

# 数据仓库技术与其在地质调查中的应用

杨晨雪, 文敏

中国地质调查局发展研究中心, 北京, 100037;

**关键词:** 数据仓库; hive; ETL; 地质调查; 业务管理

## 1 数据仓库研究背景

随着信息化、现代科技与商业智能技术的快速发展, 数据量正以指数的形式快速增长, 隐藏着巨大的商业价值。近年来, 随着传统数据库技术的广泛应用, 在同一机构中, 通常存在着多个运行在不同软硬件平台上的系统, 数据的存储彼此独立, 相互隔绝, 呈现出多元异构的特征, 难以达到数据共享的目的, 形成了一个个“数据孤岛”(陈跃国等, 2004)。为了将纷繁复杂的信息进行保存、使用和管理, 需对海量数据进行整合与集成, 将来源不同、结构不同的数据, 按照一定的逻辑进行处理和封装, 从而形成统一的数据平台, 使用户能够无障碍查询数据。在此背景下, 数据仓库应运而生。

## 2 数据仓库的特点及架构

1991 年, 比尔恩门 (Bill Inmon) 首次提出数据仓库的定义, 他指出, 数据仓库 (Data Warehouse) 是一个面向主题的 (Subject Oriented)、集成的 (Integrate)、相对稳定的 (Non-Volatile), 同时也是反映历史变化的 (Time Variant) 的数据集合, 构建数据仓库的目的是为公司管理层提供辅助决策 (Bill Inmon et al., 2002)。

数据仓库的架构主要包括数据存储层 (ODS)、数据模型层 (MDS)、数据应用层 (ADS) 等三个层次。数据存储层即 ODS 层, 目的是对源数据进行数据清洗和处理。源数据直接来自于各个业务系统, 必须按照逻辑进行加工处理, 即通过 ETL (抽取、转换、装载) 过程进入到 ODS 层中。ETL 技

术是获取数据的关键技术, 也是整个数仓系统的基础。数据模型层也是数据仓库的中间层, 通过加工, 将数据按主题进行划分, 并根据业务进行数据关联, 形成宽表。模型层数据为数据仓库的上层统计、分析、挖掘和应用提供直接的支持。ADS 层是数仓的应用层, 数据以业务线或部门划分, 基于模型层数据生成业务报表, 并能够对数据进行统计、分析、预测等一系列数据挖掘和分析工作, 从而为机构提供辅助决策。

## 3 数据仓库的建设和应用

基于上述数据仓库体系结构, 许多数据库厂商都推出了数据仓库产品, 比如 Oracle 公司的 Oracle Warehouse Builder, IBM 公司的 DB2 Data Warehouse Center, Microsoft 公司的 Data Transformation Services, Informatica 公司的 Powercenter 等, 然而随着数据规模的急剧增大, 现有的基于关系型数据库的数据仓库无法支撑海量数据下的存储和分析。目前国内较为常用的是一款基于 Hadoop 的开源数据仓库, 名为 Hive, 能够在分布式文件系统上建立数仓, 具有很高的可扩展性, 从而满足海量数据的存储、组织和管理。在我国许多行业, 例如运营商、银行、互联网、服务业等公司都建立了属于自己的数据仓库, 其中中国联通和中国移动两家运营商都基于数据仓库技术建立起“经营分析系统”, 是最具代表性的应用 (甘渊, 2018)。与此同时, 各大电力公司也纷纷研究建立了企业信息平台、电力调度中心统一数据平台、智能变电站数据中心等数据仓库应用平台 (王德文等, 2013)。此外, 在临床医疗领域, 医学数据仓库对大规模的实验和临床数据进行整合, 构建

注: 本文为国土资源部公益性行业科研专项 (201511079): 国家地质矿产调查评价专项地质调查工程 (0746) 项目的成果。

收稿日期: 2019-01-10; 改回日期: 2019-03-20; 责任编辑: 黄敏。Doi: 10.16509/j.georeview.2019.s1.145

