基于 Hadoop 的城市道路交通流量数据分布式存储与挖掘分析研究*

廖飞[1]，黄晟[1]，龚德俊[2]，安乐[2]
（1.湖南省交通科学研究院 交通运输工程信息化中心，湖南 长沙 410076；
2.长沙赛视交通科技有限公司，湖南 长沙 410076）

摘要：面对巨大而且快速增长的城市道路交通流量数据，采用基于 Hadoop 中 HBase 分布式数据库来存储采集到的城市道路路段交通流量，并采用 Hadoop 高效并行计算 MapReduce 编程模型对海量的城市交通流量数据进行了数据挖掘分析。实验结果验证了基于 Hadoop 在城市道路交通流量数据的存储与处理上比传统的方式更高效。

关键词：城市交通；交通流；Hadoop；HBase；数据处理；并行计算；挖掘分析

中图分类号：U491.1 **文献标志码**：A **文章编号**：1671－2668(2013)05－0082－05

随着人们生活水平的提高，城市汽车拥有量快速增加，城市道路交通流量急剧增长，交通堵塞、交通事故、交通违法等问题日渐显著。面对如此巨大而且快速增长的城市道路交通流量数据，如何快速存储、灵活扩展存储容量，采用何种科学、有效的数据挖掘手段对交通流量数据进行分析研究，成为城市交通领域的一大难题。

该文采用 Hadoop 中 HBase 分布式数据库存储每一时段采集到的城市道路路段交通流量，采用 Hadoop 高效并行计算 MapReduce 编程模型对海量的城市交通流量数据进行挖掘分析，得到每一时刻、每天、每月的城市各路段交通流量统计分布情

****

很好地利用交通分析软件 TransCAD 对目标年的交通量进行预测，不仅可提高交通量预测的准确性，而且能使预测更为方便、快捷。

参考文献：

[1] 马俊来，王炜，李文权，等. OD 矩阵推算技术在高速公路影响区交通需求预测中的应用[J]. 公路交通科技，2004，21(6).
[2] 王炜，过秀成. 交通工程学[M]. 第 2 版. 南京：东南大学出版社，2011.
[3] 陆化普，史其信，殷亚峰. 交通影响评价的基本思想与方法[J]. 城市规划，1996(4).
[4] 仇亮. 城市交通影响评价的理论与方法[J]. 科技情报开发与经济，2003，13(11).
[5] 谭山. 大型购物中心交通影响分析方法研究[D]. 长春：吉林大学，2007.
[6] 马玉红. 物流园区交通影响分析方法研究[D]. 长春：吉林大学，2008.
[7] 韩飞，陈昆，熊丹. 住宅小区建设项目交通影响范围研究[J]. 公路与汽运，2012(5).
[8] 赵凯. 河北省城市建设项目交通影响评价研究[D]. 天津：天津大学，2005.
[9] 薛金刚. 城市土地开发交通影响评价方法研究[D]. 北京：北京工业大学，2005.
[10] 徐维红，曾学贵. 城市建设项目交通影响评价的研究[J]. 中国人民公安大学学报：自然科学版，2004(1).
[11] 杨鑫. 建设项目交通影响评价应用研究[D]. 西安：长安大学，2012.
[12] 沈建武，刘学军，陈良琛. 城市大型商业设施交通影响分析[J]. 武汉大学学报：信息科学版，2002，27(4).
[13] 杨健. 大型超市交通影响研究[D]. 成都：西南交通大学，2009.
[14] 宋微. 交通影响范围界定理论与方法研究[D]. 大连：大连交通大学，2008.
[15] 郭春侠. 山地城市住宅建设项目交通影响分析[D]. 重庆：重庆交通大学，2011.

收稿日期：2013－03－07

*基金项目：湖南省交通科技项目(201144)；湖南省自然科学基金项目(12JJ2025)

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

况，为城市路网规划、城市交通管理与控制提供依据，为交通研究、交通规划设计、交通管理部门提供决策辅助支持，同时对缓解城市交通问题起到一定的促进作用。

1 Hadoop 简介

Hadoop 是 Apache 软件基金会开发的开源分布式计算框架，在众多大型企业都有成功的应用案例。Hadoop 实现了 MapReduce 并行编程模型，提供了分布式文件系统 HDFS(Hadoop Distributed File System)，为分布式计算提供底层存储支持。Hadoop 在可扩展性、高效性、易用性、经济行、可靠性等方面具有很强的优势。

