

# 面向大数据的高校图书馆数据集成架构<sup>\*</sup>

卓建霞 成兆珠 王丽华

(盐城工学院图书馆, 江苏 盐城 224051)

**[摘要]**从论述高校图书馆大数据的主要内容和数据特性入手,设计面向大数据的高校图书馆数据集成架构,即在数据仓库之上增加一个中间虚拟数据服务层,通过虚拟数据服务层集成数据仓库数据、实时数据及数据库变化数据,以充分发挥大数据的作用。

**[关键词]**大数据 高校图书馆 数据集成 数据仓库 数据虚拟化

**[分类号]**G250

大数据时代的来临已经毋庸置疑。大数据之“大”,并不仅仅在于数据量巨大,更多的意义在于人类可以分析和使用的数据在大量增加<sup>[1]</sup>。大数据的终极目标是对大量来自不同数据源的不同类型的数据进行分析,以识别出组织存在的风险和机会,并做出实时决策。高校图书馆作为文献信息的集散地,为教育教学服务,也为地方科技和经济发展服务,一方面使用庞大的资源为用户服务,同时在运作和服务过程中又产生了大量数据。随着近几年新兴社交媒体引入图书馆,在和用户的互动过程中又产生了大量的非结构化数据。然而,数据量大、种类繁多、分散存储几乎是所有组织的通病,由于数据结构、语义、格式转换上的较大差异,数据共享难以有效实现,高校图书馆要想迈入大数据时代,数据集成是大前提。

## 1 高校图书馆大数据的主要内容

数据是数据集成的核心,研究大数据集成,首先必须要明确哪些数据是需要获取的。大数据的来源多样,通常存储在数据库、文本文档、电子表格、电子邮件、网页文本中。归纳起来主要有:

①数据库数据。高校图书馆拥有丰富的文献资源,存储于文献管理系统及数据库中,数量巨大,持续更新,在用户服务过程中产生的读者借阅数据、数据库使用统计数据、主页访问数据等,多为结构化数据。

②用户交互数据。图书馆在文献信息服务过程中,更加重视读者的参与,如读者意见调查和反馈,资源荐购。而新

兴社交媒体的兴起也为读者参与互动提供了便捷条件,如QQ、微博、微信等,由此产生了大量的非结构化数据。

③移动互联数据。随着移动图书馆的兴起和读者阅读模式的转变,由此而产生了大量有关用户位置、移动路线和阅读爱好等方面的信息。

④主数据。主数据指系统间共享数据,与记录业务活动、波动较大的交易数据相比,主数据变化缓慢。主数据必须存在并加以正确维护,才能保证交易系统的参照完整性<sup>[2]</sup>。常用的主数据有客户、合同、供应商、合作伙伴、雇员。简言之,主数据包含了组织核心业务实体的数据,可以在组织内跨越各个业务部门被重复使用,如图书馆员工构成、部门层次关系、提供的各种服务构成的产品主数据等。将各种不同类型和格式的数据进行集成通常需要使用到与非结构化的数据相关联的键或者标签(或者元数据),而这些非结构化数据通常包含了与客户、产品、雇员或者其他主数据相关的信息。对于集成结构化和非结构化数据来说,元数据和主数据是非常重要的概念<sup>[3]</sup>。

⑤元数据。在图书馆与信息界,元数据被定义为:提供关于信息资源或数据的一种结构化的数据,是对信息资源的结构化的描述。其作用为:描述信息资源或数据本身的特征和属性,规定数字化信息的组织,具有定位、发现、证明、评估、选择等功能<sup>[4]</sup>。随着元数据的发展,如今的元数据可以用来描述各类型数据,不一定是数字形式的,可来自不同的资源。高校图书馆元数据主要包括数据库的元数据如数据集

<sup>\*</sup> 本文系江苏省盐城市图书馆学会2015年度学术研究课题“面向大数据的高校图书馆数据集成研究”(项目编号:YTX201507)成果。

的物理位置、名称、关系、字段、约束等,读者群的元数据如读者年龄、学历、专业、地理位置等,数据转换的映射关系,操作元数据的算法等。

## 2 高校图书馆大数据的特性

有关大数据时代的数据,有研究人员总结和概括出4V特征<sup>[5]</sup>,即容量、多样性、速度、价值,笔者仅对数据多样性、分布式存储特点及数据可用性再做进一步阐述。

