQL 语句如下:

altertable t1 add columns (c string);

update t1 set column26 = column1 + column2;

属性交换需要的 SQL 语句如下:

insert into newt1 select column2, column1, column3,column4,column5, column 6, column 7, column 8, column 9, column 10, column 11, column 12, column 13, column 14, column 15, column 16, column 17, column 18, column 19, column 20, column 21, column 22, column 23, column 24,column25 from t1.

由此可以看出,即使是功能简单的组件也需要多条 SQL 语句才能完成,Hive 需要将结构化数据映射成一张表,由 SQL 语句完成各步操作,越复杂的功能需要的 SQL 语句越多.SQL 语句转换成 MapReduce 需要时间,因此 ETL 比 Hive 性能优越,如图 8 所示.

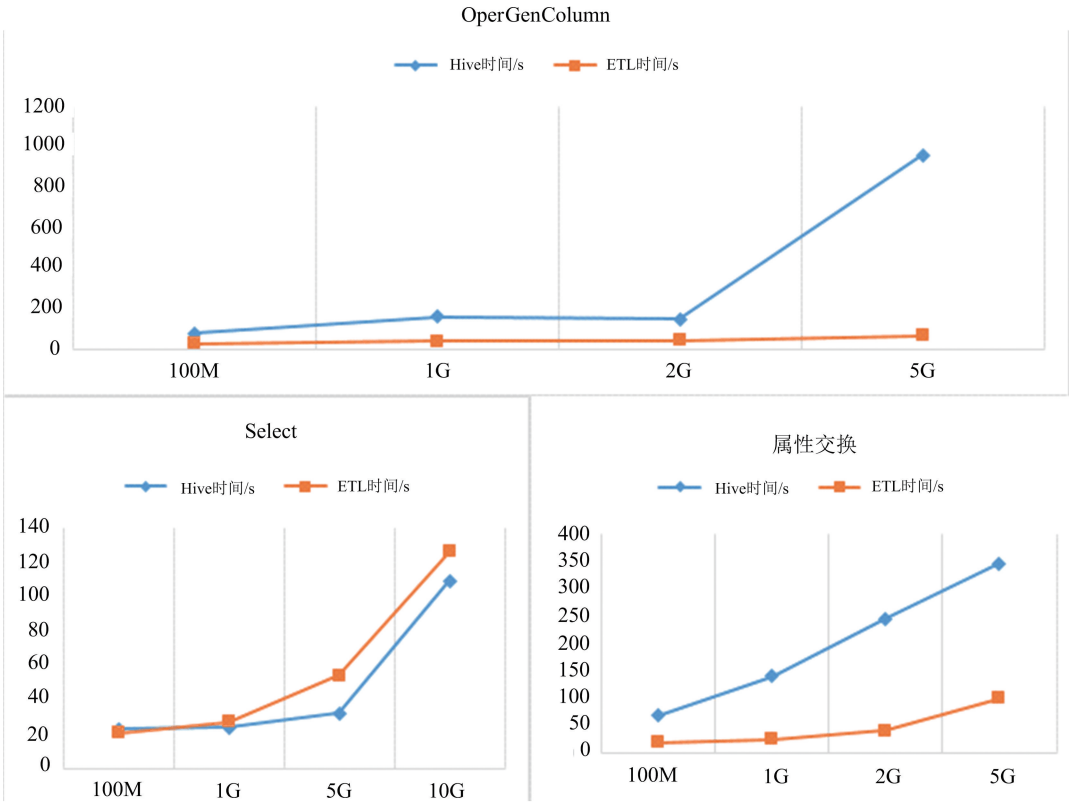

图 8 ETL 和 Hive 性能对比结果

Fig.8 Comparison of performance between ETL and Hive

4.2 算法与 MLlib

MLlib 是 Spark 对常用的机器学习算法的实现库,以 BDAP 中分类算法中的 BP 算法为例,该算法从动态调整学习率、增加动量因子、采用小批量梯度

下降法、换用交叉熵代价函数、使用 early stop 防止过拟合等多个方面对传统 BP 算法加以改进，并将改进后的算法基于 Spark 平台进行实现.现将 BP 算法与 MLlib 中的多层感知分类器(MLPC)做对比分析.

为了测试大规模数据，本文基于 covtype 通过数据复制的方式生成规模为：50 万、100 万、200 万、500 万和 1000 万这 5 组数据.结果如图 9 所示.

