基于 Hive 的智慧城市数据处理技术研究与实现

艾丽蓉,刘云峰
(西北工业大学 计算机学院,陕西 西安 710129)

摘 要: 对智慧城市系统中产生的大量数据进行有效的采集、合理的存储、高效精准的分析,进而对决策的做出提供合理的支持是在智慧城市建设过程中必须要解决的问题。对此,在充分理解智慧园区数据分析系统功能需求、性能需求的基础上,结合当前比较主流的数据分析方面的技术,提出了 Kettle+Hive+Tableau 的智慧园区数据分析解决方案。该系统具有较高的可用性、稳定性、效率以及非常高的扩展性、可移植性,不仅适合于对智慧园区的数据分析,还普遍适用于对智能化信息系统数据进行分析,具有较高的普适性、通用性。该系统通过 ETL 工具、数据可视化工具的应用,尽量减少在数据分析系统开发过程中代码的编写,能够适用于绝大部分有数据分析需求并且数据量较大的信息化系统。

关键词: 智慧城市; Hive; 数据采集; 数据可视化分析
中图分类号: TP311 **文献标识码:** A **文章编号:** 1673-629X(2018)02-0009-05
doi: 10.3969/j.issn.1673-629X.2018.02.003

Research and Implementation of Data Processing Technology of Intelligent City Based on Hive

AI Li-rong, LIU Yun-feng
(School of Computer Science, Northwest Polytechnical University, Xi'an 710129, China)

Abstract: It must be solvable for effective acquisition, reasonable storage, efficient and accurate analysis of the large amounts of data from intelligent city system, and then providing the reasonable support for decision-making in the construction of intelligent city. For this, combined with currently mainstream data analysis technologies, we propose a data analysis solution of Kettle+Hive+Tableau for intelligent park on the basis of fully understanding the functional requirements and performance requirements of the intelligent park data analysis system. The system is of better availability, stability and efficiency as well as excellent scalability and portability, which is not only suitable for the data analysis of wisdom park, but also widely applied to the analysis of data from intelligent information system. In addition, it has high universality. Through the application of ETL tools and data visualization tools, it minimizes the writing code in the development of data analysis system, which can be applied to most of the information system with data analysis needs and lots of data.

Key words: smart city; Hive; data acquisition; data visualization analysis

0 引 言

智慧城市(smart city)是把新一代信息技术充分运用到城市运行和管理的各行业,用以分析、整合城市运行核心系统的各项关键信息,从而对各行业的多种需求做出智能响应[1-3]。其实质是运用先进的信息技术,实现信息化、工业化与城镇化的深度融合,对于提高城市生活质量有显著作用[4]。

智慧城市在运行过程中产生的大量数据是智慧城市宝贵的资源,通过对这些数据的收集、汇总、分析,能够体现出城市各部门、系统的运行情况,帮助管理者做出最符合城市发展的决策。大数据与智慧城市的关系可表述为:物联网技术的运用推动大数据的发展,大数据的发展又成为智慧城市的发展的基石,智慧城市的衡量指标由大数据来体现。

1 相关技术概述

1.1 Hadoop 平台

为了应对智慧园区项目运行过程中所产生的海量数据的存储要求,以及对数据分析效率的要求,充分利用智慧园区中多核主机、大容量存储等硬件资源,特别

收稿日期: 2017-01-19 **修回日期:** 2017-05-24 **网络出版时间:** 2017-11-15
基金项目: 国家自然科学基金(61502371)
作者简介: 艾丽蓉(1970-),女,博士,副教授,CCF 高级会员(06445S),研究方向为大数据技术、智能信息处理; 刘云峰(1990-),男,硕士,研究方向为数字媒体技术。
网络出版地址: http://kns.cnki.net/kcms/detail/61.1450.TP.20171115.1425.022.html

引入Hadoop平台中HDFS(分布式文件系统)、MapReduce(并行计算框架)和Hive(数据仓库工具)。三个组件作为Hadoop的核心能够使用户轻松地架构和使用分布式计算平台,并在该平台的基础上对大规模数据进行处理与分析。图1显示了Hadoop生态圈中的各主要技术。

