一种主动实时数据仓库的体系架构分析

张 磊,王备战

(厦门大学 软件学院,福建 361005)

摘 要:传统数据仓库的主要目标是从不同的异构数据源中分析并提取商业战略信息,在传统的数据仓库中,数据不能及时从源系统导入,成功地做出决策会有延时。主动实时数据仓库的提出用来零延时地获取数据和减少正确做出商业决策的时间。本文讨论了主动实时数据仓库与传统数据仓库的区别并进行分析,最终介绍了一种改进的主动实时数据仓库体系结构。

关键词:传统数据仓库;实时数据仓库;主动实时;体系结构

中图分类号:TP311 文献标识码:A 文章编号:1007-3558(2006)04-0111-04

1 引言

在信息化时代的今天,计算机发展日新月异,电子信息数据已经以爆炸般的速度膨胀,存在于各行各业的各个环节中。数据的丰富带来的是对强有力的数据分析工具的需求,快速增长的海量数据被收集、存放在大型和大量的数据库中,如没有强有力的工具,理解它们已经远远超出了人的能力。此时,伴随着海量存储容量的不断提升、CPU 运算速率的飞跃发展以及分布集群技术的不断成熟,数据仓库技术也开始作为计算机软件技术的重要分支登上了这个时代的舞台。

数据仓库经过多年的发展,其技术也日趋成熟,已在商务活动中发挥重要的作用。数据仓库不是为了存储数据,而是为决策支持及更好地组织企业内所有可能收集到的数据。据美国国际数据公司调查,使用该技术的投资回报率平均超过 400%。然而随着市场竞争的加剧,人们越来越关注信息的实时性,“时间就是金钱”再次成为商业社会的至理名言。但是当前的数据仓库系统通常只能分析历史数据,而且数据抽取周期过长,极大地降低了企业的应变能力,难以反映瞬息万变的市场变化,实时数据仓库的出现改变了这个局面。

第一代数据仓库着眼于面向批处理的决策支持能力,第二代带来了联机分析和数据挖掘[1],下一代数据仓库[2],则是以加速信息循环周期、消除信息延时,使不同水平的用户能更用效地利用及时信息而出现的实时数据仓库。实时数据仓库是主动的、动态的数据仓库,它整合了数据仓库和业务系统,提供随需应变的业务。实时数据仓库为诸多企业业务处理过程带来好处,例如:收益管理、反欺诈、商务活动监控、业务流程管理等。根据 Gartner 的研究报告,实时企业(real-time enterprise)已成为趋势,企业具备实时决策的能力将成为市场竞争的最佳武器,目前,已经没有企业不受此影响,而实时数据仓库将成为实时企业不可或缺的基础。为此,我们有必要对主动实时数据仓库的体系结构及构建技术进行探索和研究。

2 主动实时数据仓库概念及其体系架构

2.1 基本概念

实时数据仓库 RTDW (Real-Time Data Warehouse),也就是所谓的“零延迟数据仓库环境”的一部分[3],最早是 Michaem Haisten(一名 BI 专家)提出的,他对实时数据仓库的分类和架构有详细的描述[4]。

现在逐渐提出了实时数据仓库的概念,主要的思想就是:在数据仓库中,将保存的数据分为两类,

作者简介:2006-07-08

作者简介:张 磊(1982—),男,吉林省吉林市人,硕士研究生,主要研究领域为数据仓库、数据挖掘;王备战(1965—),男,陕西西安人,副教授,博士,主要研究方向包括决策支持系统、数据仓库、分布计算系统等。

一种为静态数据,一种为动态数据,静态数据满足用户的查询分析要求;而动态数据就是为了适应实时性,数据源中发生的更新可以立刻传送到数据仓库的动态数据中,其中再经过响应的转换,满足实时的要求。

相对于传统的企业数据仓库而言,主动实时数据仓库增加了主动数据获取和实时数据分析两大主要功能。主动数据获取功能保证了系统能够即时捕捉 OLTP (On-Line Transaction Processing) 系统产生的操作数据;而实时数据分析则能够保证对即时获得的数据做出快速分析,得到决策者想要的结果[5]。

