

# 基于 MPP-Hadoop 混合架构高校数据集成系统研究

邓涵元<sup>1,2</sup>, 卢山<sup>2</sup>, 程光<sup>3</sup>

(1. 武汉邮电科学研究院, 湖北 武汉 430074;

2. 南京烽火软件科技有限公司, 江苏 南京 210019;

3. 东南大学, 江苏 南京 210019)

**摘要:**随着数字化校园的建设,传统的数据集成系统在海量数据环境下数据查询和加载的效率均有所下降,且难以对非结构化、半结构化数据进行融合和分析。针对以上情况,依托高校大数据平台,从各个异构系统中抽取数据,结合 Hadoop 和 MPP 技术的优势,设计并实现了一个基于 MPP-Hadoop 混合框架的高校异构数据集成系统,融合多种不同结构数据,提升了数据查询和加载的效率。以某高校为例,从学生的门禁刷卡系统和校园网系统中抽取学生的行为轨迹数据,载入 MPP 数据仓库,进行数据融合,并传统数据仓库产品 Oracle 搭建的现有高校数据集成系统进行数据加载和数据查询效率方面的对比评测,验证了系统的有效性并且为学生的学习生活、心理等各方面的管理工作提供一定的技术支持和指导。

**关键词:**数据集成;高校大数据;MPP;Hadoop;GreenPlum

中图分类号:TP302

文献标识码:A

文章编号:1673-629X(2018)08-0160-04

doi:10.3969/j.issn.1673-629X.2018.08.034

## Research on University Data Integration System Based on MPP-Hadoop Mixed Architecture

DENG Han-yuan<sup>1,2</sup>, LU Shan<sup>2</sup>, CHENG Guang<sup>3</sup>

(1. Wuhan Research Institute of Posts and Telecommunications, Wuhan 430074, China;

2. Nanjing FiberHome Software Technology Co., Ltd., Nanjing 210019, China;

3. Southeast University, Nanjing 210019, China)

**Abstract:** With the construction of digital campus, the efficiency of data query and loading of the traditional data integration system in the massive data environment are reduced, and it is difficult to integrate and analyze unstructured, semi-structured data in the massive data environment. For this, relying on university large data platform, combining the advantages of Hadoop and MPP technology, we design and implement a system of heterogeneous data integration based on MPP-Hadoop hybrid framework, which integrates many different structure data and enhances the efficiency of data query and loading. And taking a university as an example, the students trajectory data is extracted from the student's access card system and the campus network system and is loaded to MPP data warehouse. The system will be compared with the traditional university data integration system built by Oracle data warehouse, and its validity is verified. Technical support and guidance to students' life, study, psychology and other aspects of management is provided.

**Key words:** data integration; university big data; MPP; Hadoop; GreenPlum

## 0 引言

高校信息化从 20 世纪 80 年代开始,在 21 世纪引入了数字化校园的概念。数字化校园是构建一个包括

教学、科研、管理、服务于一体的数字环境,能够提升传统校园的工作效率,实现教学科研的全面信息化,提高教师的教学质量,提升学校对于各个机构和系统的管

收稿日期:2017-07-31

修回日期:2017-12-14

网络出版时间:2018-03-07

基金项目:国家自然科学基金(61602114);国家“863”高技术发展计划项目(2015AA015603)

作者简介:邓涵元(1994-),女,硕士,研究方向为数据集成、数据挖掘;卢山,博士,副教授,研究方向为计算机、数据集成;程光,博士,教授,博导,研究方向为网络空间安全监测和防护、网络大数据分析。

网络出版地址: <http://cnki.net/kcms/detail/61.1450.TP.20180307.1422.036.html>

理水平<sup>[1]</sup>。随着数字化校园的建设<sup>[2]</sup>,在高校的各个系统中均积累了大量的数据。由于各个异构系统的建设时期不同,采用的标准规范、技术路线编程语言等不同,因而形成了一个的“数据孤岛”。这些“数据孤岛”造成了高校资源分散、存储冗余、管理成本高、决策支持弱、各职能部门无法进行协同工作。

