实时数据仓库技术的研究

Research of Rea1 – time Data Warehouse Technology

姜 震 黄 霞 （淮海工学院计算机科学系 江苏连云港 222005）

摘要：实时数据仓库是数据仓库技术的一个新的发展方向。本文研究总结了实时数据仓库的实现技术和体系结构，并重点研究了数据的实时更新技术，在此基础上提出了一种实用性较强的实时数据仓库的实现方法。

关键词：实时数据仓库 ODS 层次化结构 数据更新 事件触发

1 引言

数据仓库经过多年的发展，其技术日趋成熟，在当今信息社会中发挥着重要作用。但在应用中也暴露出一些问题，主要体现在两方面：一：数据的更新问题。首先是缺乏实时性。随着市场竞争的加剧，信息的实时性越来越重要。人们希望没有延迟地获取信息，并据此做出分析和决策。而传统的数据仓库中大多是历史数据，数据抽取周期一般为一天甚至一周。基于传统的数据仓库很难进行实时性处理；其次是数据更新的主动性问题。传统的数据仓库采用 ETL 周期性的进行批量更新，更新的时间和数据都是既定好的，不管周期是否合适以及数据有无变化。效率低下，缺乏主动性、选择性的更新策略。二：数据仓库的使用范围和应用领域狭窄。传统数据仓库主要为制定企业中长期发展的战略性决策提供支持，服务对象是企业的高层管理者或分析员。而激烈的商业竞争要求数据仓库在提供战略性决策支持的同时，更多的给企业提供关于日常运行的战术性决策支持。而且要扩展数据仓库的使用范围，使中层管理者、操作雇员、甚至合作伙伴和客户都可以访问它，让资源得到充分利用。针对传统数据仓库的以上不足，现在开始提出了实时数据仓库的有关理论和技术。

2 实时数据仓库的有关概念和特点

实时数据仓库（Real – time DW）是数据仓库技术的一个新的发展方向。其理论还未成熟，也没有公认的严格定义。本质上实时数据仓库仍然是数据仓库，它的最大特征是实时性，主要体现在数据仓库中数据的实时性变化上。我们可以这样理解：实时数据仓库是这样一个系统：只要 OLTP 系统中的事件（如超市中商品的销售行为）完成产生了数据，这些数据就可以立即被实时数据仓库捕获，并变得可用[1,2]。与传统数据仓库的“快照”形式不同，实时数据仓库中的数据能够同步的反映业务系统（OLTP）中数据的变化，从而及时做出相关分析和决策。显然这非常有利于企业抓住瞬息万变的市场变化，在竞争中处于有利地位。

除了实时数据仓库这一概念外，目前还有一些近似的概念，如动态数据仓库（Dynamic DW）、主动数据仓库（Active DW）等[3]。主动数据仓库主要强调了新一代数据仓库中数据更新和决策支持方面的主动性；动态数据仓库则强调数据仓库中数据是动态变化的，与业务数据的变化同步。以上概念都是对新一代数据仓库技术的不同描述，在本质上是近似的。

3 实时数据仓库的实现技术研究

要实现实时数据仓库，关键技术在于实现数据的实时更新。按照对反应时间的要求，所谓的实时可以分为真正实时和近似实时两种。这里所说的反应时间是指业务系统中事件的完成时间和该事件的数据在数据仓库中可利用时间之间的延迟[1]。真正实时情况下，反应时间以秒甚至毫秒为单位，可以忽略不计；在近似实时情况下，反应时间是一段时间，以分钟为单位。根据数据实时性要求的不同，可采用不同的更新策略。

（1）以传统的 ETL 为基础，只是把周期尽量缩短。这并非真正的实时更新技术，只是模拟了实时效果。

这种方法的缺点是消耗时间和资源的代价很大。如果周期过短,频繁地进行数据仓库更新,系统资源将被大量消耗,从而严重影响到 OLTP 系统的性能和正常操作。所以它的周期一般限制在小时级。适用于实时性要求比较低或业务系统不太繁忙的情况。