如图 1 所示，Hadoop 主要包括 Hadoop Common、Avro、MapReduce、HDFS、ZooKeeper、Pig、Chukwa、Hive、HBase 等子项目。

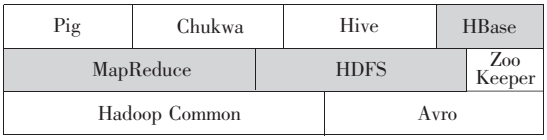

图 1 Hadoop 相关项目成员

2 城市道路交通流量数据分布式存储与挖掘分析的总体架构

根据交通流量数据自身实际特点，基于 Hadoop 设计城市道路交通流量数据分布式存储与挖掘分析的总体架构。如图 2 所示，该总体架构自下而上主要由数据采集层、数据存储层、挖掘分析层、应用服务层组成，数据存储层与挖掘分析层是该文的研究重点。

2.1 数据采集层

交通流数据主要通过部署在各个路段的感应线

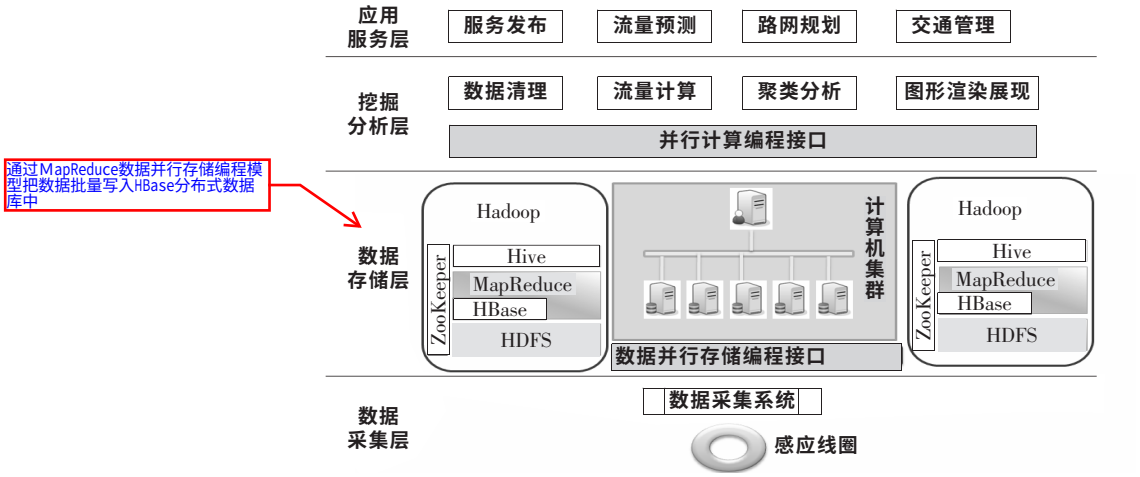

图 2 城市道路交通流量数据分布式存储与挖掘分析的总体架构

圈来采集，并把获得的数据传输给数据采集系统，然后把数据汇总，定期通过并行存储编程接口写入分布式数据库中。采集的数据包含车牌号、车辆型号、行驶方向、驶入时间等信息。

2.2 数据存储层

数据存储层主要通过 MapReduce 数据并行存储编程模型把数据批量写入 HBase 分布式数据库中，而写入的数据主要存储在 Hadoop 计算机集群中。计算机集群采用经典的主/从部署架构，也称为 Master/Slave 部署架构，其中 Master 为管理节点(NameNode)，存在一个，Slave 由存在 N 个的数据节点(DataNode)组成，并由 ZooKeeper 控制保存协调工作。NameNode 是整个集群的核心，记录每一个存储文件的切割情况，Block 存储位置等重要信息的元数据结构(MetaData)也由它进行管理和维护，DataNode 才是数据 Block 真正的物理存储位置。

HBase 通过多个进程的 Map 执行函数并行地把批量数据写入分布式文件系统中，实现海量数据的高效存储。同时，分布式数据库存储的每一个数据 Block 都采用冗余多备份的存储机制，能有效地处理单点故障，并且采用 Hadoop 内置的并行高效的 MapReduce 计算编程模型，能有效地提高存储速度，真正意义上实现海量数据的高效存储。

