

基于Hadoop与Spark的高校校园大数据平台研究

刘 萍

(江阴职业技术学院计算机科学系, 江苏 江阴 214400)

摘 要: 对校园大数据分析是校园信息化发展的新思路。Hadoop是Apache基金会开发的分布式系统基础架构,它是集分布式计算、存储和管理为一体的生态系统。目前流行的Spark框架是与Hadoop生态系统中的MapReduce类似的一个分布式计算平台,Spark比MapReduce的速度更快且提供的功能更丰富。本文以数据采集、数据存储、数据分析、数据展现为主线,结合大数据领域最流行的Hadoop框架与Spark框架提出了高校校园大数据平台架构,详细阐述了架构各层次的具体功能,并对架构中关系数据库数据的采集存储进行了详细介绍,最后设计校园大数据分析原型系统来验证架构的可行性。

关键词: 大数据; Hadoop; Spark; 校园大数据平台

中图分类号: TP391 **文献标识码:** A

Research on University Campus Big Data Platform Based on Hadoop and Spark

LIU Ping

(Department of Computer Science, Jiangyin Polytechnic College, Jiangyin 214400, China)

Abstract: The analysis of campus big data is a new way of campus information development. Hadoop is a distributed system infrastructure developed by Apache Foundation, which is an ecosystem integrating distributed computing, storage and management. The current popular Spark framework is a distributed computing platform similar to MapReduce in the Hadoop ecosystem, and Spark is faster and more functional than MapReduce. With the main line of data collection, data storage, data analysis and data presentation, this paper puts forward the big data platform architecture of university campus in combination with the most popular Hadoop framework and Spark framework in big data fields, and expounds the specific functions of the architecture at all levels in detail, and gives a detailed description of the data collection and storage of the related coefficients in the architecture. Finally, the campus big data analysis prototype system is designed to verify the feasibility of the architecture.

Keywords: big data; Hadoop; Spark; campus big data platform

1 引言(Introduction)

随着云计算、物联网、WEB搜索、电子商务、个性化信息服务和社交网络的高速发展,产生了大量的数据,其中蕴含的巨大价值成为大数据分析的强大动力。大数据技术也推动了高校校园信息化建设由数字化校园逐渐向智慧化校园转变,在持续建设过程中,校园各个阶段建设了一系列上层应用系统,例如学生管理系统、教务系统、科研系统、招生系统、就业系统、财务系统、人事系统、图书管理系统、后勤管理系统和校园一卡通等,积累的校园数据也在迅猛增加,已初步形成校园大数据环境。搭建大数据平台,对海量校园数据进行分析,集成各个时期开发的各种信息系统中的结构

化数据,并对校园网页等非结构化数据,以及实时的流数据进行采集分析,实现校园中信息的互联互通,并对师生进行个性化服务,为学生学业提供预警,如提前一年毕业资格审核,发送审核报告给学生,对学校教师和科研数据进行分析,为学校管理者制定教学管理方针提供可靠的数据依据,有着重要的意义。

2 Hadoop与Spark介绍(Hadoop and Spark introduction)

目前以Hadoop为核心的大数据解决方案占大数据市场95%以上的份额。Hadoop是一个集分布式计算、存储和管理为一体的生态系统,它为大规模并行数据处理算法提供运行

环境。Hadoop生态系统的核心是分布式文件系统(Hadoop Distributed File System, 简称HDFS)和MapReduce框架, HDFS为海量数据提供了高容错性和高吞吐量的数据存储, MapReduce框架为海量数据提供并行计算^[1]。

Spark是专为大规模数据处理而设计的快速通用的内存计算框架。Spark可以认为是MapReduce的替代方案。Spark中数据被高度抽象且存储在弹性分布式数据集(Resilient Distributed Dataset, 简称RDD)中。它将数据加载到节点内存中, 然后在内存中完成计算。Spark有许多内置操作, 可以将一个RDD转换为另一个RDD, RDD的持久化操作, 可以将它缓存在工作节点的内存中, 这样中间结果可以保存在内存, 从而不再需要读写HDFS。后续重用数据时, 直接从内存读取。它能够快速访问数据, 有效提高了迭代执行的时间效率^[2]。而MapReduce在计算过程中要将中间结果写入HDFS, 这样就需要大量的磁盘I/O操作。

