高速公路大数据处理现状与挑战①

杨仁怀[1], 郎川萍[1], 刘文美[2]

[1](四川交通职业技术学院 计算机工程系, 成都 611130)
[2](四川省交通运输厅 高速公路监控结算中心, 成都 610041)

摘 要: 高速公路在日常的运营过程中, 产生了海量的、异构的数据, 即大数据, 这些数据还在快速的增长. 大规模产生的数据, 给数据的存储和分析带来了巨大的挑战, 如何科学、高效的存储这些大数据, 并能对其进行快捷的访问和分析, 更好的服务于交通, 是一个迫在眉睫的问题. 首先讨论了高速公路中大数据的来源以及数据的特点, 然后分析了研究大数据的意义和大数据处理技术, 并分析了这些技术在高速公路大数据中分析中所面临的挑战.

关键词: 高速公路; 大数据; 并行数据库; MapReduce

Current Status and Challenges of Highway Bigdata Processing

YANG Ren-Huai[1], LANG Chuan-Ping[1], LIU Wen-Mei[2]

[1](Department of Computer Engineering, Sichuan Vocational and Technical College of Communications, Chengdu 611130, China)
[2](Sichuan Communication Department, Highway Supervision & Accounts Settlement Center, Chengdu 610041, China)

Abstract: Massive Heterogeneous Data, aka, the Big Data is being generated along with Highway daily operation, and the volume increase in fast speed. As a consequence of these massive data, data storage and analyze encoutered big challenges. How to store data in high efficency, approach of easy access and feasibility of data analyzing is an urgent issue to be addressed. This article will investigate the source of big data generated by Highway daily operation and its characteristics. Then this paper discusses the significances of big data research and data processing technologies in industry, also the challenges will be mentioned in detail.

Key words: highway; bigdata; parallel database; MapReduce

1 引言

近几年, 高速公路发展迅速. 截止 2011 年底, 我国建成通车的公路总里程为 410 万公里, 其中高速公路为8.49万公里[1]. 预计到“十二五”末, 建成通车的高速公路总里程将达到 10.8 万公里[1]. 高速公路在全国公路网中起着极其重要的作用, 有“经济大动脉”之称, 能够有力的带动和促进区域经济的发展, 有着良好的经济和社会效益[2]. 高速公路的建设目标是构建便捷、安全、经济和综合的运输体系[3], 交通部《公路水路交通运输信息化“十二五”发展规划》明确指出, 全面提高交通运输智能化、现代化的水平[3], 在高速公路交通安全应急、出行服务、市场监管、决策支持等方面进行重点建设[3]. 这就需要对产生于高速公路中的大数据进行深入的分析, 以便于从这些数据中发现知识[4], 将高速公路建成“智慧高速”, 服务于决策者, 提升高速公路的运营效率和管理水平, 服务于大众, 使得出行更顺畅.

本文在第 2 节中重点介绍了高速公路中大数据的来源以及这些数据的特点; 第 3 节介绍了现行大数据的处理技术以及大数据在高速公路中的应用; 第 4 节归纳分析了在处理高速公路大数据方面的机遇与挑战; 第 5 节对全文进行总结.

① 收稿时间:2013-12-31;收到修改稿时间:2014-03-12

2 高速公路中的大数据

2.1 高速公路中的大数据来源

高速公路中的大数据主要有以下几个来源:

(1)高速公路联网收费软件数据: 每一条高速公路上都有数量不等的收费站, 而每一个收费站又有几条甚至几十条收费车道, 收费车道上运行的收费软件产生了大量的数据, 经过长时间的累积, 数据量非常庞大.

(2)应用系统数据: 高速公路监控结算中心的清分系统、12122 呼叫系统、各个收费站和各路公司运行的稽核软件、收费站管理软件和复合卡动态管理软件等也在累积大量数据.

(3)传感器数据[4]: 遍布在高速公路上路感线圈、标识站以及收费站出入口的 RFID 传感器, 不断的对过往车辆进行感知, 持续生成数据.