于是将各信息系统的数据整合、汇聚到数据集成系统上就显得尤为重要,这有利于从不同的角度来分析学生的生活、学习、心理等各方面的成长情况,有利于学校各个职能部门的管理和协调。据调研,目前,某高校每年均有约 2 万名学生入学,积累了大量的数据;这些数据不仅数量级大,数据结构也多样,既有学生的基本信息、成绩信息、校园卡消费记录等结构化数据,也有上网记录等非结构化或半结构化数据。

传统的数据集成大多只是简单地以集成后能够查询使用为目的进行数据集成。当前研究者对校园数据集成和分析做了多方面的研究。吴振涛<sup>[3]</sup>提出了基于数据仓库的通用的校园数据集成框架,为高校的数据仓库建设提供了一个整体框架和模型;王晶春<sup>[4]</sup>对高校数据集成广泛应用的几类框架进行综合的比较,提出集线器总体架构模式;李兰友等<sup>[5]</sup>提出了基于 ODI 技术的高校数据流转运中心架构模式。然而,这些研究都是基于传统的数据集成技术,大多是主机加集中存储的架构。软件方面则主要选择 Oracle 相关数据库产品来搭建数据仓库。而随着数据的迅猛增长以及数据类型复杂程度的增加,传统的数据仓库产品的不足也逐渐凸现出来:处理数据量小,投资成本高,扩展性差,数据加载和查询效率低,针对非结构化数据的特征提取、多数据融合分析遇到困难。

针对以上问题,结合 Hadoop 和 MPP 技术,文中设计并实现了一个基于 MPP-Hadoop 混合框架的高校

异构数据集成系统,将数据融合、汇总、分析,提高数据查询和加载的效率,提高扩展性,并通过实验验证该系统的有效性。

## 1 相关技术

### 1.1 Hadoop

Hadoop<sup>[6]</sup>是一个分布式的系统基础架构,能够充分利用集群进行高速运算和存储。Hadoop 有高可靠性<sup>[7]</sup>、高效性、可扩展性、高容错性等优势<sup>[8]</sup>。

### 1.2 MPP 数据库

Hadoop 的优势在于能十分高效地处理大量的非结构化或半结构化数据。但与传统的关系型数据库相比,在处理复杂的多表关联分析、数据分析挖掘以及易操作性方面还存在差距。

MPP(massive parallel processing)<sup>[9]</sup>数据库本质上仍然是一个关系型数据库。它可以将任务并行地分散到多个工作节点上,磁盘存储系统和内存系统均为每个节点独有,不与其他节点共享,是 share-nothing<sup>[10]</sup>模式,各个节点之间通过网络互相连接,彼此协同计算,将各自的结果汇总到一起得到最终结果。与传统的关系型数据库相比,MPP 数据库在数据处理方面,具有采用分布式架构<sup>[11]</sup>、处理数据量大、更大的 I/O 能力、扩展能力好、采用列式存储<sup>[12]</sup>、节约存储空间等优势。

