DOI:10.19344/j.cnki.issn1671-5276.2020.01.028

浅析基于 Hadoop 的高校大数据云平台设计

孔德丽,屈会雪,卞志勇
(南京机电职业技术学院,江苏 南京 211135)

摘　要:为加强高校数据中心的建设,基于 Hadoop 构建的学院大数据云平台设计,对结构化与非结构化数据存储进行设计并优化。在满足学院大规模业务数据存储需求的同时,还提供了强大的云计算能力,为大数据时代的数据挖掘提供了高效的计算模型,提高了学院工作效率。

关键词:Hadoop ;云平台;大数据;设计
中图分类号:TP311.13　　**文献标志码**:B　　**文章编号**:1671-5276(2020)01-0101-02

Design of University Big Data Cloud Platform Based on Hadoop

KONG Deli, QU Huixue, BIAN Zhiyong
(Nanjing Electromechanical Vocational and Technical College, Nanjing 211135, China)

Abstract: To strengthen the construction of data centers in universities, this article designs the college big data cloud platform based on Hadoop and the structured and unstructured data storage, which meet the powerful cloud computing capabilities. The high efficient computing model is provided for data mining in the era of big data, thus improving the college's work efficiency.

Keywords: Hadoop; cloud computing; big data; design

1　现状与价值

1.1　研究现状

目前,高校数据中心建设主要分为两种形式:一是直接使用传统服务器搭建数据中心的存储及应用。服务器的架构价格高,能源消耗高,资源利用率偏低。二是采用专业的 VMware 虚拟化软件对 IT 基础设施虚拟化成资源池供各类应用部署。

1.2　建设价值

基于 Hadoop 的高职院校大数据平台[1-2],避免了传统数据中心的各种弊端,不仅可以充分发挥集群的威力,还能够充分挖掘大数据中的隐藏信息,在职业院校决策、管理与服务中可以得到更广泛的应用,以提高学院的工作效率。

2　云计算大数据平台设计

2.1　总体架构设计

本项目是在 linux 开发环境下开发的,大数据云平台总体架构及各分平台的设计思路及大数据云平台总体架构设计如图 1 所示。

2.2　数据云平台模块设计

以 Hadoop 为基础构建的大数据云平台,可以很方便地与现有各类业务信息系统实现集成。

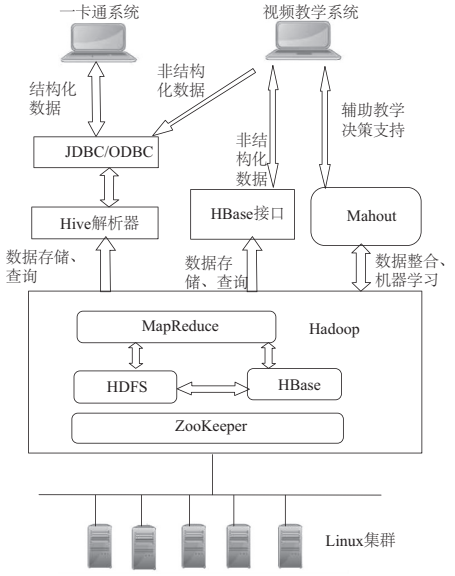

图 1　大数据云平台总体架构设计

1) 结构化数据的存储与设计

高职院校现有数据库大多为关系型数据库,HBase 数据库对结构化数据支持不足。为解决这个问题,以Apache 开源软件开发的 Hive 数据仓库工具作为接口,将结构化的数据文件映射为 HBase 数据库的一张表,并提供完整的 SQL 查询功能。还可以将 SQL 语句转换为MapReduce任务来运行,充分利用并行运算的速度优势。使用 Hive,不

基金项目:江苏省教育信息化研究 2017 年度课题(20172032)
作者简介:孔德丽(1982—),男,江苏泰州人,讲师,硕士,研究方向为计算机科学与技术、网络及软件工程、数据信息管理。

仅解决了结构化数据的存储问题,而且提高了数据查询和存储的速度,提高了工作效率。下面以一卡通信息系统中常用的结构化数据为例,针对结构化数据,使用 Hive 数据仓库平台,实现数据的存储和查询。结构化数据存储设计如图 2 所示。