2.3 挖掘分析层

在挖掘分析层，根据总体架构中数据挖掘分析的需求构建数据清理、流量计算、聚类分析和图形渲

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

染展现 4 个主要模块。由于受到车辆速度、自身状况、周围环境等因素的影响，真实的交通状况并不总能由道路线圈采集的数据很好地反映，必须对采集的交通流数据进行预处理。数据清理模块在模型运算前对采集的交通流数据进行清洗，剔除不合理及异常的数据，以剔除空值数据和重复数据作为清洗数据的主要策略。

流量计算模块根据采集的流量历史数据及流量换算系数，给定一段需要流量计算统计查询的时间段，判断记录中的车辆驶入时间是否在时间段范围内，累加在时间段范围内并经过换算后的交通流量值。再根据路段的相关参数，计算出交通流量密度。这些数据统计结果可以作为后续交通量预测、交通管理与控制的参考依据。

聚类分析是进行数据挖掘的重要分析方法，旨在对数据进行分类。聚类分析根据路段的交通流量密度，以某一初始值为计算中心，按照计算算法流程多次迭代运算，寻找新的合适的中心值，将道路交通流量密度分成若干等级，为后续的图形渲染展现模块提供数据基础。

图形渲染展现模块主要是根据聚类分析生成的数据，在城市道路路网图或统计分析图表中进行渲染展现，最后生成科学、直观的数据统计结果图，给相关单位提供辅助决策支持。

2.4 应用服务层

应用服务层直接面向城市交通行业的广大潜在用户，以一战式服务的方式提供。根据云服务中应用即服务的理念，该层的设计把所有的资源和功能都以服务的形式提供给用户，能为交通预测、路网规划、交通管理提供数据分析服务，为交通管理和控制提供参考依据，还可以提供城市道路交通量统计数据的下载。云服务平台还提供城市道路交通流量相关资源和数据文档的交换框架体系，可以与其他系统进行数据交换和服务互操作。

3 HBase 流量数据模型设计

3.1 HBase 概述

Hadoop HBase 是基于 Google Bigtable 的开源实现，属于 Hadoop 的一个子项目。Hadoop HBase 可通过利用 Hadoop HDFS 提供的文件存储系统、Hadoop MapReduce 提供的海量数据处理能力和 Hadoop ZooKeeper 提供的协同服务构建一个高可靠性、高性能、面向列、可伸缩的分布式存储系统。这种分布式存储系统的特点是实现可在廉价节点上搭建起大规模结构化存储集群。

Hbase 是一个基于列模式的映射非关系型分布式数据库，它只能表示很简单的 key/value 的映射关系。相对于传统的 Oracle、SQL Server 等关系型数据库，其优点主要体现在存储模式、数据维护、可伸缩性等方面。

3.2 数据模型设计

城市交通流量数据巨大，且增长快速。Hbase 分布式数据库数据存储是基于列存储，为非关系型分布式数据库，具有一定的哈希性质，在处理大规模数据（TB 级）的存储和处理上比主流的关系型数据库具有独特的优势，故选择 HBase 作为分布式数据存储方案。

从城市道路交通流数据的自身特点、储存与处理出发，设计基于 Hbase 的交通流量数据模型。该模型主要包括路段信息、车辆流量信息。其中，路段信息数据模型以路段 ID（RoadID）作为 RowKey，Info 列族用于存储路线的基本信息，以 Info 列族所有信息的值作为 Value。Info 列族包括路段长度（Info：RoadLength）、路段宽度（Info：RoadWidth）、路段车道数（Info：RoadLanes）、路段名称（Info：Name）等，其设计如图 3 所示。车辆流量信息数据模型以路段 ID＋倒序的 TimeStamp 时间戳作为 RowKey，Data 列族用于存储车辆流量数据的基本信息，以 Data 列族所有信息的值作为 Value。Data 列族包括车牌号（Data：CarNumber）、车辆型号（Data：CarCategory）、车速（Data：CarSpeed）、驶入时间（Data：DriveInTime）、行驶方向（Data：DriveDirect）等，其设计如图 4 所示。

RoadID (RowKey)	Column Info
	Info：Name
	Info：RoadLength
	Info：RoadWidth
	Info：RoadLanes