(2) 对第一种策略进行改进,每次只更新上次更新后发生变化的数据。这种增量更新使每次更新的数量大大减少,更新周期可以进一步缩短而不至于严重影响 OLTP 系统的性能。实时性比第一种策略高。具体方法有:①为数据源中的数据增加一个修改标记,数据修改时把标记设置为 T。更新时每次只抽取标记为 T 的数据,抽取后再把标记恢复为 F。②对数据源数据加盖时间戳,数据发生变化时修改时间戳。[4] 每次更新只抽取时间戳晚于上一次抽取时间的数据;③通过源数据库日志,决定需要更新的数据。这类策略适用于对实时性要求不高的日常事务处理,如人事管理等。缺点在于每次更新都要进行源数据库的全局扫描,导致性能下降。

(3) 程序监控:在应用程序中编写专门的线程,监控源数据库中的数据变化,一旦捕获到感兴趣的数据变化,就启动数据仓库的更新操作。这种方法实时性较高,缺点是监控线程不停的运行并查询源数据库要消耗大量系统资源,只能用于要更新的数据量很少的情况。比较适用于警告监控等领域。

(4) 建立触发机制。当某一业务事件完成,数据源中的数据发生变化时,触发器立刻负责数据仓库中有关数据的更新。这种方法也可以称为事件驱动机制。具体实现方法可以有两种:第一种是利用源数据库 DBMS 中的触发器来实现。当数据源中的数据发生变化时,建立在数据所在表上的触发器被激活,通过执行 SQL 语句或调用相应的存储过程,负责完成数据仓库的更新。这种方法简单易行,缺点是 SQL 语言功能较为简单,难以实现复杂操作;而且要求源数据库和数据仓库必须在同一物理的数据库系统中;大量触发器的存在消耗了系统资源,也带来更新的并发性和完整性问题。

第二种方法是基于规则的事件触发。用户通过规则定义工具定义好一系列的事件和规则,分别存放在事件库和规则库中,作为数据仓库元数据的一部分。可以将业务数据库的某些特定的数据变化定义为事件;动作就是一个可以执行的程序或操作序列,用于实现数据仓库的更新。可以利用原有的 ETL 工具来实现;规则可以采用主动规则,又称为 ECA(Event – Condition – Action)规则,它由事件、条件和动作三部分组成。当定义的事件发生,如果条件得到满足,即执行定义的动作[5,6]。事件可以通过监控数据库事务日志来实现检测。现在的数据库系统都支持事务日志,日志文件中记录了数据库的每一个数据变化。通过对它的监控,获取事件定义中要求的相关数据,即可实现对事件的检测。当检测到事件发生,就在规则库中寻找与之匹配的规则,若有多条规则被触发时,可以按照一定的优先级从中依次选取规则,并按照规则动作部分的定义进行数据更新。这种数据仓库的主动更新方法效率高,较好的解决了并发性和完整性问题。是最有优势的一种实时更新方法。

(5) 通过源数据库上的应用程序来实现数据仓库的实时数据更新。当业务事件完成后,应用程序将有关数据在写入业务数据库的同时,也把这些数据经过转换、集成载入到数据仓库。这种方法占用资源较少,实时性最高,有较高的参考价值。但增加了应用程序的复杂性和编写难度。而且当具备多个外部异构数据源时,容易造成数据的不一致性和并发操作的问题。

4 实时数据仓库的体系结构研究

实际上实时数据仓库中的数据往往只有部分实时性要求很高,其它大部分数据并不要求实时。基于这一点,目前实时数据仓库的体系结构主要有两种:

4.1 数据仓库的层次化结构

基本思想是:在数据仓库中,将保存的数据分为两类:静态数据和动态数据。静态数据实时性要求不高,主要用于满足用户的战略性 OLAP 和决策要求;动态数据实时性要求很高,主要用于满足实时的战术性 OLAP 和决策要求。根据数据实时性的不同把数据仓库划分为静态层和动态层两层。静态层可以采用传统的 ETL 进行周期性的更新;动态层必须根据实际需求采用前一节提到的某种策略进行更新,以满足不同的实时要求。其结构如图 1 所示:

