

●郭少友(郑州大学信息管理系 河南 450052)

浅谈企业数据集成

Abstract: The paper introduces the necessity and general method of enterprise data integration. Then the effects of XML on enterprise data integration are discussed and the main problems in the process of enterprise data integration are listed.

Keywords: enterprise; data conversion/data integration; XML

1 企业数据集成的必要性

很多企业在发展过程中,都开发或引进了许多独立运行的应用系统,每一个应用系统都有自己的运行环境和数据存储方式,从而产生了各种不同的数据源。各个应用系统彼此封闭,数据不能交换和共享,数据源之间数据格式和代码不统一、数据大量冗余,从而形成了大大小小的“信息孤岛”。

随着企业之间竞争的加剧和企业信息化进程的加快,同一企业内部各个部门越来越需要沟通,以便协调工作,提高工作效率,从而达到提高企业竞争力的目的。沟通的手段是共享信息,沟通的桥梁则是应用系统,于是一个应用系统访问其他应用系统的数据就变得越来越重要、越来越频繁了。而要使这种访问得以实现并变得快捷,就需要将企业的各种异构数据源集成起来。

从外部环境来看,随着信息技术与经济的紧密结合,电子商务已经引起了企业界和商业界的广泛关注,它是一种能使企业在新经济形势下保持竞争优势、尽快抓住商机的重要营销方式。电子商务将买方、卖方、金融部门、中间商、物流管理部门连为一个整体,企业逐渐从网络上的一个孤立的节点发展成为不断与网络上其他节点交换信息和进行商务活动的实体,企业数据集成也从企业内部集成走向了企业间集成。现在的企业比以往任何时候都需要将内部数据进行发布和交换,这必然导致越来越多的企业应用系统需要访问各种异构数据源,并且这些数据源可能分布在网络上任何地方。同样地,为了使这种访问得以实现,需要将企业内部的各种异构数据源以及商务链上合作伙伴的异构数据源集成起来。

2 企业数据集成的方法

企业数据集成的方法可分为3类:第一类为数据转换方法,这种方法实现的是企业数据的松散集成,通过转换工具实现应用系统之间的数据转换和交换,从而达到集成的目的,是一种较低层次的集成;第二类为数据聚合方

法,在各种异构数据源的基础上,借助于中间件系统构造一个虚拟的全局数据模式,是一种集中式管理、分布式存储的较高层次的集成模式;第三类是析取、转换和装载(Extract、Transform and Load, ETL)方法,通过对异构数据源中的数据进行分析、转换和装载,建立一个数据仓库,是一种面向企业决策的数据集成方法。

2.1 数据转换

数据转换方法是一种传统的数据集成方法,相对于其他方法来说,技术上较为简单,比较容易实现,目前在很多领域仍然是一种主要的数据集成方法。数据转换方法通过转换工具在数据库之间进行模式映射,将一个数据库中的数据复制、转换为另一个数据库中的数据,从而实现数据库之间的集成。

目前能实现数据转换的工具很多,大致可以分为3种:一是各种数据库管理系统(DBMS)自带的转换、迁移工具;二是应用系统内部集成的转换工具;三是通用的、集成的数据转换工具。

目前绝大多数DBMS都带有数据转换、迁移工具,以实现本系统与其他DBMS甚至非结构化数据库系统的数据交换。如Oracle的Migration Workbench, Microsoft SQL Server的数据转移服务(DTS),VFP的导入导出工具和升迁工具等。这些数据转换工具大都能完成常见种类的异构数据库之间的转换,但是也有一定的局限性,即它们都属于特定的数据库系统,通用性不强。

应用系统内部集成的数据转换工具是本系统与其他应用系统之间的数据接口。从规范性的角度来看,这类数据接口分为两种情况:一是在相关的应用系统之间进行数据转换的接口,接口参数完全由设计人员自行规定,主要用于企业内部各应用系统之间的数据转换,而且要求交换数据的双方应用系统由相同的设计人员设计,如果出自不同设计人员之手,则应用系统所有者双方应取得共识,就数据转换达成一致的意见;二是遵循某种国家标准或国际标准的转换接口,应用系统通过转换接口将数据转换为标准格式,并传递给目标应用系统,目标应用系统中遵循同样

标准的转换接口再将其转换成内部数据格式。第一种类型的转换接口局限性较大,只适合在指定的范围内转换数据,没有通用性;第二种类型由于遵循一定的标准,使用范围较广,典型代表是目前仍在使用的各种电子数据交换(EDI)软件,主要用于企业之间的数据转换。