图 3 路段信息数据模型（Roads）

上述所提到的 Key 和 Value 在数据库都以二进制的形式存储。

在 HBase 中，对于一个 Rowkey 来说，只需要指定相关的列簇名就可以获取相关业务的全部记

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

RoadID (RowKey)	TimeStap	Column Data
	T5	Data: CarNumber
	T4	Data:CarCategory
	T3	Data:CarSpeed
	T2	Data:DriveInTime
	T1	Data:DriveDirect

图 4 路段交通流量数据模型(TrafficFlow)

录。比如只要给定一个路段 ID 的值,就能获取该路段的所有 Info 列族的所有信息,同样也可以获取路段交通流量数据。

4 关键算法设计与实现

4.1 并行写入 HBase 中的 MapReduce 处理流程

由于道路交通流量数据是通过数据采集层传输过来的,是批量且大容量的。为了将数据高效地存储至 Hadoop 分布式存储集群中,需要开启多个 MapReduce 任务跟踪并行执行。

Map 阶段,解析传递过来的数据记录,提取路段基本信息及交通流量数据基本信息,把得到的 Key/Value 数据传递给 Reduce 阶段。其中 Key/Value 的格式是(路段 ID,[路段名称、路段长度、路段宽度、路段车道,车牌号、车辆类型、车辆速度、驶入时间、行驶方向])。

Reduce 阶段,把 Map 阶段得到的 Key/Value 数据按 Key 合并,并通过 HBase API 写入 HBase 的两个数据表中。

4.2 数据清洗中的 MapReduce 处理流程

数据清洗处理流程比较简单,MapReduce 算法设计也不复杂。

Map 阶段,解析 Hbase 查询出来的数据记录和数据格式并提取要验证的信息值,检查是否合理,如果异常,则遗弃该记录。将正常的数据发送至存储路径,进入 Reduce 阶段。

Reduce 阶段,将清洗过的数据按照路段编号进行归并,具有相同路段编号的记录被归并为一个文件,得到新的数据集。

4.3 流量密度计算中的 MapReduce 处理流程

根据设定的时间段范围,比如 1 h、1 d 或 1 个月,统计出各个路段的交通流量累计和。

Map 阶段,由于传递过来的记录是在数据清洗中根据路段 ID 合并后的,根据循环路段对应的交通流数据,判断车辆驶入时间是否在设定的时间范围内,如果是,则根据车辆换算系数,计数器加上换算后的车辆数。最后将数据传递给 Ruduce 阶段。Map 阶段输出的数据格式为(路段 ID,车道数、车道长度、车道宽度、车辆数)。

Reduce 阶段,根据 Map 阶段传递过来的输入数据及路段 ID 对车辆数进行累加计算,得到路段在时间段范围内的最终车辆数,根据流量计算模型,即可得到路段的流量密度。

4.4 聚类分析中的 MapReduce 处理流程

聚类分析过程较为繁琐、耗时,需要进行多次迭代运算,才能获得较为准确的分类结果。

Map 阶段,首先给定 N 个作为初始运算的基本参照的预定义初始的中心值,循环每个城市交通路段,分别计算路段流量密度与这 N 个中心值的距离,取其中距离最小的对应类别 M 作为它的同类。Map 阶段输出的数据格式为(类别,[路段 ID,流量密度])。

Reduce 阶段,把 Map 阶段计算归类得到的所有同类类别中的值进行重新计算,获得其新的中心值。按照 Map 阶段的处理流程,进一步地迭代计算,得到更精确的分类。

按照上述步骤,如此迭代循环计算,直到所有类别中的中心值不再发生变化为止。该处理流程最后得到的数据结果可以为随后的图形渲染展现提供数据支持。

5 实验

5.1 模拟环境

利用 VMWare 9.0 虚拟机建立计算机集群,计算机系统为 Ubuntu12.10(Linux 内核)。集群中共有 4 台计算机,其中 1 台作为 NameNode,另外 3 台作为 DataNode。所有计算机上都装备有 JDK1.6、Hadoop、HBase 和 ZooKeeper 等开发包。采用 Eclipse 作为 Java 开发环境,ArcGIS 作为 GIS 可视化展现平台。

5.2 流量计算模型及相关数据的准备

流量计算采用的计算模型为:

$$P_i=\frac{Q_i}{N_iL_iW_i}$$

式中:P_i 表示 i 路段的交通流量密度(辆/m^2);Q_i 表示 i 路段的交通流量总和(辆);N_i 表示 i 路段的车道数;L_i 表示 i 路段的车道长度(m);W_i 表示 i

