谈数据仓库建设中的 ETL 过程

ETL processing in data warehouse project

张 云 （浙江机电职业技术学院 310013）

摘要：本文介绍了数据仓库建设中的 ETL 过程，包括 ETL 的概念、目标以及如何正确实施 ETL 以保证数据仓库成功。

关键词：商业智能 数据仓库 元数据 ETL

1 引言

随着市场竞争的加剧和信息社会需求的发展，企业需要从海量信息数据中提取（检索、查询等）制定市场策略的信息就显得越来越重要了。这种需求既要求联机服务，又涉及大量用于决策的数据，而传统的数据库系统已无法满足这种需求。其具体体现在三个方面：

（1）历史数据量很大，而业务系统的数据模型是针对事务处理设计的，不适合做分析。

（2）辅助决策信息涉及许多部门的数据，而不同系统数据之间的不一致性突出。

（3）由于访问数据的能力不足，它对大量数据的访问性能明显下降从而会影响现有系统的运行。

商业智能（BI，Business Intelligence）正是在这种背景下逐渐兴起的，它利用海量的数据构建信息系统，是现代企业运用科学管理、决策分析的基础。商业智能就是为企业把数据转换为信息、知识，相应蕴育而出的 IT 技术。企业级 BI 的基础就是一个完整的、准确的、统一视角的数据平台，即数据仓库（DW – Data warehouse）。而在整个数据仓库项目中 ETL（数据抽取）规则设计和实施是工作量最大的，其工作量要占整个项目的 60% –80%，这是国内外在众多数据仓库实践中得到印证的。因此正确的实施 ETL 是关系到数据仓库建设成功的关键。

2 相关术语介绍

（1）DW：Data Warehouse（**数据仓库**），W. H. Inmon 对数据仓库的定义为：数据仓库是支持管理决策过程的、面向主题的、集成的、稳定的、不同时间的数据集合。

（2）Metadata：元数据。元数据是描述数据仓库内数据的结构和建立方法的数据，是描述数据的数据。元数据的典型表现为对象的描述，即对数据库，表，列，列属性（类型，格式，约束等）以及主键/外键关联等等的描述。

（3）ETL：ETL 分别是三个单词的首字母缩写（Extract、Transform 、Load ）也就是抽取、转换、装载，我们通常简称其为数据抽取。ETL 包含了三方面，首先是‘抽取’：将数据从各种原始的业务系统中读取出来，这是所有工作的前提。其次‘转换’：按照预先设计好的规则将抽取得数据进行转换，使本来异构的数据格式能统一起来。最后的‘装载’：将转换完的数据按计划增量或全部的导入到数据仓库中。

3 如何正确实施 ETL

ETL 按照统一的规则集成并提高数据的价值，是负责完成数据从数据源向目标数据仓库转化的过程，是实施数据仓库的重要步骤。以下就如何正确实施 ETL 进行介绍。

3.1 统一的元数据

目前在业务应用系统的异构性与分布性越来越普遍的情况下，统一的元数据就显得愈发重要了。随着企业的信息系统不断增加，而形成的一个个”信息孤岛”是很多企业目前应用的现状。统一、合理的元数据则能有效的描绘出信息的关联性。

元数据对于 ETL 过程产生影响的集中表现为：

（1）定义数据源的位置及数据源的属性。

（2）确定从源数据到目标数据的对应规则。

(3) 确定相关的业务逻辑。

(4) 在数据实际加载前的其他必要的准备工作。

元数据贯穿整个数据仓库项目,ETL 的所有过程必须最大化的参照元数据。在 ETL 实施过程应该特别注意,以下几个有关元数据的实际问题:

① 元数据在一个大型项目中很容易被虚化。为了迁就现有的业务系统,为了能在指定的时间内系统上线,元数据又不像数据展现等功能可以让业务人员直接看到工作结果,往往元数据就因此被舍弃掉。

② 如何提交一份可以真正帮到 ETL 实现的元数据,以便利用到项目前期的需求分析及业务系统调查结果。

③ 在赶工期的时候,如何协调模型的变化及 ETL 的关系。

3.2 合理的业务模型设计

合理的业务模型设计对 ETL 至关重要,数据仓库是企业唯一、真实、可靠的综合数据平台。数据仓库的设计建模一般都依照三范式、星型模型、雪花模型。数据仓库的模型设计通常都采用星型(STAR SCHEMA)模型设计方法。