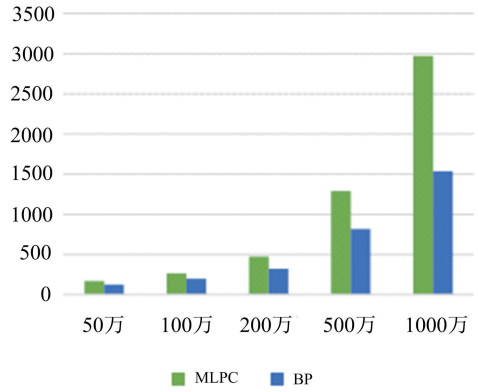

图 9　BP 与 MLPC 运行时间对比

Fig.9　Comparison of running time between BP and MLPC

此外，还对比逻辑回归、线性回归和文本朴素贝叶斯与 MLlib 的对比.实验结果显示，逻辑回归、线性回归、文本朴素贝叶斯的性能均好于 MLlib，原因在于 BDAP 中的算法，有些对 MLlib 的流程进行优化，有些在算法上进行改进，提高算法的并行度，提升算法的运行速度，如表 3～5 所示.

表 3　逻辑回归

Tab.3　Comparison of running time between logistic regression and MLlib

Logistic Regression	200M	2G	5G	10G
行数	10 000	1 000 000	3 000 000	5 100 000
BDAP 时间(s)	2 002	6 579	22 191	40 653
MLlib 时间(s)	2 109	7 783	22 032	39 271

表 4　线性回归

Tab.4　Comparison of running time between linear regression and MLlib

Linear Regression	200M	2G	5G	10G
行数	10 000	1 000 000	3 000 000	5 100 000
BDAP 时间(s)	1 977	7 010	24 009	42 591
MLlib 时间(s)	2 098	7 204	23 264	39 657

表 5　文本朴素贝叶斯

Tab.5　Comparison of running time between Naive Bayes and MLlib

文本朴素贝叶斯	10M	100M	1G	5G
行数	290 000	606 887	14 450 000	68 250 000
BDAP 训练时间(s)	57	42	97	209
MLlib 训练时间(s)	30	59	99	310
BDAP 测试时间(s)	54	81	160	652
MLlib 测试时间(s)	51	75	256	1 053

4.3　SparkBIRCH 算法的加速比实验

一般情况下，大数据集的加速比与 CPU 核数是成正比关系的，随着 CPU 核数增加，加速比也随之增加.并且在核数一定的情况下，随着数据点增多，其加速比也随之增大.数据量越大，加速比的值越大，并且随核数增加，增大趋势越明显，如图 10 所示.

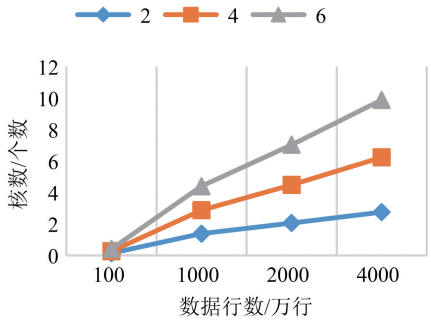

图 10　Spark BIRCH 在不同核数和不同数据行数下的加速比

Fig.10　The speedup of Spark BIRCH under different number of cores and different number of data rows

5　应用实例

5.1　电网数据预处理

5.1.1　数据的获取和初步处理

(Ⅰ)获取输入数据.调接口得到的线路故障前线路两端的 ems 数据和气象数据，包括可能造成故障的 11 个因素：大风风向、风速、能见度、阵风风速、相对湿度、降水量、阵风风向、风向、气压、极大风风速、温度；包括三个线路状态量：P—有功功率、Q—无功功率、I—电流.

(Ⅱ)处理结果.将同线路同时刻的线路各项数据指标整理成同一条数据.

(Ⅲ)处理方式.将通过接口获得的 ems 根据线路的 id 和时刻进行整合,把 P、Q、I 放到同一条数据中,并保证线路的同一端数据在同一列;将通过接口获得的气象数据进行计算,把测量的时间维度和 ems 数据保持一致(ems 数据隔五分钟测一次,气象数据间隔 10 分钟),然后再根据线路 ID 和时刻将气象数据和上述 ems 数据处理成一条;对得到的数据进行缺值处理(把离缺值最近的非缺值赋给缺值);最后生成数据并写入文件.