Hadoop的容错方式与Spark很不一样, Hadoop通过复制多份数据进行容错。而Spark无需备份数据, RDD为Spark提供了良好的容错性能, Spark记录RDD上执行的一系列操作, 构建一个有向非循环DAG图, 若数据出错或丢失, 就依据DAG追溯重算^[3]。Spark比MapReduce有诸多的优势^[4], 特别适用于批处理、迭代计算(如机器学习、图计算等)、交互式处理(如数据挖掘)和流式处理(如点击日志分析)等^[5]。

Spark仅仅是计算框架, 它的生态系统并没有提供资源管理系统, 因此需要使用Hadoop YARN, 为上层应用提供统一资源管理和调度, 它为集群在利用率、资源统一管理和数据共享等方面带来了巨大好处^[6]。Spark本身也并没有提供分布式文件系统, Spark的分析大多依赖Hadoop的分布式文件系统HDFS, 它可兼容HDFS和Hive等分布式存储层, 可融入Hadoop生态系统^[7]。因此下文结合Hadoop与Spark构建高校校园大数据平台。

3 基于Hadoop与Spark的校园大数据平台架构 (Campus big data platform architecture based on Hadoop and Spark)

对学校各阶段开发使用的应用系统中的结构化数据进行收集存储, 对校园内海量的半结构化和非结构化数据也进行统一管理和存储, 能够实现学生、教师、科研、招生、就业、财务、人事、图书、后勤、一卡通等信息的统一采集、存储、处理、分析。研究科研数据, 分析最近几年学院的科研动态, 对科研发展方向进行预测; 研究教师数据, 及时关注学校的教学动态, 为学校教育教学方针政策的制定提供数据依据, 通过关联规则分析教务系统中教师的教学工作量与科研系统中教师的科研进展之间的关系; 研究学生数据, 分析学生的学习状况、食宿消费、图书借阅等行为, 对学生毕业资格提前进行审核, 对可能无法正常毕业的学生进行预

警。对校园一卡通的消费记录进行分析, 得出学生的消费水平和消费习惯, 为学院助学金政策的制定和实施提供有效的数据支撑^[8]。

下文结合伯克利APM实验室打造的Spark生态圈(BDAS), 因其是在算法(Algorithms)、机器(Machines)和人(People)之间通过大规模集成来展现大数据应用的一个平台。该生态圈已涉及机器学习、数据挖掘、信息检索、数据库、自然语言处理和语音识别等领域。但该生态圈并没有提供分布式文件系统, 因此Spark的分析可借助与Hadoop的分布式文件系统HDFS; Spark还可以使用Hadoop生态系统中的Yarn进行集群资源管理和调度^[9]。因此结合Hadoop和Spark技术, 提出了基于Hadoop与Spark的校园大数据平台架构, 如图1所示。



Fig.1 Campus big data platform architecture diagram

3.1 校园大数据平台架构分析

针对校园大数据平台架构的各层次功能实现给出如下阐释:

数据源该层包括结构化数据、非结构化数据和实时流数据, 结构化数据主要包括高校校园信息化建设过程中的各应用系统中的数据, 如学生管理系统、教务系统、科研系统、招生系统、就业系统、财务系统、人事系统、图书管理系统、后勤管理系统、资产管理和校园一卡通等; 非结构化数据一般包括网页、微博、微信等; 实时流数据主要包括监控、门禁等。

大数据采集与存储, 该层实现从数据源中抽取数据到HDFS中进行数据分析。数据采集采用Sqoop将关系数据库, 如Oracle、MySQL等中的数据导入Hadoop的HDFS或HBase中, 当然也可将HDFS的数据导出到关系数据库中。Sqoop通过Hadoop的MapReduce导入导出, 因此整个平台拥有很高的并行性, 以及良好的容错性。在HDFS上储存时, 文件会分成64MB为一单位的数据集合, 通过冗余镜像的方式在计算机集群中进行分布式储存。对于结构化明显的可以进行集中储存, 对于结构化不明显的以文件和目录方式进行分布式储存。对于非结构化数据和流数据通过专业的工

具,如网络爬虫等,收集教师、学生的数据,如教学过程、教育管理过程产生的视频、文档、PPT、教案,学生的作业、试卷等信息,以及物联网产生的信息如监控、门禁信息等,这其中包含了大量的文件、照片、视频等数据。

大数据分析,该层使用内存计算框架Spark,Spark提供了大量的库,包括Spark SQL、Spark MLlib、Spark GraphX、Spark Streaming、DataFrames等。本平台将无缝组合使用这些库。使用Spark SQL读取HDFS,并进行分布式数据查询操作,使用Spark Streaming将流式数据转换成RDD,使用Spark GraphX解析运行图计算,使用Spark MLlib设计和实现机器学习算法。运用它协同实现校园大数据的分类、聚类、关联规则等数据挖掘算法。部分使用MapReduce进行离线分析。