这种体系结构可以满足企业不同层次的信息处理需要,提高系统开发的速度和效率。缺点是系统既要支持基于实时数据的查询分析和战术性决策,还需要

分析大量的历史数据，进行战略性决策，服务器负担比较重且效率低；动态数据的更新频繁，会影响系统的性能和响应速度。

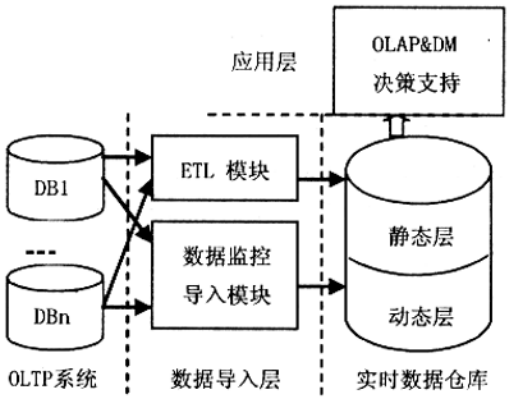

图 1　层次化结构的实时数据仓库

4.2 利用 ODS 技术

ODS(Operational Data Store)即操作型数据存储，是一个存储区域，用于实现对业务数据库中的实时或准实时数据的暂时存储。ODS 的数据具有面向主题、集成的、可变的和实时或接近实时的 4 个基本特征[7]。ODS 中的数据组织方式和 DW 一样也是面向主题的和集成的，但 ODS 只存放当前或接近当前的数据，而且可以动态的对数据进行增、删和更新操作，即和 DW 区别主要体现在数据的可变性和当前性上。ODS 可以看作是介于 DB 和 DW 之间的一种数据存储技术。实时数据仓库设计者可以在常规的、静态的数据仓库之外，建立一个实时的分区(ODS 系统)。其在物理上和管理上独立于传统的数据仓库。ODS 处于业务系统和数据仓库之间，如图 2 所示：

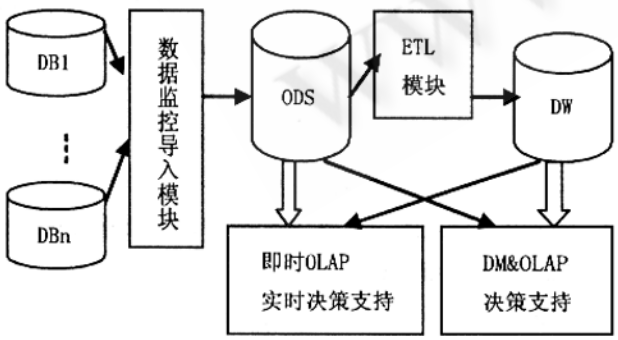

图 2　引入 ODS 的实时数据仓库结构

业务系统产生的数据首先通过某种实时更新策略进入 ODS。ODS 中只保留近期(一般为 1－3 个月)的数据，超过这一期限的数据被 ETL 周期性的导入数据仓库。用户可以直接在 ODS 上进行即时的 OLAP 和实时性战术决策。

这种体系结构的优点：ODS 使操作型环境(DB)和分析型环境(DW)完全隔开，而且 ODS 中的数据也是面向主题和集成的，所以减轻了 DW 导入数据的转换、集成等工作，并降低了数据的管理难度；即时 OLAP 和战术性决策建立在 ODS 上，减轻了 DW 的负担，并且因为 ODS 数据量相对较少，查询和处理的效率高。不足之处：由于中间多一个 ODS 层次，容易造成数据的不一致性。对一些同时关联到历史数据与实时数据的 OLAP 和决策支持不够好，还需要前端工具的支持，才能够实现无缝查询。