(4)视频监控系统的数据: 高速公路两侧、隧道中以及收费站的视频监控探头, 特别是高清探头产生了大量的视频数据.

国家经济快速发展, 人们生活水平不断提高, 购买私家车的家庭越来越多, 近几年快递业也在不断发展壮大, 这些都促使汽车保有量不断增长. 同时, 高速公路里程也不断增加, 大数据的生成速度也在加快, 需要处理的数据越来越多.

以上的大数据从其内在结构上可以分为结构化数据和非结构化数据[5]两大类. 结构化数据主要是高速公路联网收费软件、各应用系统产生的数据, 这些数据主要存储在关系数据库中, 如 SQL Server 和 Oracle 中. 没有存放在关系数据库中的数据如视频监控数据、图片数据等称为非结构化数据. 据统计, 在整个高速公路网中, 非结构化的数据所占的比重高达 80%以上, 传统的关系数据库处理非结构化数据非常困难.

2.2 高速公路中大数据的特点

高速公路中的大数据满足 4V 定义[6]: 即规模大(volume)、变化多样(variety)、价值密度低(value)和高速性(velocity)[7].

(1)数据规模: 以四川高速为例, 每一天高速路网中, 经过收费站的过往车辆达到了 200 万辆, 车辆经过收费站时, 监控系统都要拍照, 每张图片大小约为 500k, 每年产生的静态图片大小约为 300T. 四川高速公路联网收费始于 2001 年, 加上收费系统、各应用系统、传感器以及视频监控系统累积的数据早已达到了 PB 级以上, 并仍在持续增长中.

(2)数据多样: 高速公路中的数据种类繁多. 有收费系统、各应用系统产生的结构化数据, 视频系统产生的视频数据, 收费站出入口抓拍的静态车牌图片以及传感器等非结构化数据. 此外还有诸如交通事件、交通环境、交通管制以及高速公路相关联的其他信息, 如服务区数量, 服务区打开还是关闭、收费站是否正常放行、收费车道数量及周边道路是否畅通等信息, 以及是否有地质灾害, 天气信息等数据.

(3)数据价值密度: 数据量大, 但是有用的数据却较少. 例如, 视频监控系统全天 24 小时进行监控, 产生了大量的数据, 如发生了交通意外、地质灾害等, 有用的数据只有几秒钟或几分钟. 因此, 对异常数据的检测则可分析出路网的异常状况.

(4)高速性: 数据的处理速度要及时高效. 如收费站的收费数据要及时、快速的到达结算中心, 某路段发生了交通拥堵, 系统要能及时获得这些信息并分析处理, 不能有较长时间的延长[8].

对大数据的分析并不是指简单的生成报表. 传统的数据仓库技术和各种 BI 工具可以舍弃不重要的数据建立数据集市能对数据进行分析[9], 主要进行 OLAP 分析(上卷、下钻、汇总、切片和旋转等)[4], OLAP 分析已经明显不够用了. 同时数据仓库具有“面向主题”[10] 的特性, 这就决定数据仓库的主题是不易变化的, 这种模式难以分析处理变化的业务环境. 数据仓库是典型的关系型数据库技术的延伸, 难以应当对海量的数据.

3 大数据处理技术

3.1 大数据研究的意义

随着社会的进步, 人们需要对大量累积的数据进行连续的、宏观的分析, 以发现知识, 给人们的决策提供有力的数据支撑.

在国外, eBay 购物平台每天产生的数据量达到了 100PB, eBay 使用大数据处理技术对每一条数据进行跟踪分析, 以便准确掌握用户的购物行为[11]. 沃尔玛是最早利用大数据的企业之一, 其在大数据方面的投资, 正在逐渐产生回报, 如沃尔玛使用大数据技术分析 Source 和 Carlie Brown 两个超市顾客的购买意向正在向高档产品转移, 并及时调整了两家店的库存, 一举将销售业绩提升了 40%[12]. 在国内, 阿里旗下的淘

宝网每天新增的数据达到了 10TB, 2013 年的“双十一”购物节, 淘宝支付宝成交额达到了 350.19 亿元人民币, 比 2012 年增长了 159.19 亿[13], 淘宝之所以能取得如此好的成绩, 主要归功用其对历史数据的分析, 特别是用户的消费习惯、搜索习惯以及浏览习惯等数据所进行的综合分析[14]. 国内 B2C 电商京东商城, 凭借多年以来累积的销售数据, 通过对用户购买行为的分析, 充分的利用和整合了电商和手机生产厂商的资源, 推出了“JDPhone”计划, 既节约了制造成本, 又为用户提供了最佳体验的产品, 恰到好处的满足了用户的需求[15].