## 2 系统设计与实现

### 2.1 系统框架设计

该面向高校的异构数据集成系统结合 Hadoop 和 MPP 两种技术的优势,架构设计如图 1 所示,总体上分为数据层、应用层和数据源层。

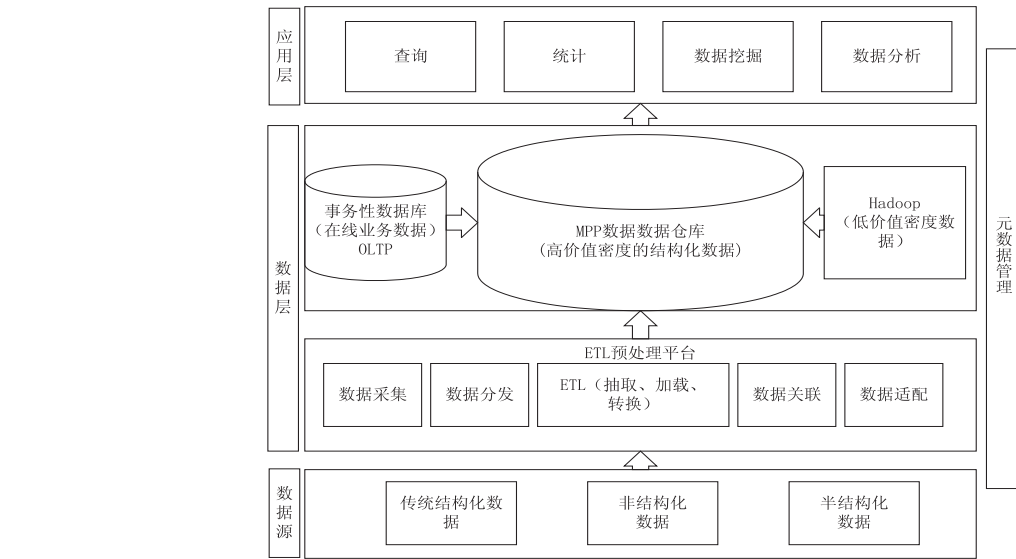


图 1 异构数据集成系统框架

(1)数据源即包含来自各个异构系统的数据,大致可以分为两块,一是来自传统的结构化数据,二是来自海量非结构化或者半结构化的大数据(如上网日志数据)等。

(2)数据层包括数据存储平台和 ETL 数据预处理平台两部分 ETL 数据预处理平台从本地 FTP 服务器中抽取相关基础数据,抽取方式分为全量抽取和增量抽取两种。全量抽取将所有的初始数据抽取到目标数据库中,增量抽取针对新增数据,时间间隔为 24 小时自动抽取。抽取后的数据需要进行清洗、转换和加载,去除噪声数据、转换数据格式、生成新的数据等。处理后的数据一部分进入传统的数据库中,一部分存储到 Hadoop 的 HDFS 中,再将两部分的数据整合、汇聚到 MPP 数据仓库中,完成数据的基本存储功能。

(3)在大数据背景下,简单地将异构的数据源集成起来实现查询已经不能满足现实要求,应用层的功能除了查询和展示结果外,还有分析和挖掘、生成报表等。在高校应用场景下,主体主要是教师、学生以及职工,应用层将以不同的主体,不同的数据分析需求实现不同的功能。

(4)元数据管理对各个来源的数据进行集中管理,构建元数据模型,能够更加有效地对数据质量进行把控,更高效地进行数据集成分析和挖掘。

## 2.2 系统实现

传统的数据仓库基本上都基于 Oracle 产品<sup>[13]</sup>,数据加载速度、数据查询效率在海量大数据情况下变慢甚至无法加载或者查询。GreenPlum 是 EMC 推出的大规模并行处理(MPP)的数据库软件,是一个基于 PostgreSQL 的开源分布式数据引擎,是目前业界研究和应用比较广泛的数据仓库引擎之一。它具有高并发支持、良好的线性扩展能力、高性价比、高可用性等优势。结合学生行为轨迹数据的特点,选用 GreenPlum 作为数据仓库来搭建面向学生行为轨迹数据分析的数据集成系统,并传统数据库 Oracle 产品搭建的数据集成系统对比。

在 x86 平台上分别搭建 Hadoop 集群平台和 MPP 数据库平台。Hadoop 集群由 1 台控制节点和 3 台数据处理节点组成,操作系统采用 Linux5.5,Java 环境的 JDK 版本为 jdk-1.7。MPP 分布式环境由一个 Master (主节点)和多个 Segment (数据节点)组成,每个节点配置 2 个 CPU 内核,8 GB 内存,节点之间使用千兆网络连接。操作系统为 Linux5.5,选择的 MPP 数据库版本为 Greenplum-db-4.1.1.3,将具有高并发性、高可用性优势的 MPP 数据库作为数据仓库对海量数据进行集中的管理和存储,结合 Hadoop 集群的高速存储和运算的特点,搭建系统的物理组网架构,如图 2