EDI已经成为商业、零售业、外贸部门、进出口业、工业、化工、石油、汽车等众多行业进行数据转换、交换的工具。传统EDI的工作流程是:用户应用系统从数据库中取出用户格式数据,通过映像程序将数据展开成平面文件,再由翻译器按照EDI标准将平面文件翻译成EDI报文;通信软件将EDI报文通过网络传送到网络中心;贸易对方通过通信线路从网络中心读取EDI报文,经过EDI翻译器转换成平面文件,再经过映像程序转换成用户格式数据,存入本地数据库中。随着因特网的发展和XML技术的引入,基于因特网的XML/EDI方式正逐渐取代传统的EDI方式。其工作流程是:用户从自己的数据库中提取出所需的数据并将其转化为标准的XML文档(该XML文档需遵循行业内共同遵守的一套XML Schema),应用系统通过HTTP协议将XML文档传送到交换伙伴的电子邮箱或指定的FTP目录中,交换伙伴从相应的位置接收XML文档,按照约定的XML Schema对传来的数据进行校验,通过XML解析器取出XML文档中的数据并保存到自己的应用系统之中^[1]。

集成转换工具是脱离具体的DBMS和应用系统,通用性很强的独立软件,可以在任意两个常见的数据源之间进行数据转换。目前成熟的集成数据转换工具比较多,一般都采用向导驱动方式和GUI图形用户界面,可以进行Fox-Pro、Access等桌面数据库与Oracle,SQL Server等大型数据库之间的数据存储、转换和调用。一般来说,这类工具都允许把一个数据库中的数据(一个或多个表中的部分或全部行)转入至另一个数据库的一个表中(这个表可以存在或不存在)。

2.2 数据聚合

数据聚合方法是将多个数据库集成为一个统一的数据库视图的方法。可以认为,数据聚合工具产生的数据聚合体是一种虚拟的企业数据库,它包括了多个实体的物理数据库。

数据聚合方法利用中间件集成异构数据源,该方法并不需要改变原始数据的存储和管理方式。负责数据集成的中间件系统位于异构数据源(数据层)和应用程序(应用层)之间,向下协调各数据库系统,向上为访问集成数据的应用系统提供统一的全局数据模式和数据访问的通用接口。各数据库的应用仍然完成它们原来的任务,中间件系统则主要为各种异构数据源提供一个高层次检索服务。任

何对其他应用系统数据源的访问,都将通过中间件系统,并由中间件系统负责数据的模式映射和转换工作。显然,基于中间件系统的数据聚合模式是实现异构数据集成较理想的解决方案^[2]。随着中间件技术的不断成熟和推广,数据聚合将会成为企业数据集成的一种重要方法。

2.3 ETL

ETL方法是一种实现异构数据源的集中式管理、集中式存储的方法。ETL工具从多个数据源中抽取数据,然后进行数据转换和加载,最终得到统一的、完备的数据仓库。原来分散的应用系统仍然独立运作,原来存在的异构数据源仍然为各自的应用系统提供数据服务。这种集成方法的特点是:不会破坏企业原有的应用架构,比较适合于大量数据的迁移,可以提供复杂的数据转换功能,可以集成多种数据源和复杂的商业规则,能容忍数据在时间上的延迟等。

ETL工具与数据仓库技术密切相关。数据仓库是一种面向主题的、集成的、稳定的、包含历史数据的数据集合,它能够分布在企业网络中的不同站点的商业数据集成到一起,为决策者提供各种类型的、有效的数据分析,起到决策支持的作用。数据仓库在各种异构数据源(包括结构化数据源和非结构化数据源)的基础上建立统一的全局模式,用户可以通过数据仓库提供的统一的数据接口进行决策支持方面的查询。

专门提供ETL工具的厂商包括Ascential, Acta, Information, SAS, Iway等公司,它们的ETL产品一般都提供多种数据析取、转换适配器,同时允许大批量的数据转换和实时的数据操作。与前两类数据集成工具不同,ETL是基于数据库级的集成工具,用它进行数据集成不涉及应用的集成。

3 XML在企业数据集成中的作用

不同的应用系统(尤其是不同企业的应用系统)开发语言不同、部署平台不同、通信协议也可能不同,对外交换的数据格式更是可能有着巨大的差异。这些不同和差异给系统集成带来了巨大的困难。从1998年开始发展的XML技术及其相关技术是解决这些困难的初步尝试。XML技术的提出,其初衷是为了改善HTML的无结构化状况而造成的全球Web信息的结构混乱。XML规范的开发小组为了使得全球Web信息能够迈向结构化的方向,基于强大的SGML语言制定了XML 1.0规范。由于XML解析器在各种平台上都被开发人员使用,所以大家不约而同地发现,使用XML在不同的异构系统之间交换数据是一件非常方便的事情。首先,XML格式具备描述各种类型数据的能力;其次,DOM/SAX针对XML文档封装了一套有效

的处理方法,开发人员可以使用 DOM/SAX 对 XML 文档进行处理,不必自行开发文档格式处理模块;再次,XML、DOM 是 W3C 规范,大家都会共同遵循,在不同平台上的处理方式是完全一致的。XML 解决了在不同平台/系统之间的数据结构模式的差异,使得数据层在 XML 技术的支持下统一起来。因此,XML 很快就成为应用范围极为广泛的数据交换的工具^[3]。目前使用 XML 进行数据交换已经成为计算机软件领域,尤其是电子商务应用领域的标准技术模式。