①数据多样性。包括来源多样、存储格式多样、数据类型多样。图书馆数据来源多种多样,有的来自历史数据,有的来自读者互动的实时更新数据;在存储格式上,或为数据库,或为Excel,或为HTML;除了可以从传统的关系型数据库获取大量的结构化数据之外,庞大的可用外部数据通常来自社交媒体,而这些数据往往是非结构化的,不同结构的数据给图书馆数据集成带来了困难,但这又是不可避免的问题,因为图书馆作为信息服务机构,不能闭门造车,要以用户的需求为第一位。从社交媒体或者移动设备上获取的数据,如果能够挖掘出其中的价值,对于推进和优化图书馆服务至关重要。

②分布式存储。不同来源、不同格式的数据有时会分散存储在不同的服务器上,数据的使用、更新等操作不在同一处或者所有者、权限管理者不同,当数据的容量非常庞大时,单一的线性合并数据集的方案耗费时间和空间,已无法满足大数据集成的需要。另一方面,由于分布式权限问题,我们必须要考虑数据的安全访问层次问题。

③数据可用性。诚然,大数据的价值是巨大的,每个组织存储的数据量也非常可观,然而信息劣质、数据错误、数据重复的问题也普遍存在,这是信息化社会固有的问题。一个正确的大数据集至少应该满足5个性质:一致性、精确性、完整性、时效性、实体同一性<sup>[6]</sup>。具体就图书馆而言,应用系统的不断更新升级,图书馆从业人员的素质参差不齐,对数据的重视程度不够,导致数据冗余、重复、错误,图书馆要想从大数据中挖掘价值,对于现有数据的集成整治是首要课题。

## 3 面向大数据的高校图书馆数据集成架构

完整的数据集成过程包含了对数据的访问、解析、转换和清洗,以及抽取和交付数据等,核心功能是对数据的抽取、

转换和加载(ETL),即从源数据存储系统获取数据之后,转换成目标系统所兼容的格式,再将其导入目标系统中。目前常用的数据集成方法有联邦数据库方法、中间件集成方法、数据仓库方法。

数据仓库技术可以将组织多年积累的历史数据唤醒,不仅为组织管理好这些海量数据,而且挖掘数据潜在的价值。对于高校图书馆而言,数据库包含了大量结构化数据如文献数据、读者借阅数据,如能充分集成分析,将有助于图书馆馆藏资源的优化和读者服务的提升。因而现有的数据仓库技术无疑应当成为高校图书馆大数据架构的重要组成部分。然而,由于组织每天都有大量的数据产生,向数据仓库加载新的数据源总是需要很长的时间,一方面无法做到实时响应,另一方面也对组织的存储能力提出挑战。基于此,在数据仓库之上增加一个中间件,在中间件层上存在一个虚拟数据服务层,将数据仓库作为数据虚拟化服务器的数据来源之一。中间件层既能集成结构化数据,也能集成非结构化数据,将其构建于数据仓库之上,以实时的方式集成数据仓库中的数据 and 当前数据,用户基于全局视图通过中间件层访问数据,充分发挥大数据分析的作用。

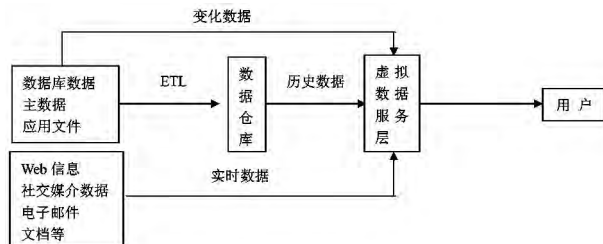


图1 面向大数据的高校图书馆数据集成架构

### 3.1 数据仓库集成架构

将图书馆应用数据库数据、主数据及其他应用文件经ETL工具集成到数据仓库中,使数据仓库成为中间虚拟数据服务层的数据来源之一,用户通过中间层访问数据仓库中的数据。数据仓库中的信息具有稳定性和历史性,图书馆应用数据库中的文献数据和用户阅读数据系统记录了图书馆从使用该数据库以来到当前阶段收录的文献信息情况及用户使用情况,依据这些信息,可以对图书馆文献信息的发展历程和未来趋势做出定量分析和预测。主数据如图书馆员工构成、部门层次、信息服务产品,应用文件如来自外部的供应商合作文件、某一节点上发生的事件等都具有稳定性,一旦进入数据仓库,一般将被长期保存下来,供用户查询。