从上可以知道大数据在电子商务行业已经成功运用, 并取得了良好的效果, 但在交通行业, 大数据技术的应用还需要进一步加强. 通过对高速公路大数据的分析能够全面提高高速公路的智能化、现代化的水平, 可以在疏导交通, 缓解拥堵、道路安全预警、应对自然灾害(如台风、大暴雨、冰雪路面等)、联网收费结算、绿色出行, 节能减排、突发事件应急等等方面发挥重要的作用.

3.2 并行数据库

并行数据库(如 Oracle、DB2 等)起源于 20 世纪 80 年代[9], 以关系的形式存储结构化数据, 都支持标准的 SQL, 通过 SQL 语言, 并行数据库与外界可以很好的交互. 在过去的 30 年里, 并行数据库取得了很好的发展, 直到现在其功能也在不断扩充和增强. 然而, 随着高速公路网越来越大, 其累积的数据已经大大超出了并行数据库的处理能力, 同时, 并行数据库对非结构化数据(图片、视频)支持较差. 并行数据库在大数据面前已经显得力不从心, 最为主要的原因是并行数据库的扩展性较差. 并行数据库性能扩展主要通过纵向扩展(scale up)和横向扩展(scale out)来实现. 纵向扩展(scale up)是指提升单个节点的硬件性能, 如增加或更换性能更好的 CPU, 扩大内存和硬盘, 这种方式并不能无限制的提升单个节点的处理能力. 横向扩展(scale up)是指增加计算机节点数量, 形成集群, 将数据库部署到集群上以提升并行数据库的处理能力. 这种方式对单个节点硬件的要求较为苛刻, 如果某一节点的性能较其他节点低, 则会影响这个集群的处理能力, 极端情况下会出现, 集群处理能力还不如单个节点的情况. 如果想要达到规模较大的集群, 代价比较高昂.

3.3 云计算

云计算是在分布式计算(Distributed Computing)、并行计算(Parallel Computing)和网格计算(Grid Computing)的基础上发展起来的[16], 其核心的技术是海量数据的存储以及数据的并行处理, 是一种可行的处理大数据的技术.

云计算的数据存储技术主要有两种: (1)Google 公司的分布式文件系统 GFS(Google File System)[17], 使用廉价的服务器搭建的集群, 具有良好的性能、高可用性以及高扩展性, 但并不开源; (2)Hadoop 的 HDFS(Hadoop Distributed File System)[18], 具有和 DFS 相似的功能, 是开源系统. 很多互联网公司, 包括雅虎、淘宝等都使用 HDFS 技术存储数据[19].

为了能更好的处理大数据, 需要使用特定编程模型. MapReduce[20]是 Google 在 2004 年提出的用于处理大数据的编程模型. MapReduce 简化了分布式编程的复杂性, 程序员只需要关心程序的逻辑实现, 而复杂的并行处理以及任务调度交由系统完成. 用户在编程时只需要实现 Map 函数和 Reduce 函数, Map 函数指定需要处理的数据块, Reduce 函数则对分块数据进行处理, MapReduce 框架自动对数据分块、调度并执行, 其执行流程如图 1 所示. Google 通过 GFS 和 MapReduce 每天能处理的数据高达 20PB[21]. MapReduce 在数据分析、日志分析以及商业智能分析等方面有很好的应用效果.