5.12 正常数据的范围计算

(Ⅰ)输入数据.5.1.1 节处理步骤(Ⅰ)中得到的线路故障之前一短时间(可变)的各项指标信息.

(Ⅱ)处理结果.线路不同时刻的各项指标的变化范围.

(Ⅲ)处理方式.将输入文件进行整合,对相同线路相同时刻不同日期的数据逐项进行计算,算出各项数据在不同时刻的平均值和方差,并得到正常状态下不同线路不同时刻各项指标的粗略变化范围(其中 ems 是得到某一端的变化范围和两端差值的变化范围,气象数据得到两端的变化范围).

5.1.3 故障前的不正常项关联规则的输入

(Ⅰ)输入数据.上述处理步骤(Ⅰ)中得到的线路故障之前半个小时的各项指标信息和上述处理步骤(Ⅱ)中得到的各项指标的变化范围信息.

(Ⅱ)处理结果.线路故障前半个小时各时刻的各项指标情况,把不正常的指标整合成关联规则的输入数据.

(Ⅲ)处理方式.首先要广播标准的数据表(这个标准表每一项的范围以及它的格式是前一步生成的),根据标准的数据表,判断每一条输入数据的特征是否符合合标准,如果不符合标准,就输出这个特征,这样每一条输入就对应一行异常特征的输出,最后就生成关联规则的输入.然后根据需要还可以进行各特征量之间的回归分析,以此来查看各特征之间的正负相关性,判断是否与物理规律类似,为之后的预测奠定基础.再运用关联规则的算法:Apriori 算法、FpGrowth 进行分析,可以得到相应的分析结果,如图 11 所示.

5.2 视频挖掘实例

视频挖掘功能包括:分类,检索,识别,跟踪,检测.KTH[17]数据集包含 6 种人类动作:行走、慢跑、奔跑、拳击、挥手和拍手.25 个实验者分别在 4 种场景下录制 6 类行为视频,如图 12 所示.s_1 是户外小范围运动场景,s_2 是有一定尺度变换的户外大范围场景,s_3 是穿着不同衣服的户外场景,s_4 是室内场景.

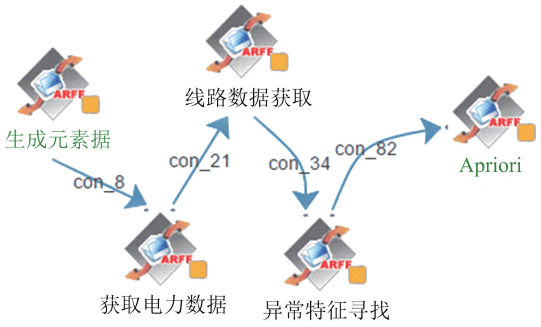

图 11 电网数据预处理过程

Fig.11 Grid data preprocessing process

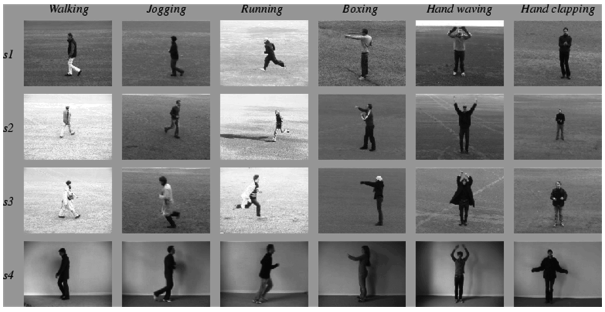

图 12 KTH 数据集视频截图

Fig.12 KTH data set video screenshot

实验中,训练集选取 16 个实验者,测试集取 9 个,即训练集包含 16 * 6 * 4 个视频,测试集包含 9 * 6 * 4个视频,本实验中使用到的算法组件均是 BDAP 提供的.

(Ⅰ)训练数据集生成特征词汇

图 13 和 14 展示流程的建立和面板设置信息.

图 13 训练集生成视觉词汇

Fig.13 Training set to generate visual vocabulary

(Ⅱ)训练 SVM 分类模型

利用提取出的原始视觉特征,转换成 bag of features 向量,然后使用 SVM 进行训练.这里的 SVM 组件是本平台已经集成的并行 SVM 算法.