2.2 体系架构分析

传统数据是由 ODS(Operational Data Store)、数据仓库、数据集市和 BI 工具组成。OLTP 系统中的操作型数据在数据仓库的非响应期批处理载入到 ODS 中。ODS 中的数据经过晚间的批处理通过分段传输和集中处理存入数据仓库。BI 工具则利用数据仓库、联机分析处理 OLAP 工具和数据挖掘等技术将数据转化为知识。由此可见,传统数据仓库的一个重要步骤是操作型数据经由 ODS 由 OLAP 转入到数据仓库中去。

而主动实时数据仓库的架构则提倡可省略 ODS 这一中间步骤,即操作型数据在事件产生时由 OLTP 系统中直接载入到数据仓库,免去了批处理作业的麻烦,保证了数据仓库的实时更新。基于 EAI 实时数据仓库实现了滴漏式的数据加载,是真正意义上的实时数据仓库,其数据模型如下[6]:

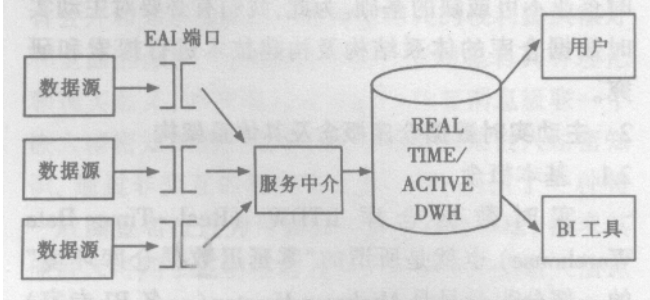

图 1 实时/主动数据仓库模型

3 传统数据仓库与主动实时数据仓库的对比分析

传统的数据仓库不包含当前的数据。它通常是每周或每天个别时间由操作系统加载,但在任何时间都是一个面对过去的窗口。而(主动)实时数据仓库的设计正是为了改善这一情况而提出。这就造成了二者结构及实现的差别。通过对传统数据仓库和实时数据仓库的比较和分析[7][8],就有助于我们区别二种数据仓库的体系结构和设计实现过程。

表 1 传统数据仓库与主动实时数据仓库的比较

	传统数据仓库解决方案	主动实时数据仓库解决方案
战术性和战略性	只能支持战略决策	支持战略决策和战术决策
是否具备主动获取数据能力	不具备主动获取数据的能力	具备主动获取数据的能力
时间粒度	时间粒度较大	可以精确到以分钟为周期获取数据
使用者	分析员、决策者等内部用户	还可为操作员、客户代表和外部用户等
数据载入手段	用 ETL 作为数据载入、更新的手段	可用 EAI 等技术进行数据载入和更新
分析数据方式	基于批处理,它用于脱机性质的分析	提供最新数据实时分析
主要组成部分	采用 ODS、数据仓库、数据集市	将 ODS、数据仓库、数据集市整合到一个大的数据仓库中,基于它存储查询
性能保证	尽可能地提供信息与保证性能	确保信息的可用性及性能

通过上面的比较分析,我们容易看出,实时数据仓库扩展了传统数据仓库的适用范围,能给企业提供关于日常战术决策的技术支持。在当今的商业决策越来越要求实时的情况下,主动实时数据仓库可用于精确地回答用户提出的针对实时数据的问题。可针对长时间的大量数据作分析,来决定如何更好的服务于每一个客户,或者鉴别出潜在的、欺诈的或非法的行为。所有这些都要求把现有系统和应用运转在实时数据上。

建设数据仓库的最难的部分之一就是将源数据从不同的系统进行数据导入的抽取、转换、清洗和装载的 ETL 部分。传统的数据仓库采用 ETL 工具或开发自己的数据库脚本,这个处理过程显著的影响数据仓库的非响应期。在进行导入的过程中不允许用户访问数据仓库,一般在晚间进行数据载入的批处理操作。而主动实时数据仓库一般采用 EAI 等技术[9],在事务完成后即触发,对数据仓库进行载入或更新。