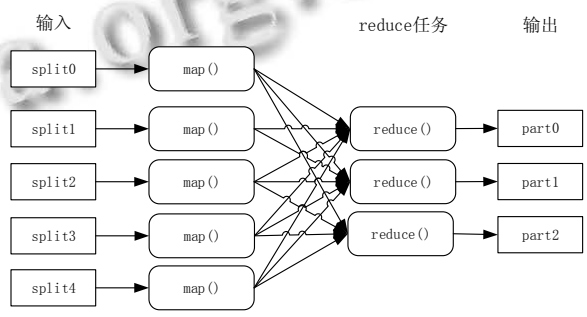

图 1 MapReduce 并行执行流程

4 高速公路大数据的机遇与挑战

4.1 大数据的存储

Hadoop 的 HDFS 系统虽然可以用来存储高速公路中的大数据, 但是难以满足其对实时性的要求. 因此需要对高速公路中产生的大数据进行分类, 采用不同的方式进行存储. 实时性要求较高的数据存储到实

时数据库中, 实时系统处理后的数据以及其他对实时性要求不高的数据或各业务系统产生的数据采用并行数据仓库存储, 大量的历史数据和非结构化数据存储到 HDFS 系统中. 高速公路大数据存储架构如图 2 所示.

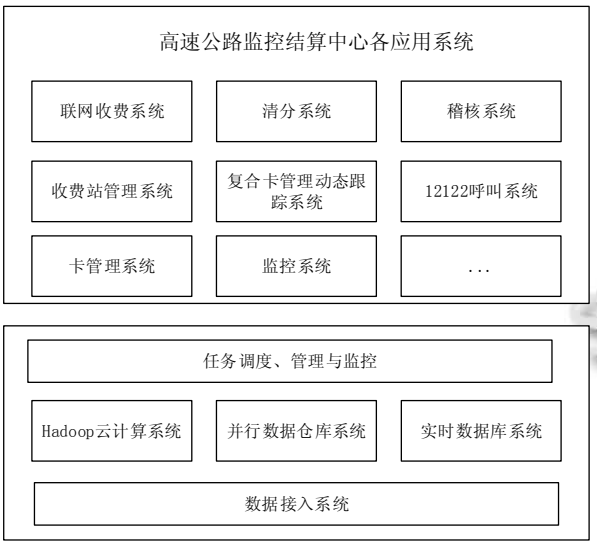

图 2 高速公路大数据存储架构

另外, 高速公路网中的大数据与传统电子商务系统中的大数据有着很大的不同, 具有数据生成速度快等特点, 且要求分析响应及时. 因此需要研究面向高速公路网大数据存储的格式, 从而有利于后续的数据分析和, 是一项艰巨的任务.

4.2 大数据的展现

高速公路中的大数据经过分析后, 将会有许多信息呈现给用户, 这是数据处理的最后一个环节. 如何将信息直观、有效的进行呈现是一项非常有挑战性的工作[22]. 没有好的人机交互界面就不会有好的用户体验, 系统的实用性将大打折扣. 因此科学的设计人界交互界面是高速公路大数据处理重要的技术.

4.3 大数据在高速公路中的应用

大数据处理技术在高速公路中的应用还处于起步阶段. 文献[23]分析了大数据背景下高速公路收费面临的问题, 讨论了未来高速公路收费系统的发展趋势. 文献[24]分析了现有高速公路网数据存储模式在抗风险、资源有效利用、大数据处理方面均面临着问题, 讨论了将云计算应用于高速公路中, 可以提高数据可靠性和存储能力, 增加数据安全, 降低成本方面, 但仅限于理论讨论, 尚未实现. 文献[25]介绍了浙江省“智慧高速”的规划, 依托大数据处理技术, 将来在交通预案、交通状态分析、交通事故分析和决策服务等方面将发挥重要的作用.