无论哪种设计思想,都应最大化的含盖关键业务数据,把业务系统中杂乱无序的数据结构统一成为合理的、关联的、分析型的新结构,而 ETL 则依照模型的定义去提取数据源,进行转换、清洗,并最终加载到目标数据仓库中。业务模型的重要之处在于对数据做标准化定义,实现统一的编码、统一的分类和组织。

ETL 按照业务模型按以下顺序进行数据集成:初始加载、增量加载、缓慢增长维、慢速变化维、事实表加载等,根据业务需求制定相应的加载策略、刷新策略、汇总策略、维护策略。每一个环节都是艰巨而复杂的任务。

3.3 高质量的数据

经验告诉我们,纵然整个数据仓库的业务模型构架非常理想,但是数据的质量往往是最致命的。一个数据仓库成功是非常依赖于数据的质量。

影响数据质量问题的因素有很多,由源数据造成的因素包括:

(1) 数据格式错误,例如缺失数据、数据值超出范围或是数据格式非法等。要知道对于处理大数据量的数据源系统,通常会舍弃一些数据库自身的检查机制,例如字段约束等。

(2) 数据一致性,同样,数据源系统为了性能的考虑,会在一定程度上舍弃外键约束,这通常会导致数据不一致。例如在航空运单表中会出现一个航空地理表中没有的机场 ID,有些价格类型代码在代码表中找不到等。

(3) 业务逻辑的合理性,通常,数据源系统的设计并不是非常严谨,尤其当客户的业务系统本身对业务逻辑的约束就很弱的情况下,数据之间的逻辑联系就更加难以保证。

而 ETL 过程中对数据准确性产生重大影响的因素包括:

① 规则描述的错误。设计人员对数据源业务系统理解的不充分,导致规则理解错误,这是一方面;另一方面,是规则的描述,如何无二义性地描述规则也是要探求的一个课题。规则是依附于目标字段的,但是规则总不能总是用文字描述,必须有严格的数学表达方式。

② ETL 开发的错误。即使规则很明确,ETL 开发的过程中也会发生一些错误,例如逻辑错误、书写错误等。例如对于一个分段值,开区间闭区间是需要指定的,但是开发人员没注意,一个大于等于号写成大于号就导致数据错误。

③ 人为处理的错误。在整体 ETL 流程没有完成之前,为了图省事,通常会手工运行 ETL 过程,这其中一个重大的问题就是不按照正常流程去运行了,而是按照自己的理解去运行,发生的错误可能是误删了数据、重复装载数据等。

3.4 正确的数据集成

为确保获取高质量的数据,ETL 过程必须进行正确的数据集成,包括对数据进行预定的转换处理,转换处理中的具体方法有:

(1) 字段映射值:指定两个数据源。比较第一个和第二个数据的值。当两者的值相同的时候,就将指定的映射值复制到输出数据中。

(2) 计算值:通过 SQL 语言来生成一些新的字段。

(3) 过滤字段:可以通过指定一些字段,然后选择对这些字段是保留下来,还是保留除这些字段以外的其他字段。

(4) 字段丢失值: 就是指定一个或多个字段,对每个字段搜索丢失值,并给予一个值。例如:航线类型字段为空的话,我们可以给它一个"国内"的值。

(5) 字段无效值: 就是对于一个字段的值,不是一个有效值的话,我们可以给它一个缺省的值。例如:对于航班状态字段,如果是"正常"、"延误"、"取消"值之外的话,我们可以给它一个缺省值"异常"。

(6) 聚集值: 使用它来产生输出数据,其中包含输入数据中的聚集值。可以同时提供多个聚集表达式和多个新字段名。

(7) 空值的处理: 可捕获字段空值,进行加载或替换为其他含义数据,并可根据字段空值实现分流加载到不同目标库。

(8) 规范数据格式: 可实现字段格式约束定义,对于数据源中,时间、数值、字符等数据,可自定义加载格式。

(9) 拆分或者合并数据:依据业务需求对字段可进行分解或是合并。

(10) 验证数据正确性:进行数据验证。

(11) 获取随机样本:通过指定要抽取的样本的比例,从原始数据中采样出一个较小的样本。

(12) 建立 ETL 过程的主外键约束:以保证主键唯一记录的加载,实现数据规则过滤。

ETL 过程中数据的正确性很大程度上受客户企业对源数据的理解程度的影响,因此从业务的角度看数据集成非常重要。确实,谁也不能绝对把握数据的正确,不仅是系统集成商,包括客户也无法确定。准确的东西需要一个标准,但首先要保证这个标准是准确的,至少现在还没有这样一个标准。客户也许会提出一个相对标准,例如将你的 OLAP 数据结果和客户现有的统计报表结果进行对比,但显然这是一种不公平的比较,因为客户现有的报表数据本身可能就是不准确的。