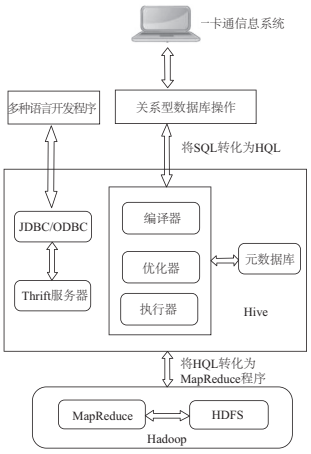

图 2　结构化数据存储设计

2）非结构化数据的存储设计

在视频教学系统中会产生大量视频或图片文件,常规存储方式速度太慢。以典型的 7 200 r/min 的硬盘为例,其最大传输速度约为 22.3 MB/s,这极大地限制了映像文件的存储和查询速度,而以分布式存储能解决这个问题。本文以视频教学系统视频和图片等文件存储为例,阐述如何解决非结构化数据存储过程中的核心问题。

非结构化数据存储系统设计在视频教学系统中,单个课堂教学视频文件往往会有数百兆大小的文件。以 HBase 存储保证数据安全性的同时不仅可以实现数据的快速查询和存储,还能实现对历史视频或图片文件的快速查询。非结构化数据存储系统架构设计如图 3 所示。

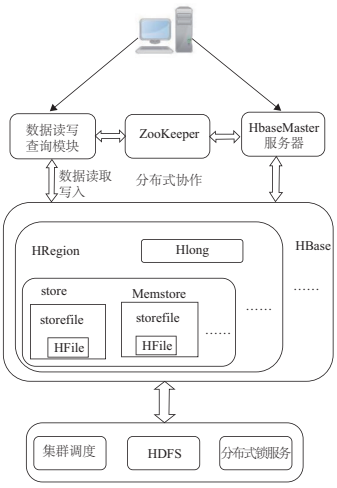

图 3　非结构化数据存储系统架构设计

3）机器学习平台的设计

以 Hadoop 为基础的分布式计算对大数据的机器学习具有更高的效率。在本项目中使用开源项目 Mahout 提供的机器学习工具[3],开发机器学习模块,在此模块基础上采用机器学习的结果为教学工作者提供建议,优化教学流程,还可以作为数据挖掘的基础为更高级功能提供基础架构。基于 Hadoop 的机器学习平台如图 4 所示。

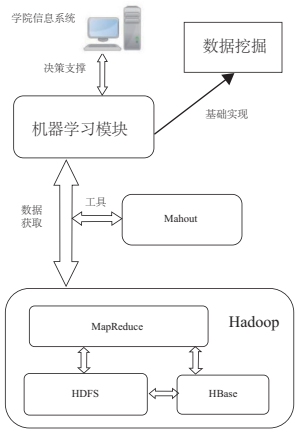

图 4　机器学习平台设计

机器学习的系统底层是 Hadoop 框架,从 HBase 和 Hive 中获取数据,通过 MapReduce 进行分布式计算提供机器学习算法,利用 Mahout 数据进行整合、清理,然后进行机器学习[4]。获取数据之后,使用 Mahout 提供的算法,这些算法包括朴素贝叶斯分类、支持向量机、随机森林等分类算法和 EM 聚类、K-means 等聚类算法,利用其 API 实现对数据中心数据的有效利用。

2.3　数据共享交换平台

数据共享交换平台主要包括以下几个部分:数据交换引擎;安全管理服务;系统管理服务;Web 服务管理;Service接口;中心数据库。将分散建设的若干应用信息进行整合,进行数据传输与共享。此平台可以接入学院所有业务系统的数据,在中心数据库汇聚,并为学院各部门提供数据业务协同功能,推动智慧校园建设。

3　结语

基于 Hadoop 构建的学院大数据云平台,对结构化与非结构化数据存储的优化设计,不仅满足了学院大规模业务数据存储的需求,还提供了强大的云计算能力。学院数据中心存储了教务信息、学工信息、科研信息、招生信息、就业信息等各类业务系统的海量信息,利用这些信息可以对教学改革、科研方向规划、招生宣传、专业设置、就业导向等提供数据支持,辅助学院高层领导决策。

参考文献:

[1] 唐燕, 刘仁权, 王苹. 基于 Hadoop 的高校大数据平台的设计与实现[J]. 信息技术, 2017(12):113-117.

[2] 苏叶健. 教学资源共享平台云存储性能优化设计[J]. 电脑知识与技术, 2018,14(33):24-25.

[3] 熊刚. 基于 Hadoop 的大数据存储系统的设计与实现[D]. 南昌:江西师范大学, 2014.

[4] 郭双宙. 基于 Hadoop 和 Mahout 的分布式推荐引擎的设计[J]. 科技情报开发与经济, 2014, 24(7):119-121.

收稿日期:2019-11-26