所示。

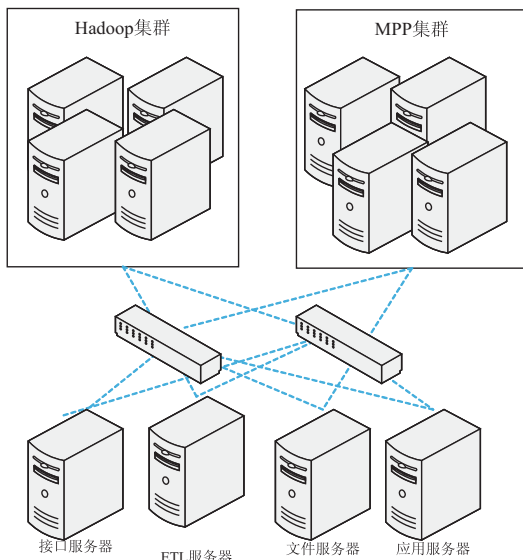


图 2 物理组网架构

## 3 学生行为轨迹数据应用分析

大学生群体作为一个特定的、庞大的社会群体,其轨迹行为具有很强的规律性。对于个体而言,掌握其行为轨迹规律对于掌握其学习、生活的规律和习惯有一定的帮助,对于出现的异常能够起到一定的指导作用。而对于群体而言,了解校园热点地区分布<sup>[14]</sup>,对学校的后勤工作有一定的指导作用。对学生轨迹的相似度进行分析,对于了解学生的线下社交<sup>[15]</sup>、好友发现、排除孤独症患者等有一定的意义,为学生心理健康的管理工作提供一定的依据。据了解,某高校有在校本科生 2 万余名,每名学生每天均产生大量的轨迹数据,在数据加载和查询方面进行对比评测。

### 3.1 数据源选择

目前,某高校宿舍、图书馆、体育馆均设有门禁设备,但是某些区域如教学楼没有门禁,而且门禁可能存在代刷、漏刷的现象,所以,单以门禁记录来研究学生的轨迹不够准确。通过调查发现,该高校教学区、宿舍、食堂、休闲区域均有 wifi 部署,随着校园移动设备用户的爆炸式增长,产生了大量的 wifi 位置记录数据。而且,wifi 数据对研究学生的校园行为的位置,具有覆盖范围广、定位精度高、成本低等特点。但是,单以 wifi 数据来研究学生的轨迹数据,则可能存在学生进入某区域未使用 wifi 连接网络造成轨迹数据缺失的情况。故结合校园卡门禁刷卡数据和 wifi 上下线信息,能够比较真实地反映学生的轨迹信息。

该实验选取的数据来自某高校大一学生 2016 年一学期即 4 个月的 wifi 上下线日志数据,约 1 500 万条,以及学生校园卡刷卡记录,约 600 万条,总量大小为 65.4 GB。

3.2 数据清洗及预处理

由于获取的校园学生 wifi 登录数据信息量大且复杂,包含字段较多,而真正有价值的只有几个字段,因此首先需要对原始数据进行过滤和筛选。为获取学生行为轨迹数据,主要有用的字段是学生学号、wifi 登录时间、校园位置 Id、刷卡时间、刷卡位置 Id。

(1)对获取到的原始数据进行统计分析,系统中存在一定时间内反复刷卡或者反复登录的情况,设置时间阈值  $\Delta t = 1\text{ min}$ ,过滤掉连续刷卡以及在某个时间反复连接 wifi 的数据。

(2)根据用户名过滤到教职工以及其他账号信息,只保留学生的数据。

(3)门禁刷卡记录的原始数据中记录了用户登录时所在校园位置的 ID 号;wifi 上下线日志数据中存储的位置信息用经度 (longitude) 和纬度 (latitude) 字段表示,结合百度地图 API<sup>[16]</sup> 和文献[14]提出的基于校园环境的逆地址解析算法进行校园位置的转换。

经过清洗后存储到数据库中的变量名及变量的含义如表 1 所示。

表 1 变量名及含义

变量名	变量含义
sId	学生学号
LoginTime	wifi 登录时间
LocalId	校园位置 Id
SwipTime	刷卡时间
SwipLocal	刷卡位置

3.3 性能对比评测

3.3.1 数据加载

将预处理后的数据采用外部表的方式分别加载到原系统和现系统中,加载速度对比如表 2 所示。

表 2 数据加载速度对比

指标	原系统		现系统	
数据节点个数	2	4	8	
数据加载所需时间/s	918	128	83	51

3.3.2 数据查询

(1)在硬件配置相同的情况下,该系统采用 4 个节点与现有系统进行查询复杂度对性能影响的比较,在查询复杂度  $Q_3 > Q_2 > Q_1$  的情况下,查询效率对比如图 3 所示。