由于 XML 具有极强的适应性并得到多方支持,使其可以实现对资源的快速包装和集成发布。XML 技术与全局数据模式相结合可以使异构数据源集成中间件系统能更好地适应于开放、发展环境(如企业的动态联盟环境)中的数据集成。许多著名的异构数据源集成研究项目和实用系统都引入了 XML 及相关技术,例如中科国际软件有限公司的异构数据源集成中间件系统 A2E - DataIntegrator^[4]、IBM 的 GARLIC 项目等。

4 企业数据集成过程中应注意的问题

异构数据源集成是数据库领域的经典问题,并随着 XML 技术和中间件技术的兴起,再次成为该领域的一个研究热点。尽管目前已经有不少比较成熟的数据集成方法和相应的工具投入到实际的应用中,然而由于企业应用系统的复杂性、异构数据源的多样性等诸多因素的制约,使得企业数据集成过程变得相当复杂。为了使企业数据集成工作能顺利进行并能取得比较满意的效果,应该认真考虑以下问题。

1) 集成范围问题。企业应用系统涉及到的数据源可能包括结构化数据库、文本文件、HTML 文件、多媒体文件等多种形式,企业数据集成并不是将所有数据源中的所有数据全部集成,那么集成哪些数据源以及数据源中的哪些数据,则是首先要考虑的问题。

2) 数据资源所有权问题。不同的数据资源可能属于不同的独立经济核算部门,如何在访问异构数据源数据基础上保障原有数据资源的权限不被侵犯,实现对原有数据资源访问权限的隔离和控制,也是企业数据集成过程中应该解决的问题。

3) 全局模式问题。数据聚合方法和 ETL 方法分别需要建立统一的虚拟数据库和数据仓库,这两种数据集成方法都需要在各个异构数据源局部模式的基础上充分做好元数据工作,从而建立全局模式。不同的数据源针对同一对象建立起来的局部数据模式往往差别较大,如何在不变更现有异构数据源结构的基础上建立一个能与它们兼容的全局数据模式,应该说是一项比较艰巨的任务。

4) 模式映射问题。这是对前一个问题的补充。如果经过各方协调已经建立有全局数据模式,应用系统的数据库与虚拟数据库或数据仓库之间的模式映射问题相对简单一些。如果采用数据转换方法进行数据集成,由于双方的数据模式不同、语义不同,较难进行准确的模式映射,转换后的数据往往不能完全准确地表达源数据的信息(EDI 方式由于有共同遵守的 EDI 报文标准,模式映射相对比较准确,除非用户的 EDI 系统没有按 EDI 报文标准来规划输入数据)。

5) 数据动态集成问题。企业数据集成的目标并不是建立一个大的统一的数据库并抛弃旧有的数据库,而是在将各个数据源的数据集成到某个数据库(某个应用系统的数据库或者整个企业的数据仓库)的同时还保留原来的数据库,并继续为相应的应用系统提供数据服务。因此,企业数据集成应该是一个动态的过程。集成数据库需要经常从变化的数据源中集成新的数据,数据库管理员需要选定合适的周期来刷新集成数据。□

参考文献

- 1 李翔.电子商务.北京:机械工业出版社,2002
 - 2 周竞涛.企业异构数据源集成. Available from: <http://www.ccw.com.cn/html/center/tech/02.8.30.2.asp>
 - 3 柴晓路.Web 服务带来了新集成.中国计算机报,2002(28)
 - 4 <http://www.chinatech-bj.com/page/product/a2edate/a2edate.htm>
- 作者简介:郭少友,男,1965 年生,讲师。发表论文 10 余篇。
- 收稿日期:2002-11-08

欢迎订阅《情报学进展》第四卷

由中国国防科技信息学会编辑出版的《情报学进展》第四卷(2000—2001 年度评论)已于 2001 年 10 月正式出版。《情报学进展》自 1995 年首次在我国问世以来,每两年出一卷,至今已出版了 4 卷。该书跟踪报道和研究分析情报学最新学术动态和研究成果,受到我国情报界人士广泛欢迎,并被一些高等院校作为考博参考用书。《情报学进展》第四卷内容包括:洛特卡定律的研究,社会信息学的形成与发展,国外因特网信息服务发展概述,国外信息查寻行为研究进展,网络环境下科学信息交流的研究,网络科技信息服务技术现状和发展趋势,数字图书馆的发展,XML 技术和应用,元数据研究,Z39.50 标准、协议及应用和网络信息查寻相关缩略词等。全书约 30 万字。定价 25 元(含邮费)。

欲订购《情报学进展》第四卷者,请向北京 2413 信箱 10 分箱《情报理论与实践》编辑部索取订单,邮编:100089。