5　一种实用的实时数据仓库设计方法

前面两种体系结构都有各自的优缺点，可以将两者结合形成一种新的实时数据仓库设计方法：即在引入 ODS 的同时也把数据仓库分层，其结构如图 3 所示。

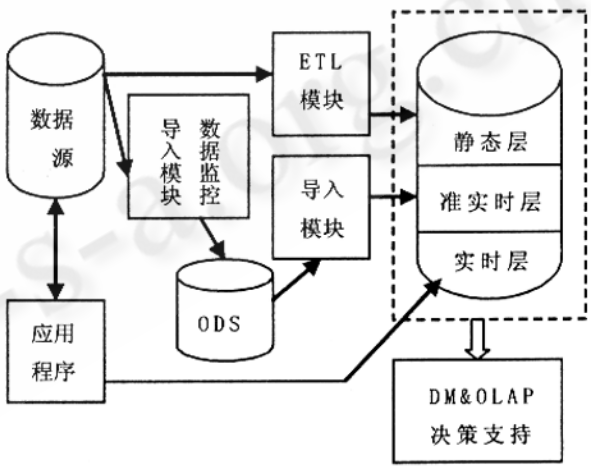

图 3　一种实用的实时数据仓库结构

数据仓库分为静态层、准实时层和实时层三层。静态层中是没有实时性要求的数据，如企业长期发展制定战略性决策有关的历史性数据；准实时层包含实时要求不是很高的数据，如人事变动数据；实时层是一些实时性要求非常高的数据，如警告监控数据，数量较少。在不同的层次中数据的更新技术不同。

静态层可以采用传统的 ETL 进行周期性的更新；

准实时层借助于 ODS 实现。业务系统中产生的准实时数据首先采用某种更新策略(如策略 4 中的事件驱动机制)进入 ODS。然后再视具体的实时性要求,以一定的周期从 ODS 导入准实时层。这一过程可以采用策略 2 实现,即利用 ETL 进行增量更新;实时层采用策略 5 进行更新。由业务系统上的应用程序在产生数据的同时就完成数据的转换和载入,以保证实时性。

这种实现方法对不同实时性要求的数据采用不同的更新策略,数据仓库的三层结构既能够最大限度的满足不同数据的实时性要求,又不会过分增加 DW 更新的负担、影响系统的性能;ODS 的引入简化了到 DW 的数据传输接口,减轻了数据仓库导入数据的负担。另外所有的查询分析和决策都是建立在数据仓库上,这样就有效的解决了基于 ODS 进行查询分析的数据不一致性问题。在电力营销决策支持系统 - 泰安项目的开发中,应用这种设计方法取得了良好的效果。

6 结束语

作为数据仓库技术的一个新的发展方向,实时数据仓库有效地克服了传统数据仓库实时差、难以为企业提供灵活及时的战术性决策等弊端,有着广阔的发展前景。很多数据库厂家都在致力于它的实现,其中 TeraData 取得的进展较大[3]。本文重点研究了实时数据仓库的体系结构和数据更新技术策略,但在实时数据仓库的建模以及基于实时数据仓库的决策支持等方面还没有深入研究,这也是我们下一步的工作方向。

参考文献

1 Simon Terr. Real – Time Data Warehousing www. DMReview. com.

2 John Vandermay, DataMirror Corporation . Considerations for Building a Real – time Data Warehouse . http://www. grcdi. nl/considerations. pdf

3 Stephen Brobst ,Carrie Ballinger . Active Data Warehousing. http://www. teradata. com/t/page/87127/index. html.

4 陆剑峰、张浩,数据仓库数据更新的研究及基于 Oracle 数据库的开发与应用,计算机工程与应用,2004. 6.

5 Thalhammer T, Schrefl M. , Mohania M . Active data warehouse: complementing OLAP with analysis rules. Data and Knowledge Engineering, Volume 39,? Number 3, December 2001 .

6 李庆忠、张抗抗、杨少军、郑永清,一种数据仓库的主动更新方法,系统仿真学报,2003. 2.

7 朱鹏翔、刘文煌,基于 DB – ODS – DW 的 CRM 动态数据仓库,计算机工程与应用,2002. 20.