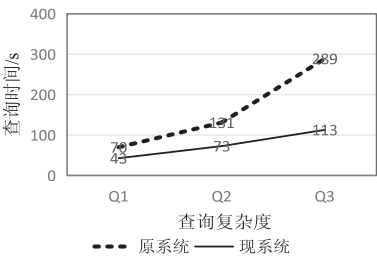


图 3 数据查询效率与复杂度的关系

(2)将同样的数据分别加载到现有系统和文中系统后,测试对比多表连接查询语句 (join) 的查询性能,如表 3 所示。

表 3 查询效率对比

数据库类型	现有系统		文中系统	
数据节点个数	2	4	8	
查询时间/s	176	113	87	42

从实验结果来看,与传统数据库的处理方式相比,文中系统在数据加载和查询效率上有明显的提升,并且具有良好的扩展性,查询效率随着数据节点的增加近乎呈线性增长的趋势。但是在数据量不大的情况下,现有系统查询处理效率不比文中系统低,在海量数据处理的情况下,文中系统能够体现出强大的处理能力。故搭建的异构数据集成系统在高校的海量数据环境下在数据处理和分析上具有明显的优势。

3.4 学生行为轨迹数据分析

为了分析学生轨迹变化,引入统计学中相似度的概念。相似度使用以度量两组数据变化趋势相似程度的一个数值亮度,其取值范围为  $[-1, 1]$ 。相似度的计算方法基于统计学中相关系数的概念。

相关系数是变量之间相关程度的指标,相关系数  $(r(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}[X] \text{Var}[Y]})$  的取值范围是  $[-1, 1]$ 。该参数的值表示变量之间线性相关的程度。对学生每月的轨迹数据进行拟合,观察相关系数的变化  $i$ ,若是波动较大,则可查看学生当前周和月的轨迹变化曲线,判断学生的行为轨迹是否正常。

另外,结合可视化技术比较不同学生的行为轨迹数据,构建班级社交网络,发现学生的社交关系,避免大量的问卷以及人工调查的繁琐工作,分析出学生的社交情况,便于对社交能力弱的学生提供帮助,结合心理资源库中的测评结果,筛选较为孤僻的学生,能够为高校中的学生心理健康教育工作提供指导。

4 结束语

结合 MPP 和 Hadoop 技术,将数据从业务系统中抽离出来,提出一种基于 MPP-Hadoop 混合架构的高校数据集成的系统框架,实现业务系统间的数据共享,充分发挥数字化校园的整体协同功能,解决了传统数据库在海量数据情况下数据加载慢、数据查询效率低、难以融合多种异构数据源进行分析等问题。并以学生行为轨迹数据的分析为例,验证了系统的有效性,为学生的管理工作提供支持。

参考文献:  
[1] 方 园,高润生,徐国徽. 数字化校园环境下一卡通系统  
(下转第 169 页)

4 结束语

在物联网技术不断发展和空巢老人无人看护的环境下,设计实现了一种面向智能家居的老人看护系统。采用 Zigbee 技术、异常值处理、特征提取实时地获取传感事件序列,基于 Zigbee 开源协议栈 Z-Stack 2007 和智能网关实现传感节点部署采集和异构网络互联互通;利用固定时间滑动窗口法和隐式马尔可夫模型预测老人基本日常行为 ADLs,并提供历史记录查询和远程监控功能。实验结果表明,该系统能有效减少通信开销,降低传输时延,实现了智能家居环境下老人的看护。

参考文献:

[1] 秦 丽.《2016 中国智能家居产业发展白皮书》发布[J]. 电器,2016(4):32.

[2] 王 怡,鄂 旭.基于物联网无线传感的智能家居研究[J]. 计算机技术与发展,2015,25(2):234-236.

[3] ESCH J. A survey on ambient intelligence in healthcare[J]. Proceedings of the IEEE,2013,101(12):2467-2469.

[4] BILGIN B E, GUNGOR V C. Performance evaluations of ZigBee in different smart grid environments[J]. Computer Networks,2012,56(8):2196-2205.

[5] WANG Wei, HE Guangyu, WAN Junli. Research on Zigbee wireless communication technology[C]//International conference on electrical and control engineering. Yichang, China: IEEE,2011:1245-1249.

[6] 张 毅,徐菲菲,雷景生,等.基于 CC2530 和 ZigBee 技术智能家居系统的设计与研究[J]. 上海电力学院学报,

2017,33(2):191-195.