5 小结

高速公路网已经产生了大量的数据, 如何科学存储和分析这些大数据是必须要解决的问题. 本文提出了云平台、并行数据仓库以及实时数据库共同来存储高速公路网产生的数据, 成本低、扩展性高、数据处理速度快, 满足了数据处理的实时性又能存储海量的数据. 但大数据在高速公路中的运用还处于初级阶段, 高速公路收费结算中心和路公司对大数据运用都有着宏伟的蓝图, 目前大都停留在分析阶段, 并没有进入实际的运用. 在如何提高数据的存储、提高数据的可靠性以及有效分析利用数据方面还存在着较多的问题, 需要大家去探索、解决.

参考文献

1 杨进欣.我国加速编织公路网通车总里程将突破 410 万公里.新华社,2012-12-29.

2 交通部规划研究院.国家高速公路网规划. http://www.moc.gov.cn/2006/06tongjisj/06jiaotonggh/guojiagh/guojiajt/200608/t20060815_46064.html.[2005-01-13].

3 中华人民共和国交通部.公路水路交通运输信息化“十二五”发展规划.http://www.gov.cn/gongbao/content/2011/content_1992578.htm.[2011-04-27].

4 覃雄派,王会举,杜小勇,王珊.大数据分析—RDBMS 与 MapReduce 的竞争与共生.软件学报,2011,(9):32–45.

5 宋亚奇,周国亮,朱永利.智能电网大数据处理技术现状与挑战.电网技术,2013,(4):927–935.

6 Grobelnik M. Big-data computing:Creating revolutionary breakthroughs in commerce science and society. http://videolectures.net/eswc2012_grobelnik_big_data/.[2012-10-02].

7 孟小峰,慈祥.大数据管理:概念、技术与挑战.计算机研究与发展,2013,(1):146–169.

8 窦万春,江澄.大数据应用的技术体系及潜在问题.中兴通讯技术,2013(4).

9 王珊,王会举,覃雄派,周烜.架构大数据:挑战、现状与展望.计算机学报,2011,(10):1741–1752.

10 Inmon WH. Building the Data Warehouse(3rd). New York: John Wiley and Sons Inc, 2002.

11 Dealing with data. Science, 2011, 311(6018): 639–806.

12 天极网.沃尔玛等传统企业大数据投资开始产生回报. http://cio.yesky.com/317/34677317.shtml. [2013-04-27].

13 王元卓,靳小龙,程学旗.网络大数据:挑战、现状与展望. 计算机学报,2013,(6):1125–1138.

14 新华网.“双十一”电商丰收—阿里单日销售额超 350.19 亿元. http://news.xinhuanet.com/fortune/2013-11/12/c_118096396.htm.[2013-11-12]

15 网易科技.京东发布 JDPhone 计划通过数据挖掘定制手机 .http://tech.163.com/13/1119/17/9E2F86BF000915BF.html.[2013-11-19].

16 中国云计算网.什么是云计算? http: //www.cloudcomputing-china.cn/Article/ShowArticle.asp?ArticleID=1.[2009-02-27].

17 Ghemawat S, Gobioffh LPT. The Google file system. Proc. of the 19th ACM Symposium on Operating Systems Principles. New York. ACM Press. 2003. 29–43.

18 ApacheHadoop. Hadoop. http://hadoop.apache.org/. [2009-03-06].

19 陈全,邓倩妮.云计算及其关键技术.计算机应用,2009,(9): 2562–2567.

20 Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters. Proc. of the 6th Symposium on Operating System Design and Implementation (OSDI04). San Francisco, California, USA. 2004. 137–150.

21 Dean J, Ghemawat S. MapReduce: Simplified data processing on large clusters. In: Brewer E, Chen P, eds. Proc. of the OSDI. California: USENIX Association. 2004. 137–150.

22 Wong PC, Shen HW, Chen C, et al. Top ten interaction challenges in extreme-scale visual nalytics. Computer Graphics and Applications, 2012, 32(4): 63–67.

23 杜玉辉,蒋姣丽.大数据背景:高速公路收费系统数据的现状、分析与展望.电脑知识与技术,2012,(5):3752–3754.

24 高祥.高速公路新型数据处理结算中心云计算模式探讨. 中国交通信息化,2011,(4):113–117.

25 杨志杰.高速公路再掘金.IT 经理世界.2013(17).