数据应用展示,该层为用户提供个性化服务定制功能,可以为用户进行信息推送,用户可以查询分析结果,可以展示科研数据分析、学生数据分析、教师数据分析,以及学生毕业资格审核等。通过建立校园大数据平台得到有价值的信息并以可视化方式,例如通过门户网站、移动APP等展现给用户,并为学生提供学习情况和消费情况分析、为老师提供教科研分析等,分析结果以曲线、二维图像或文件报告等方式展现。大数据分析结果也可在其他系统中展示和应用^[10]。

3.2 关系数据库数据导入HDFS

由于校园大数据平台中很大一部分数据是来源于各个阶段开发的各类信息系统,下面介绍一下在本文提出的校园大数据平台架构中如何将关系数据库导入到HDFS文件系统中,以Oracle数据库为例,其他关系数据库类似。

使用Sqoop将关系数据库数据迁移到HDFS中,首先下载Sqoop压缩包上传到Hadoop集群的节点上,将它解压缩即可使用,然后将连接Oracle的驱动包拷贝到Sqoop目录下的lib目录中,并配置好Oracle数据库的远程连接。准备工作完成后,使用如下步骤和命令把关系型数据库中数据导入HDFS中。

(1)设置Oracle的连接字符串,包括Oracle的主机地址、SID和端口号。

CONNECTURL=jdbc:oracle:thin:@192.168.14.253:1521:ORCL

(2)访问Oracle数据库的用户名和密码。

ORACLENAME=lp ORACLEPASSWORD=123456

(3)设置需要从Oracle数据库中导入的表名,例如此处导入学生表信息。

oracleTableName=Student

(4)需要从Oracle中导入的学生表的字段名。

columns=studentId,studentName,sex,age,classId,telephone

(5)将Oracle中的数据导入到HDFS中的存放路径。

hdfsPath=apps/as/hive/\$oracleTableName

(6)执行sqoop导入逻辑,将Oracle中的数据导入到HDFS中。

```
sqoop import--connect$CONNECTURL--username
$ORACLENAME--password$ORACLEPASSWORD
--target-dir$hdfsPath--num-mappers 1--table
$oracleTableName--columns$columns--fields-
terminated-by'\001'
```

4 原型系统(Protosystem)

为验证校园大数据平台架构的合理可行性,基于Hadoop集群和Spark集群构建大数据处理平台,并采用JavaEE平台的Struts框架和Hibernate框架进行原型系统开发。下面以原型系统中的毕业资格审核模块为例,来验证平台架构的合理性。

学生在毕业前一年,通过原型系统的毕业资格审核模块进行毕业资格的提前预审,并以邮件的形式发送报告给学生,对有科目未通过,或其他影响毕业的问题的学生进行提前预警,以便让学生可以好好利用最后一学年查漏补缺,争取正常毕业。用户通过浏览器访问原型系统,可按系部、班级进行整班查询毕业资格预审,也可具体到某个学生,通过学号或姓名等查询某学生毕业资格预审情况,如图2所示,查询出来的结果集里包含学生基本信息、课程成绩是否全部通过,若有未通过课程,给出未通过课程名称;学生考证情况,会给出具体考了哪几个证书;学生奖惩情况和积分情况,积分若未通过,会给出具体缺多少;还会对学费缴纳情况、图书借阅情况、体育器材借阅情况等进行说明,这些都会影响毕业证书的领取,这些信息分别来源于高校招生系统、教务系统、学生管理系统、财务系统、图书管理系统等,原来这些信息是分别存在于各个应用系统中的,这些应用系统之间数据冗余,不统一,存在一个个“信息孤岛”,不能实现数据共享。通过本文设计的大数据平台架构,把各个应用系统的数据都被抽取到了HDFS文件系统,然后通过大数据分析统一处理,实践表明,该本文提出的大数据平台架构是可行的,也是合理的。

图2 原型系统运行界面

Fig.2 Protosystem running interface

5 结论(Conclusion)