4 改进的主动实时数据仓库架构模型

4.1 基于 EAI 的实时/主动数据仓库模型不足

基于企业应用集成的实时数据仓库工作原理是:将实时数据从数据源系统中抽取出来,并将 ODS、数据仓库和数据集市整合到一个庞大的数据

仓库中。由于涉及到载入、更新、查询、分析等多种任务及融合了数据仓库与集市处理的多种技术，必将在体系结构、数据建模、数据库设计及存储方面存在着重大不足，主要体现在以下几个方面：

1. 服务器负载：由于融合了 ODS、数据仓库及数据集市，就需要处理各种实时数据（如事实表、维表及其索引）更新，同时还需要对历史数据进行 OLAP 和数据挖掘。这使得服务器超负载运行，将严重影响系统性能。

2. 触发机制：在事务处理中，需要大量的触发器来检查事务，同时需要较多的等待来保证更新的并发性和完整性，同样也严重降低了系统的性能。

3. 同步机制：在实时数据仓库中，消息和事件的出现具有异步性，这样的同步在实时条件下很难实现，为了尽可能达到‘准实时’的效果，将以提高系统的复杂度和牺牲系统性能为代价。

4. 数据库设计：在传统的数据仓库中，ODS 由于载入更新较为频繁且批量数据不大，一般采用较小的数据块。而数据仓库中由于存放庞大的事实表且较少进行更新，一般采用较大的数据块，这样可以提高性能及减少 I/O 访问次数。如果整合到一起，块大小的设定就会对数据库空间利用率及数据库查询效率产生严重的影响，无论块区大还是小，都将影响到系统运行效率。

5. 恢复机制：安全问题是我们建设和使用数据库要一直面对的问题。要确保在出现数据灾难时不丢失关键数据。由于实时数据载入是 24*7 无间断在运行，因此对灾难恢复提出了较高的要求，同时也增加了系统的难度和负担。

6. 数据验证：在大多数情况下，实时数据表的更新需要验证其数据的有效性与合法性。数据仓库发展到今天，我们经常会遇到事实表远远大于维度表的情形。比如，一家大型连锁超市的数据记录很容易达到上百万条。在如此大的维度结构情况下来验证每条实时记录，必然会造成资源的短缺与性能的下降。

4.2　改进的架构模型

主动实时数据仓库有着很广泛的市场需求，符合企业的信息化需求，应加以研究和利用。直接构建主动实时数据仓库还很困难，笔者认为应当结合传统数据仓库的一些优点加以改进。

在这个主动实时数据仓库架构中，我们利用 EAI 平台提供的实时数据交换的功能，利用建立在面向服务架构体系上的企业应用集成将实时数据从源数据系统中抽取出来，监控得到的实时数据将被转存到增强的 ODS 模块中，用来解决传统的企业数据仓库体系结构中数据仓库与数据源耦合程度高，数据源端的数据库系统负载繁重的问题。在这个增加的 ODS 模块中，不仅可以承担大量的报表工作和简单查询工作，而且还因为添加了实时组件和基于事件的数据可以进行一些战略查询用于企业决策和数据挖掘分析。

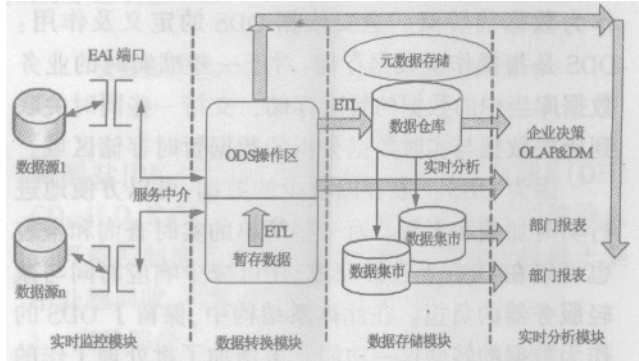

图 2　改进的主动实时数据仓库架构模型

结合文献[10]本文提出一种较为完善的主动实时数据仓库模型，如上图所示。其中主要的模块如下：

1.实时监控模块：体现系统实时性的关键模块，主要通过 EAI 平台提供的实时数据交换，通常以事件触发的形式来主动获取数据，即一旦有事务发生并产生操作型数据，该模块立即将新增的操作数据传送到 ODS 中。