图 14　特征抽取和视觉词汇生成配置面板

Fig.14　Feature extraction and visual vocabulary generation configuration panel

图 15　训练分类模型

Fig.15　Training classification model

(Ⅲ)测试集预测

图 16 是使用测试集进行预测的过程,测试视频数据也要进行视频特征抽取和 bag of features 转换,然后使用训练阶段已经训练出来的 SVM 分类模型进行分类预测.

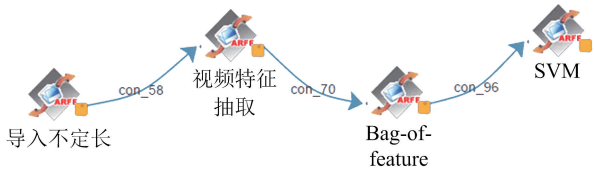

图 16　测试集预测

Fig.16　Test set prediction

6　结论

本文构建了一种基于 Hadoop 和 Storm 的大数据分析平台——BDAP.本文的主要成果有:

(Ⅰ)研究大数据分析平台的基础架构,并提出大数据平台的 7 层架构.

(Ⅱ)分析了 BDAP 的特点和优点.

(Ⅲ)对平台与 Hive 和 MLlib 进行性能对比,得到结论:ETL 越复杂,其性能优于 Hive 就越明显;经过改进,BDAP 算法也比 MLlib 性能好.

(Ⅳ)分析两个应用实例.一是根据传感器数据获得电路故障与电路状态参量与气象的关联,从而对故障聚类、根据环境量预测故障.二是视频挖掘实例,用已有的视频数据集测试该组件功能.

参考文献(References)

[1] Dimitrios Sarlis, Nikolaos Papailiou, et al. Datix: A system for scalable network analytics [J]. ACM SIGCOMM Computer Communication Review, 2015, 45(5): 21-28.

[2] 卓安. 基于 P2P 可伸缩架构的大数据分析平台研究与实现[D]. 北京: 清华大学, 2012.

[3] LI H, LU K J, MENG S C. Bigprovision: A provisioning framework for big data analytics[J]. IEEE Network, 2015, 29(5): 50-56.

[4] 何清, 庄福振, 曾立, 等. PDMiner: 基于云计算的并行分布式数据挖掘工具平台[J]. 中国科学:信息科学, 2014, 44(7): 871-885.
HE Qing, ZHUANGFuzhen, ZENG Li, et al. PDMiner: A cloud computing based parallel and distributed data mining toolkit platform[J]. Scientia, Sinica information, 2014, 44(7): 871-885.

[5] YU L, ZHENG J, SHEN W C, et al. BC-PDM: Data mining, social network analysis and text mining system based on cloud computing[C]// Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China: ACM Press, 2012: 1496-1499.

[6] 董西成. Hadoop 技术内幕深入解析 MapReduce 架构设计与实现原理[M]. 北京: 机械工业出版, 2013: 92-93.

[7] Welcome to Apache™ Hadoop ©! [EB/OL]. http://hadoop.apache.org/.

[8] DEAN J, GHEMAWAT S. MapReduce: Simplified data processing on large clusters[J]. Communications of the ACM, 2008, 51(1): 107-113.

[9] HBASE. A distributed, scalable, big data store[EB/OL]. https://hbase.apache.org/, 2016.03.28.

[10] 方宸.基于 YARN 网络数据分析系统实现与应用研究[D]. 武汉:华中科技大学, 2014.

[11] Alluxio. Formerly known as Tachyon, is a memory speed virtual distributed storage system [EB/OL]. http://www.alluxio.org/. 2016.03.28.

[12] Apache Spark. Lightning-fast cluster computing [EB/OL]. http://spark.apache.org/, 2016.03.28.

[13] 沈超. 基于 Web Services 的工作流系统关键技术的研究与实现[D]. 南京:东南大学, 2005.

[14] Apache Storm - A free and open source distributed realtime computation system, http://storm.apache.org/.

[15] GARG N. Apache Kafka[M]. Packt Publishing, 2013: 5-9.

[16] 石山园. Spark 入门实战系列: 10.分布式内存文件系统 Tachyon 介绍及安装部署[EB/OL]. http://www.cnblogs.com/shishanyuan/p/4775400.html, 2015.9.10.

[17] KTH video dataset. Recognition of human action[EB/OL]. http://www.nada.kth.se/cvap/actions/, 2016.03.28.