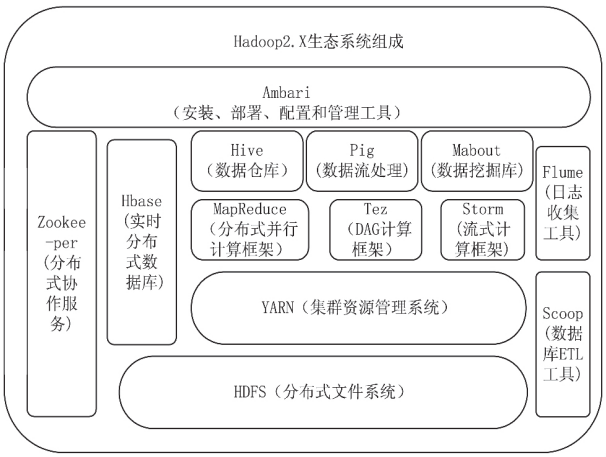

图1 Hadoop生态圈

1.1.1 HDFS分布式存储

HDFS(Hadoop distributed file system)是Hadoop项目的核心子项目,是Hadoop生态系统中所有组件的基础,具有容错性高、可靠性高、可扩展性高、吞吐率高等特性[4]。HDFS在系统架构上采用master-slave结构,可以用廉价的硬件实现大规模数据的可靠性并可实现对数据的高吞吐量的访问,非常适合于需要对大量数据进行存储与处理的应用场景。

1.1.2 数据仓库Hive

Hive是一个以Hadoop为基础的,建立在Hadoop生态系统之上的数据仓库,能够对大量的结构化数据进行存储与处理[5]。它将大量的数据存储在HDFS中,采用类SQL语言HQL对数据进行操作与管理。Hive中的解释器负责对HQL进行解析和转换,将其解释为map/reduce任务,并通过执行map/reduce任务返回对Hive中数据的查询或处理的结果。

1.1.3 MapReduce分布式计算

作为Hadoop的核心组件之一,MapReduce是一种分布式计算框架[6]。该框架由编程模型和运行时环境两部分组成,其中编程模型由谷歌于2004年发表的分布式计算框架MapReduce的论文提出,为用户提供了非常易用的编程接口,用户只需像编写串行程序一样实现几个简单的函数就可以实现分布式程序。通过运行时环境来完成如节点间通讯,数据切分和节点失效等复杂工作,用户无需关注框架的运行细节,可以轻易地完成大规模数据的处理任务。

1.2 ETL

ETL过程的主要作用为从各个数据源(如业务系统数据库、文本文件等)将数据抽取到中间层,之后根据制定的数据清洗规则对抽取到的数据进行清洗、转换,最后加载到数据仓库中,为进行数据分析打下基础。

ETL过程是构建数据仓库过程中非常关键的一部分,起到了承前启后的作用[7-9]。智慧园区各个业务系统的数据均存放在自己的业务数据库中,其中存放的数据是面向业务的,数据粒度较细,存储的信息较为详细,不适于直接对其中的数据进行分析并且各个业务数据库是相对独立的,直接进行分析需要面对多表联结、数据格式不一致等相关问题,给分析工作增加难度。

为了解决上述问题,引入ETL与数据仓库。通过ETL过程提取不同数据库中的数据,按照数据分析需求制定数据清洗策略,完成对数据的清洗、转换之后将数据加载到数据仓库。ETL过程的工作直接关系数据仓库中数据的质量,同时也关系到数据分析的质量与结果。

1.3 数据可视分析

Google首席经济学家Hal Varian教授指出"数据正在变得无处不在、触手可及;而数据创造的真正价值,在于我们能否提供进一步的稀缺的附加服务;这种增值服务就是数据分析[10]。"数据是信息化系统最宝贵的财富,在数据中蕴含着大量可为企业进行决策提供支持的关键信息。而蕴藏在数据中的信息只有采用相关的数据分析技术进行深入挖掘才能得到,仅仅凭借经验与直觉并不能充分利用数据中的信息。Thomas和Cook在文献[11]中对可视化的定义是:可视分析是一种通过交互式可视化界面,来辅助用户对大规模复杂数据集进行分析推理的科学与技术。可视分析的运行过程可看作"数据→知识→数据"的循环过程,中间经过两条主线:可视化技术和自动化分析模型。从数据中洞悉知识的过程主要依赖上述两条主线的互动与协作[12-13]。