[7] 孙建书. ZigBee 网络能耗优化机制的研究与实现[D]. 北京:北京邮电大学,2015.

[8] HU Yang, TILKE D, ADAMS T, et al. Smart home in a box: usability study for a large scale self-installation of smart home technologies[J]. Journal of Reliable Intelligent Environments,2016,2(2):93-106.

[9] COOK D J, CRANDALL A S, THOMAS B L, et al. CA-SAS: a smart home in a box[J]. Computer,2013,46(7):62-69.

[10] COOK D J, YOUNGBLOOD M, HEIERMAN E O, et al. MavHome: an agent-based smart home[C]//Proceedings of the first IEEE international conference on pervasive computing and communications. Fort Worth, TX, USA: IEEE, 2003:521-524.

[11] KWAPISZ J R, WEISS G M, MOORE S A. Activity recognition using cell phone accelerometers[J]. ACM SIGKDD Explorations Newsletter,2010,12(2):74-82.

[12] 李云洁. 基于连续型传感器数据的人体动作识别[D]. 上海:东华大学,2015.

[13] SINGLA G, COOK D J, SCHMITTER-EDGEcombe M. Recognizing independent and joint activities among multiple residents in smart environments[J]. Journal of Ambient Intelligence & Humanized Computing,2010,1(1):57-63.

[14] 宋 涛. CS 体系的传统二层结构与流行三层结构的比较分析[J]. 硅谷,2012(9):135.

[15] SHI Weisong, CAO Jie, ZHANG Quan, et al. Edge computing: vision and challenges[J]. IEEE Internet of Things Journal,2016,3(5):637-646.

(上接第 163 页)

[J]. 华中师范大学学报:自然科学版,2017,51(S1):156-160.

[2] 沈培华,王映雪,蒋东兴,等.清华大学数字校园建设与思考[J]. 管理信息系统,2002,75(2):18-19.

[3] 吴振涛. 基于数据仓库技术的数据集成在数字化校园中的应用[J]. 电子设计工程,2016,24(9):28-31.

[4] 王晶春. 数字化校园数据集成总体架构浅析[J]. 长春理工大学学报:自然科学版,2015,38(3):148-151.

[5] 李兰友,陈 立,陈建红. 基于 ODI 的数字校园数据集成研究与应用[J]. 南京工程学院学报:自然科学版,2016,14(2):29-34.

[6] 王 峰,雷葆华. Hadoop 分布式文件系统的模型分析[J]. 电信科学,2010,26(12):95-99.

[7] GAUTAM J V, PRAJAPATI H B, DABHI V K, et al. Empirical study of job scheduling algorithms in Hadoop MapReduce[J]. Cybernetics and Information Technologies,2017,17(1):146-163.

[8] JENA B, GOURISARIA M K, RAUTARAY S S, et al. A survey work on optimization techniques utilizing map reduce

framework in Hadoop cluster[J]. International Journal of Intelligent Systems and Applications,2017,9(4):61-68.

[9] CHANG F, DEAN J, GHEMAWAT S, et al. Bigtable: a distributed storage system for structured data[J]. ACM Transactions on Computer Systems,2008,26(2):205-218.

[10] 音 春. 大数据时代数据库技术研究[J]. 广东通信技术,2015,35(3):12-14.

[11] 辛 晃,易兴辉,陈震宇. 基于 Hadoop+MPP 架构的电信运营商网络数据共享平台研究[J]. 电信科学,2014,30(4):135-145.

[12] 周润松. 大数据 MPP 产品测评研究[J]. 软件和集成电路,2016,23(8):36-37.

[13] 赵 闯. 构建数字化校园数据仓库的方案研究[D]. 长春:东北师范大学,2009.

[14] 杜胜兰,李 枫,黄长青,等. 基于轨迹数据的武汉大学学生行为规律分析[J]. 测绘地理信息,2017,42(1):91-95.

[15] 鲁鸣鸣,张 丹,王建新. 基于校园一卡通数据好友发现及应用[J]. 大数据,2017,3(2):78-91.

[16] 杜传明. 百度地图 API 在小型地理信息系统中的应用[J]. 测绘与空间地理信息,2011,34(2):152-153.