作者简介: 杨晨雪, 女, 1993 年生, 硕士, 助理工程师, 地理信息系统专业, Email: cxyang910@163.com。

了生物医学平台,进行了数据挖掘,为发现疾病、基因、蛋白质间的关系网络提供了有效途径(孙慧媛等,2017)。

#### 4 数据仓库在地质调查业务管理中的应用

中国地质调查局现已建设了一套较完善的地质调查业务管理信息化系统,包括地质调查规划部署、项目管理、装备管理、安全生产管理、指挥调度、行政办公等管理领域,均部署了信息化系统,实现了管理在线化与数字化(谭永杰等,2018)。然而随着各个系统版本的更迭,跨业务的系统之间存在壁垒,使得业务管理数据表现出存量大但分散存储、数据价值高但利用率低的特点。因此建设地质调查业务管理数据仓库,不仅有利于机构掌握整体的运营情况,还能通过打通数据,分析和发现业务管理中存在的问题,从而促进业务的开展和改进。

地质调查业务管理数据仓库采用了一种混合的技术架构,结合传统的关系型数据库以及 Hadoop 集群组件构建而成,根据数据量大小和数据的属性分别在两个平台间进行数据传递。源数据需通过 ETL 工具抽取到数据存储层中,我们采用 Hadoop 集群中的 Scoop 组件,将非结构化数据集成至分布式文件系统 HDFS 中。由于 Scoop 并不适合数据量不大的结构化数据,我们采用 ETL 工具 Kettle 对结构化数据进行选取、清洗和转换,并存储于 ods 层。根据地质调查局具体业务,确定模型层主题域,包括项目质量管理主题、工作进度评估主题、人力资源评估主题等。接着对数据按主题建模,根据维度表和事实表的关系来确定模型,一般有星型模型、雪花模型。在应用层,引入了分布式查询分析引擎 Kylin 以及智能报表平台 Tableau,使用户能够进行自主数据查询,从而实现数据共享,减少业务人员

的管理负担,提升了管理效率。同时基于数据集成治理结果,针对业务管理需求,构建数据分析模型,开展预测与评价等相关研究,发挥数据价值,支撑地质调查管理与决策。

截止目前,地质调查业务管理数据仓库已集成各类结构化数据 2000 万条,非结构化数据近 20 万档,实现了地质大调查开始以来,项目监管、项目统计、经济管理、人事管理、安全生产等多年度、多来源、多类型数据的高效集成、实时采集同步与一体化有效组织治理。数据仓库作为数据汇集、共享、分析的一种有效、可行的技术解决方案,在政府部门存在较大的应用空间(赵贤,2018)。同时,数据仓库也可结合大数据分析方法与技术,如分布式计算、自然语言处理等技术,更好的辅助决策,推动政务数据的治理、共享与应用。

#### 参 考 文 献 / References

- 陈跃国,王京春. 2004. 数据集成综述. 计算机科学, 31(5): 48~51.
- 甘渊. 2018. 全国土地确权登记报送系统的架构优化设计与实现. 导师: 段晓峰. 北京邮电大学电子与通信工程系硕士学位论文: 1~69.
- 孙慧媛,孙瑞华,李友林. 2017. 数据仓库的建设与发展及医学领域的应用现状. 北京中医药, (8): 20~22.
- 谭永杰,屈红刚,文敏. 2018. 论地质调查工作大数据. 地理信息世界, 25(2): 7~11.
- 王德文,肖凯,肖磊. 2013. 基于 Hive 的电力设备状态信息数据仓库. 电力系统保护与控制, (9): 125~130.
- 赵贤. 2018. 基于数据仓库技术的政务数据汇集共享解决方案. 计算机产品与流通, (2): 11.
- Inmon, William H. 2002. Building the Data Warehouse.
- YANG Chenxue, Wen Min: Data warehouse technology and application in geological survey**
- Keywords: data warehouse; hive; ETL; geological survey; business management**