随着信息化系统产生的数据量的增大,传统的数据分析技术已经不能满足对大规模数据集进行有效分析的需求。而如何对数据量大但价值密度较低的大数据进行有效分析是每一个现代化信息系统必须面对的问题。人类获得的绝大部分信息来源于视觉,将现有的大数据平台与数据可视化分析技术结合起来,借助于大数据平台具有的数据处理能力,将数据以更直观的形式(图片等)展示出来,能够帮助决策人员更好地理解数据中所蕴含的信息。

因此,大数据可视化是对大数据进行分析的最有效、最重要的环节,数据可视化技术在大数据分析中扮演着非常重要的角色。

2　智慧园区数据分析系统设计

2.1　系统架构设计

如图 2 所示，底层采用虚拟化技术，虚拟化实现了 IT 资源的逻辑抽象和统一表示，在大规模数据中心管理和解决方案交付方面发挥着巨大作用，是支撑云计算最重要的技术基石。ETL 过程则采用 ETL 工具从智慧园区各个业务系统的业务数据库抽取数据，经过清洗、转换之后加载到大数据系统的 Hive 数据仓库中，之后利用可视化数据分析工具对加载到 Hive 中的数据进行分析。

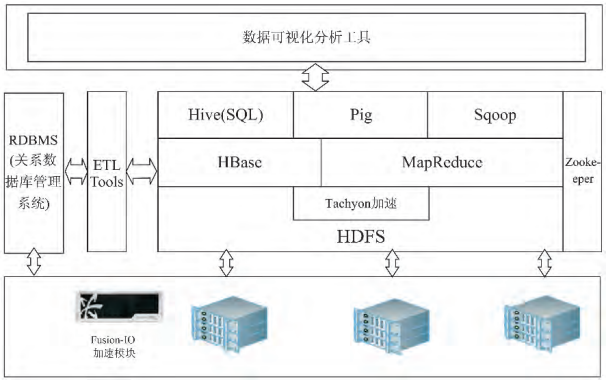

图 2　智慧园区数据分析系统架构

在硬件方面，针对数据分析系统对高可用性、可伸缩性、高吞吐量、高效性以及部分应用的低时延的需求，在网络交换传输使用层使用 Infiniband 交换机; 在数据传输协议层，将原有的 Hadoop 中的 socket 传输机制替换成 RDMA(远程直接内存存取) ，从而极大地缩短了传输时延，并充分发挥 Infiniband 交换机的数据传输能力，同时减少了 CPU 在节点间数据交换所耗费的资源。

2.2　系统模块划分

根据智慧园区数据分析业务的功能需求，在原有大数据平台的基础上，充分满足业务可伸缩性的要求，建设数据分析系统。系统主要分为三个功能模块: 数据采集模块、数据存储模块、数据可视化分析模块。数据清洗模块的主要功能为数据抽取、数据转换、数据加载。数据存储模块的主要功能为将数据采集模块处理过的数据存储到 Hive 数据仓库中。数据可视化分析模块的主要功能为数据源链接、数据整合、数据可视化展示。具体的模块功能划分和模块关系如图 3 所示。

3　数据采集系统设计

3.1　系统功能概述

为了实现和现有数据库的无缝连接，建立不重不漏、互通互联的数据分析系统，开发数据采集子系统，主要用于连接各种已有业务数据库(如: Oracle、SQL Server 和 MySQL) ，并且可以对变化更新的数据进行捕捉然后提取数据。具体功能主要包括:

(1) 数据导入: 首次运行时，与现有的数据库进行系统对接，将各数据库中的数据提取后并进行清理与治理，最后存入大数据库中;

(2) 自动抓取: 系统正常运行后，在不影响现有系统正常运行的基础上，对各库进行实时监控，并对新数据进行增量抓取，最后存入大数据库中;

(3) 数据录入: 对于部分动态数据库中的数据库，开放接口，方便用户通过其他终端录入数据;