诚然,进入数据仓库集成的基本都是结构化数据,基于

结构化数据的数据仓库有确定的生命周期,数据从源系统抽取出来,装入暂存区并进行清洗和优化,再依据转换表进行数据转换,最后加载进数据仓库,读取效率高。然而,对于一部分非结构化的应用文件的集成则需要借助主数据和元数据。例如,一份读者意见调查表通常反映的是读者对于图书馆某项服务提出的意见和建议,我们首先搜索到该数据,通过分析文本,明确其关联的是与图书馆员工或者服务产品等主数据相关的信息,进而给该数据贴上主数据的元数据标签,再进行数据的转换和加载。

### 3.2 虚拟数据服务层架构

除了集成数据仓库数据之外,虚拟数据服务架构还集成另外两个来源的数据:实时交互数据、数据库中有变化的数据。大数据背景下,新兴社交媒体不断涌现,图书馆服务主动化,充分利用各种渠道增进与用户的交互性,让用户更多地参与到图书馆的服务和管理中来,于是产生了大量的Web信息和社会媒介数据、电子邮件等一道道“消息”构成的实时互动数据;数据仓库中存储稳定的数据,但源数据库中的数据是定期加载、刷新的,如果将新的数据源不断增加到数据仓库,往往都要重复复杂的加载过程,需要耗费很长的时间。利用变化数据抓取工具(CDC, Changed Data Capture)从数据库日志中提取变化数据,并且变化的数据被保存在数据库的变化表中,等待进一步集成处理。

虚拟数据服务层通过不同的适配器与数据层的各种数据源实现链接,将数据源中的各种数据实体映射成中间件的虚拟数据层的表,虚拟数据层中的表都只有元数据,而不存储实际的生产数据。用户可以在虚拟数据层上采用可视化图形界面定义数据映射关系,进行数据加工整合,这些数据加工逻辑一般会以文件或者数据库方式存储。当用户通过中间件访问虚拟数据层的数据时,虚拟数据层根据系统定义的逻辑首先将需要加工的细节数据从各个数据源抽取到虚拟数据层,然后中间件根据设计时的数据加工逻辑对其进行加工,最后中间件将加工好的数据以调用接口要求的格式返回。

### 3.3 元数据管理

从前文所述来看,元数据贯穿整个大数据架构。数据仓库架构中的元数据管理主要集中在对数据仓库ETL过程的管理,包括数据源元数据,主要记录源数据的含义、描述信息、物理状态、版本信息等;操作型元数据,包括数据的使用、更新记录、数据抽取转换规则、数据检查和清洗规则等;技术

型元数据,包括数据的来源、系统响应时间记录、许可及安全数据等。与实时数据集成有关的元数据和数据仓库集成元数据非常相似。

虚拟数据服务层在访问每个不同的数据源时,都需要导入和集成相关的元数据,因而完整描述数据的元数据应当随着抽取数据一起传输。元数据对于大数据集成架构至关重要,对元数据的管理应该形成机制。

## 4 结语

大数据环境下的图书馆数据集成系统构建是一项复杂而困难的工程。技术层面上,除了大数据集成架构外,对主数据和元数据的管理、数据的安全等也有待进一步研究;组织层面上,不同类型的数据分散在各个部门,由不同的人管理和负责,很多图书馆人虽然知道“大数据”一词,但对于将不同部门的数据进行整合利用尚缺乏主动意识;人才层面上,目前绝大多数图书馆员多是业务专家,在原始数据的清洗和质量检查环节能够发挥很好的专业指导作用,但尚不具备数据整合的能力,图书馆应加强数据挖掘和分析专业人才培养。此外,系统构建应坚持经济的原则,各高校图书馆可以通过合作联盟的方式共同建设。

## 参考文献:

- [1] 涂子沛.大数据[M].桂林:广西师范大学出版社,2013:57.
- [2] 百度百科.主数据[EB/OL].[2015-07-23].<http://baike.baidu.com/view/402047.htm>.
- [3] April Reeve 著;余水清,潘黎萍译.大数据管理:数据集成技术、方法与最佳实践[M].北京:机械工业出版社,2014:8.
- [4] 百度百科.元数据[EB/OL].[2015-07-23].<http://baike.baidu.com/view/107838.htm>.
- [5] 樊伟红,等.图书馆需要怎样的“大数据”[J].图书馆杂志,2012(11):63-68,77.
- [6] 李建中,刘显敏.大数据的一个重要方面:数据可用性[J].计算机研究与发展,2013(6):1147-1162.

卓建霞 女,1983年生,馆员。研究方向:图书情报。

(收稿日期:2015-09-08;责编:张欣。)