因此为了能更好的实现 ETL,我们给出以下几点建议:

① 推荐利用数据准备区对业务数据进行预处理,以保证集成与加载的高效性。

② 开始 ETL 之前,应事先制定流程化的配置管理和标准协议;而绝大多数情况下,数据更新过程就是一个流程化的处理过程。

③ 关键数据的标准至关重要。ETL 面临的最大挑战是当接收数据时其各源数据的异构性和低质量,以某航空公司为例:A 系统按照运输角度管理,B 系统按照财务角度管理,而 C 系统按照 SITA 系统管理。而 ETL 需要对这三个系统进行集成以获得对客户和业务的全面视角。这一过程需要复杂的匹配规则、航班类型/机场匹配正常化与标准化。而 ETL 在处理过程,会定义一个关键数据标准,在此基础上,制定相应的数据接口标准。

④ 在数据仓库项目中,和多个外围系统进行数据接口是一个不可避免的模块。为数据接口建立概念模型是一项有益的工作,因为在数据仓库项目中,数据接口的工作其实占用了项目组非常多的非技术时间。因为接口的定义不清晰,职责不明确,涉及到本方、客户还有若干外围厂商。而数据接口在项目进程中还处于上游,它的定义不清给后续的 ETL 工作带来的麻烦不小。

3.5 影响 ETL 性能的因素和异常原因分析

在 ETL 实施中,也许会出现 ETL 性能的问题,我们可以查找可能的性能瓶颈的因素,包括:硬件;操作系统;网络;数据库;ETL 应用程序的效率等等。

而导致 ETL 异常发生的原因,大致可以分为以下五类。有一些是硬性的,有一些是软性的,有一些是环境导致的,有一些是流程导致的。注意下面提到的原因是针对整个 ETL 过程的,从抽取到转换到装载包括:

(1) 硬件、操作系统、网络导致异常;

(2) 数据源数据传输、质量导致异常;

(3) ETL 过程处理导致异常;

(4) 目标数据模型导致异常;

(5) 开发、维护阶段人工干预导致异常;

因此在数据仓库实施过程中,必须注意以上可能导致问题的各种因素。

3.6 常用 ETL 工具介绍

借助专业 ETL 工具可以帮助我们提高通用性和开发效率,减少项目风险,降低后期维护成本。目前专业 ETL 工具厂商和产品包括:Ascential DataStageXE;Sagent Solution;Informatica. 而整体 ETL 方案提供商和产品包括:Oracle Warehouse Builder;IBM Warehouse Manager。

(下转第 83 页)

(上接第 79 页)

4 结束语

总之,优秀的 ETL 必须做到:采用统一的元数据方法,集中进行管理;合理的业务数据模型;接口、数据格式、传输有严格的规范;尽量不在外部数据源安装软件;抽取的数据及时、准确、完整;可以提供同各种数据系统的接口,系统适应性强;提供软件框架系统,系统功能改变时,应用程序很少改变便可适应变化;可扩展性强;管理简单;数据抽取系统流程自动化,并有自动调度功能。

ETL 过程是企业信息数据集成的必经过程,也是揭示业务数据潜在价值的唯一途径。本文则介绍了一个如何成功实施 ETL 的方法。

参考文献

1 W. H. Inmon. Building the Data Warehouse, 2nd edition. John Wiley, 1996.

2 Jose Samos, Felix Saltor, Jaume Sistac and Agusti Bardes. Database architecture for data warehousing: an evolutionary approach. In DEXA 1998: 746 –756.

3 王珊等,数据仓库技术与联机分析处理,科学出版社,1998.6 。

4 Harjinder S. GILL 等著,王仲谋、刘书丹 译,数据仓库 –客户服务器计算指南,清华大学出版社,西蒙与舒斯特国际出版公司,1997.10。