(4) 数据清洗: 对获得的数据，根据业务以及用户的规定进行清洗、去重与治理。

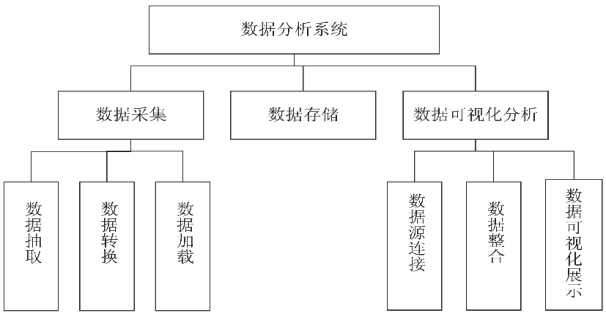

图 3　数据分析系统功能模块

数据采集系统的数据主要来自于正在运行的业务系统数据库(如: 惠民系统数据库、城管系统数据库、环卫系统数据库等) ，数据采集系统与这些业务数据库建立连接，从中抽取需要的数据，经过转换、数据清洗等操作，最后加载到 Hadoop 大数据平台的 Hive 数据仓库中。数据的抽取、转换、加载的过程称之为 ETL 过程。

3.2　系统流程

数据采集首先经过数据抽取，数据抽取的数据来源为业务系统数据库、文件系统。抽取的数据按照制定的清洗规则进行清洗后到达数据转换模块。数据转换模块按照转换规则对数据进行转换后，加载到数据仓库或者输出到文件。

3.2.1　数据抽取

数据抽取是进行数据清洗、数据转换的前提。数据抽取是一项艰难的工作，因为数据是多样和复杂的。这一部分需要在调研阶段做大量的工作，首先要搞清楚数据是从几个业务系统中来，各个业务系统的数据库服务器运行的是何种 DBMS，是否存在手工数据，手工数据存量有多大，是否存在非结构化的数据等等。等收集完这些信息之后再进行数据抽取设计。抽取的方式分为全量抽取和增量抽取。

该系统是智慧城市的一个模块，智慧城市还包括各个业务系统(城管系统、执法系统、环卫系统、惠民系统等) 。在数据采集系统第一次运行时，需要将各个业务数据库中的数据全部抽取出来，经过后续的数

据清洗、数据转换后加载到数据仓库中。由于系统是持续性运行的，会不断有新的数据进入到业务系统数据库，此时再对数据进行全量抽取是不现实的，不仅会加重整个系统的负担，更可能造成网络的拥塞，使整个系统的延迟增大[10]。此时便需要采用另外一种数据抽取方式—增量抽取。该系统中采取的增量抽取方式为基于时间戳的抽取方式。

数据抽取流程如图 4 所示。首先加载数据库驱动程序，连接到要抽取的数据源。判断是否连接成功，如果连接失败，写日志，记录失败的原因。如果连接成功，查询表中记录，进行数据的抽取工作，将抽取的数据放入数据缓存区，留待进行数据清洗转换。此外，由于后续要进行基于时间戳的 CDC(数据增量抽取工作)，系统需要维护 CDC 表来保存此次数据抽取的时间，获取当前系统时间，更新 CDC 表中 load 字段(上次抽取时间)。在进行增量抽取时，需要先读取 CDC 表中的 load 字段，之后只抽取业务库中更新时间大于 load 字段值的记录。

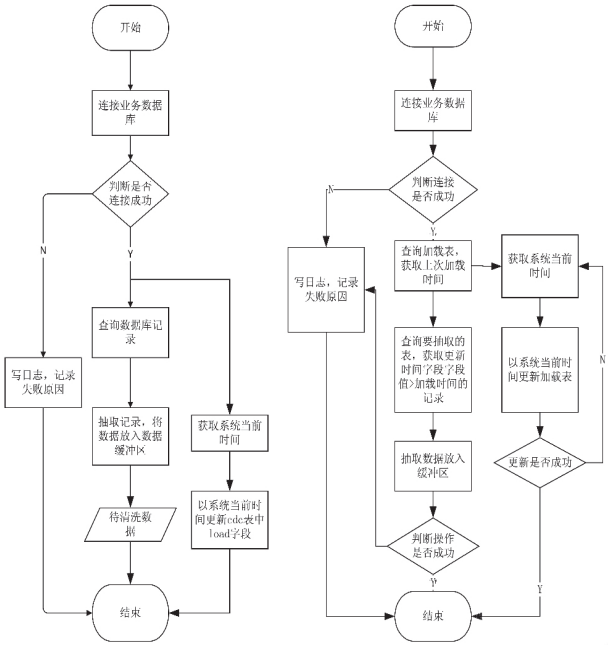

图 4 全量抽取、增量抽取流程

3.2.2 数据清洗转换

如图 5 所示，数据清洗转换是数据采集系统的核心。由于数据来源的多样性，业务系统不能完全保证存储在业务数据库中数据的真实有效性与准确性。而数据仓库是用来进行辅助决策的，要求存储在数据仓库中的数据都是正确且真实有效的，避免脏数据出现在数据仓库中。