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

路段的车道宽度(m)。

根据 CJJ37－2012《城市道路工程设计规范》,交通量换算采用小客车为标准车型,车辆换算系数如表 1 所示。

表 1 车辆换算系数

车辆类型	换算系数	车辆类型	换算系数
小客车	1.0	大型货车	2.0
大型客车	1.5	特大型货车	3.0
小型货车	1.0	拖挂车	3.0
中型货车	1.5	集装箱车	3.0

利用 ArcGIS for Desktop(高级)绘制模拟长沙市主要道路网,并设置相关的路段基本信息。采用数据采集小工具 TrafficDataCollector 24 h 不间断、随机性地向路段填充交通流量数据,持续 15 d 后,数据容量已达 10 G。

5.3 结果

基于 Hadoop 及算法模型,计算出 8:00—10:00 范围内路段交通流量情况,结合 GIS 可视化技术,得出的结果如图 5 所示。

图 5 长沙市主要路段 8:00—10:00 交通流统计信息

6 结语

从实验结果上看,针对海量的城市道路交通流量数据,利用 Hadoop 技术进行数据存储和处理是合理、可行、高效的。利用 Hadoop 的分布式文件系统灵活扩展的特性,可解决交通流量数据的快速增长存储问题;利用 Hadoop 的并行快速计算模型,在海量交通流量数据分析处理方面可发挥极大的优势。快速、准确的城市道路交通流量数据的深入挖掘分析,可为交通研究、交通规划设计、交通管理部门提供决策辅助支持,对缓解城市交通问题具有一定的促进作用,具有很强的应用前景。

参考文献:

[1] Tom White. Hadoop:the definitive guide[M]. O′Reilly Medial Yahoo Press,2009.
[2] Jason Venner. Pro Hadoop[M]. Apress,2009.
[3] Chunk Lam. Hadoop in action[M]. Manning Publications Co,2010.
[4] 马国旗. 城市道路交通流量特征参数研究[D]. 北京:北京工业大学,2004.
[5] 朱钥,贾思,奇张俊,等. 基于 Hadoop 的城市交通碳排放数据挖掘研究[J]. 计算机应用研究,2011,28(11).
[6] 王敬昌. 基于 Hadoop 分布式计算架构的海量数据分析[J]. 数字技术与应用,2010(7).
[7] 陈勇. 基于 Hadoop 平台的通信数据分布式查询算法的设计与实现[D]. 北京:北京交通大学,2009.
[8] 王振宇,郭力. 基于 Hadoop 的搜索引擎用户行为分析[J]. 计算机工程及科学,2011,33(4).
[9] 孙福权,张达伟,程勖,等. 基于 Hadoop 企业私有云存储平台的构建[J]. 辽宁工程技术大学学报,2011,30(6).
[10] 杨锋,吴华璃,朱华吉,等. 基于 Hadoop 的海量农业数据资源管理平台[J]. 计算机工程,2011,37(12).
[11] 谢桂兰,罗省贤. 基于 Hadoop MapReduce 模型的应用研究[J]. 软件天地,2009,8(2).
[12] 周红伟,李琦. 基于云计算的空间信息服务系统研究[J]. 计算机应用研究,2011,28(7).
[13] 李勇君. 基于 Hadoop 的海量期货数据的分布式存储和算法分析[D]. 天津:天津大学,2012.
[14] 朱珠. 基于 Hadoop 的海量数据处理模型研究和应用[D]. 北京:北京邮电大学,2008.
[15] 丁辉,张大华,罗志明. 基于 Hadoop 的海量数据处理平台[A]. 2011 电力通信管理暨智能电网通信技术论坛论文集[C]. 2011.
[16] 万至臻. 基于 MapReduce 模型的并行计算平台的设计与实现[D]. 杭州:浙江大学,2008.
[17] CJJ37－2012,城市道路工程设计规范[S].
[18] 谢桂兰,罗省贤. 基于 Hadoop MapReduce 模型的应用研究[J]. 微型机与应用,2010(8).
[19] 封俊. 基于 Hadoop 的分布式搜索引擎研究与实现[D]. 太原:太原理工大学,2010.
[20] 邓自立. 云计算中的网络拓扑设计和 Hadoop 平台研究[D]. 合肥:中国科学技术大学,2009.

收稿日期:2013－03－13

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net