研究大数据技术Hadoop和Spark在教育信息化领域的应用,进一步丰富数字化校园和智慧校园的技术体系,提出校园大数据平台架构,利用大数据技术,对校园信息进行综合的分析整理,更好帮助教师管理学生。在提出的校园大数据平台中采用HDFS完成庞大数据的分布式存储。该平台基于开放架构,具备较好的扩展处理能力,能为多种数据源提供全面可靠的处理。通过该大数据平台对学生历史数据和实时数据进行分析,对学生异常状态进行预警,尤其可以对学生提前进行毕业资格预审,可以促使更多的学生按时毕业。研究科研数据,分析学校科研动态发展;研究教师数据,关注学校教学动态,为教学方针政策的合理制定提供依据。该平台能够充分集成院校不同阶段使用的各信息系统。下一步将重点研究大数据相关算法,例如对图书馆的图书研究推荐算法,读者可以轻松获取与本人兴趣相关的推荐图书等,以及架构中数据利用过程中的安全问题和隐私保护问题。总之,本文提出的校园大数据平台可以加快推进智慧校园的建设。

参考文献(References)

- [1] 常涛.改进型MapReduce框架的研究与设计[D].北京:北京邮电大学,2011:19-22.
- [2] Koliopoulos A K,Yiapanis P,Tekiner F,et al.A parallel distributed weka framework for big data mining using

Spark[J].2015,26(7):9-16.

- [3] 闫梦洁,罗军,刘建英,等.IABS:一个基于Spark的Apriori改进算法[J].计算机应用研究,2017,34(8):2274-2277.
- [4] Gao Yanjie.Data Processing with Spark[M].Beijing:China Machine Press,2015.
- [5] 李涛,刘斌.Spark平台下的高效Web文本分类系统的研究[J].计算机应用与软件,2016,33(11):33-36.
- [6] Landset S,Khoshgoftaar T M,Richter A N,et al.A survey of open source tools for machine learning with big data in the Hadoop ecosystem[J].Journal of Big Data,2015,2(1):24.
- [7] 毛国君,胡殿军,谢松燕.基于分布式数据流的大数据分类模型和算法[J].计算机学报,2017(1):161-175.
- [8] 何清,李宁,罗文娟,等.大数据下的机器学习算法综述[J].模式识别与人工智能,2014,27(4):327-336.
- [9] Shan S.Big data classification:problems and challenges in network intrusion prediction with machine learning[J].ACM Sigmetrics Performance Evaluation Review,2014,41(4):70-73.
- [10] 潘奇.基于Hadoop技术的高校学生行为分析系统研究与实现[D].北京:北京邮电大学硕士论文,2014:1-2.

作者简介:

刘萍(1982-),女,硕士,讲师.研究领域:软件工程,系统集成,大数据分析.

(上接第30页)

机转速,如图5所示。除此之外,还让电机处于高速和低速状态,并按上述要求进行多次实验。实验证明,此系统在30米内有遮挡物的环境下仍能有效地接受数据,具有很高的可靠性。



图5 PC上位机实时接收数据

Fig.5 PC host computer receiving data in real time

5 结论(Conclusion)

针对工业生产中在旋转设备中转速控制和转速采集上的需求,本文提出了一款基于51单片机和无线传输的直流电机调速测速系统的设计方案。该系统运行可靠稳定、成本低且易于实现,能够满足工业生产中电机测速的需求,且利用nRF24L01模块实现了信息的无线传输,摆脱了有线传输带

来的局限性,大大提高了系统的灵活性。基于以上的各种优点,该系统可投入到实际生产应用中去,具有一定的实用价值和推广价值。

参考文献(References)

- [1] 李健,周建.直流电机调速和测速系统设计[J].仪器仪表用户,2016,23(08):28-29;65.
- [2] 尹冬梅.基于单片机控制的直流电机调速测速系统[J].科技风,2011(14):29-30.
- [3] 时志云,盖建平,王代华,等.新型高速无线射频器件nRF24L01及其应用[J].国外电子元器件,2007(08):42-44.
- [4] 王天娇.基于nRF24L01的高效无线传输系统设计[A].中国高科技产业化研究会智能信息处理产业化分会、中国高科技产业化研究会信号处理专家委员会.第十一届全国信号和智能信息处理与应用学术会议专刊[C].中国高科技产业化研究会智能信息处理产业化分会、中国高科技产业化研究会信号处理专家委员会,2017:7.
- [5] 徐灵均,李志宏.基于串口通信的控制系统设计[J].电子技术与软件工程,2017(15):48.

作者简介:

陈庭伟(1997-),男,本科生.研究领域:电子信息工程.

林艺帆(1997-),女,本科生.研究领域:信息管理与信息系统.

刘巧(1980-),女,硕士,讲师.研究领域:嵌入式自动控制.