数据清洗转换主要包括数据去重、不一致数据转换、数据粒度转换等过程。数据清洗转换是一个长期、反复的过程，不是短时间内能够完成的，是一项持续性的工作[11]。

3.2.3 数据加载(加载事实表)

事实表是用于分析的详细业务数据的集合。它的数据量大，会消耗大量的存储。图 6 展示了数据加载流程。由于数据抽取分为全量抽取和增量抽取，在加载事实表时也分为增量加载与全量加载。但其中对于数据的处理方式是相同的。此外，在进行数据加载时，必须先查询维度表中是否存在相对应的代理键，如果不存在，先进行维度表的更新，之后进行事实表的加载。

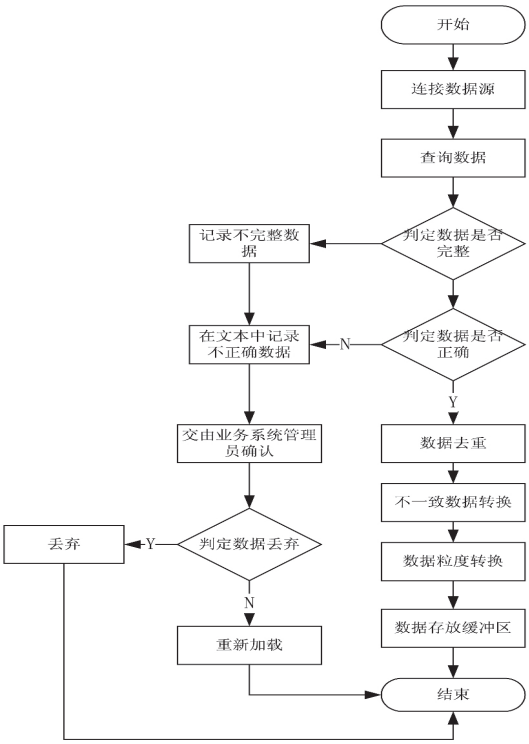

图 5 数据清洗转换流程

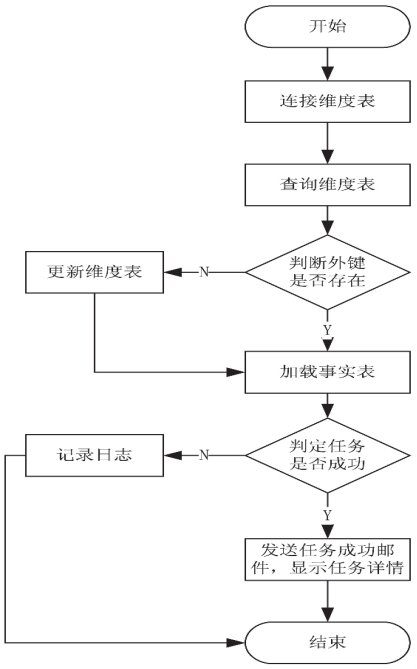

图 6 数据加载流程

4　数据可视化分析系统

数据分析系统的功能流程如图 7 所示。

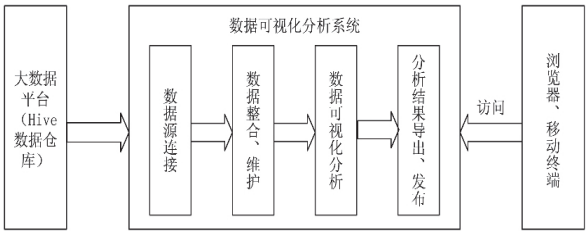

图 7　数据可视化系统业务流程

数据可视化分析系统具有以下功能:

(1) 数据源连接: 数据是进行数据分析的基础。第三节介绍了数据采集系统,其将数据从业务数据库中经过数据清洗转换之后加载到大数据系统中的 Hive 数据仓库中,为数据分析提供了数据源。因此,数据可视化分析系统应当具有连接到 Hive 并从中提取数据的功能。

(2) 数据整合、维护: 用来进行分析的数据可能来自于同一张表,也可能来自于多张表或者不同的服务器。因此数据可视化分析系统应当具有数据整合功能,用以实现同一数据源的多表联结、多个数据源的数据融合。同时由于分析时对数据有不同的需求、数据源中数据会发生变化,因此应当对数据进行筛选,限定数据的分析范围并且可以刷新数据源,保持数据更新。

(3) 数据可视化分析: 是数据可视化分析系统的核心功能。可以将分析结果以条形图、直方图、饼图、折线图等形式展示。并且具有统计分析、基于时间序列预测等功能。

(4) 分析结果导出、发布: 数据可视化分析系统可以将分析结果导出到文件,如 Excel、PDF 或图片。并将成果发布到服务器上,通过浏览器或者移动终端进行交互式访问。

5　结束语

现阶段城市的发展遇到了各种各样的挑战,如健康与环境、交通、水资源、能源利用、城市管理等问题,已经不能通过传统意义上的城市规划设计予以解决。通过运用新一代信息技术来管理城市的运行,将城市中不同的部门结合起来,共享不同部门间的信息。但是新一代信息技术的引入势必带来数据量的剧增,为信息系统的数据服务器带来巨大的压力,与此同时如何对海量数据进行有效的分析,为智慧城市决策人员的决策提供依据也是智慧城市在发展过程中需要解决的问题。为了解决海量数据的分析问题,对数据进行有效的分析,并保证数据分析的效率,文中在智慧园区信息系统的开发过程中引入 Hadoop 大数据平台、Hive 数据仓库等大数据技术,为海量数据的存储、分析提供支持。通过充分了解智慧园区数据分析系统的功能需求,并考虑到系统的稳定、易用等因素,决定采用 Kettle+Hive+Tableau 的方式来实现对智慧园区数据的分析。经过充分的技术验证与测试,证明了该方案能够解决智慧园区对数据分析的业务要求。

参考文献:

[1] 甄　峰,秦　萧.大数据在智慧城市研究与规划中的应用[J].国际城市规划,2014(6):44-50.
[2] 陈红松,韩　至,邓淑宁.智慧城市中大数据安全分析与研究[J].信息网络安全,2015(7):1-6.
[3] 李光亚,张敬谊,童　庆.大数据在智慧城市中的应用[J].微型电脑应用,2014,30(12):1-4.
[4] SHVACHKO K,KUANG H,RADIA S,et al.The Hadoop distributed file system[C]//26th symposium on mass storage systems and technologies.[s.l.]:IEEE,2010.
[5] THUSOO A,SARMA J S,JAIN N,et al.Hive: a warehousing solution over a map-reduce framework[J].Proceedings of the VLDB Endowment,2009,2(2):1626-1629.
[6] DEAN J,GHEMAWAT S.MapReduce[J].Communications of the ACM,2008,51(1):107.
[7] 宋旭东,刘晓冰.数据仓库 ETL 任务调度模型研究[J].控制与决策,2011,26(2):271-275.
[8] 张　宁,贾自艳,史忠植.数据仓库中 ETL 技术的研究[J].计算机工程与应用,2002,38(24):213-216.
[9] 徐俊刚,裴　莹.数据 ETL 研究综述[J].计算机科学,2011,38(4):15-20.
[10] 任　磊,杜　一,马　帅,等.大数据可视分析综述[J].软件学报,2014,25(9):1909-1936.
[11] THOMAS J J,COOK K A.Illuminating the path: the research and development agenda for visual analytics[M].[s.l.]:National Visualization and Analytics Ctr,2005.
[12] 陈　聪,张国惠,马晓磊,等.利用大数据挖掘和知识发现技术辅助智慧城市发展[J].大数据,2016(3):39-48.
[13] 官思发,孟　玺,李宗洁,等.大数据分析研究现状、问题与对策[J].情报杂志,2015,34(5):98-104.