2.数据集成模块：数据集成模块是该体系架构的核心模块，ODS 则是它功能的核心体现。它介于操作数据存储和数据仓库存储之间的一个过渡性存储区域，通常用于存放从数据源中抽取出的操作数据，并对操作数据进行整合、转换等清洗工作，因此可看作数据仓库的数据准备区。分析工作可在增强功能的 ODS 直接进行，这样不仅节省了数据进入数据仓库的传输时间，同时还省略了数据转换等复杂步骤。

3.数据存储模块：指常规意义的数据仓库系统，主要用于存放历史数据来进行常规分析，同时该模块也负责接收实时数据，并将其与已有历史数据进行归档保存。

4.实时分析模块：该模块主要包括各种分析工具和展现工具，如 OLAP 工具、数据挖掘工具等。通

过该模块，企业可对实时及历史数据进行战略分析。

整体的系统描述如下：通过 EAI 提供的实时监控模块来监控操作型事务中改变的数据。监控模块捕捉的数据传输到增强的 ODS 模块中。数据在 ODS中存放一定的时间，其作用可归结为两个方面：一是与数据仓库中的数据进行整合处理；二是进行时段性分析。最终，实时数据将被导入非实时部分中，执行传统数据仓库的功能，整合历史数据，用于战略查询。

5 评价与分析

在这个主动实时数据仓库体系中，增加了 ODS 作为数据的转储。主要依据 ODS 的定义及作用：ODS 是指操作型数据存储，对于一些准实时的业务数据库当中的数据的暂时存储，支持一些同时关联到历史数据与实时数据分析的数据暂时存储区域。

因为 ODS 一般存储较少的数据，可以方便地进行实时加载和更新。而一些简单的实时查询和报表也可以在 ODS 上直接完成，还可减少响应时间与减轻服务器的负担。在此体系结构中，保留了 ODS 的作为数据的转储这一功能，虽增加了批处理工作的复杂性及实时更新的时间，但由于添加了实时组件的获取也增加了 ODS 的功能，完全可弥补这一缺陷。而且与传统数据仓库结合更紧，同时也更方便进行 OLAP 分析及数据挖掘，对 BI 工具也提供了良好的接口。

前文所述的原实时数据仓库体系结构的弊端，如触发机制、同步机制及恢复机制的实施难度，在改进后的体系结构中难度也有所降低，同时服务器的负载也有所减轻。由于数据量相对较小，处理起来相对简单，易于设计和实现。

6 小结与展望

本文对主动实时数据仓库和传统的数据仓库进行了比较和分析，根据其差别与架构的不同，提出一种改进的主动实时数据仓库的体系结构并分析其主要模块的作用。利用现有的技术来实现新型数据仓库是有可能的，尽管构建路上还存在很多的技术细节和构建策略的问题。但是我们应看到实时数据仓库代替传统数据仓库的必然趋势。在主动实时数据仓库优势不断显现的今天，真正实现还需大家共同的努力。

参考文献：

[1]Inmon,W.H,《数据仓库》Building the Data Warehouse. 数据仓库 (3rd Edition)[M].王志海等译.北京：机械工业出版社，2003.3.

[2]Samuel S.Conn, OLTP and OLAP Data Integration: A Review of Feasible Implementation Methods and Architectures for Real Time Data Analysis [C].Regis University School for Professional Studies, 2005.

[3]Bruckner, R.M., Tjoa, AM.: Capturing Delays and Valid Times in Data Warehouses- Towards Timely Consistent Analyses [J] Journal of Intelligent Information System(JIIS), 2002.

[4]王霓虹，刘美玲. ODS 数据仓库新技术的研究与应用[J].信息技术. 2004,28: (11).

[5]唐小燕，李斌，许有志. 一种 Agent 协同规范及其应用研究[J]. 计算机应用研究. 2005, (1) : 82- 84.

[6]Raj Basu, Challenges of Real - Time Data Warehousing [EB/OL] http://www.dmreview.com,2003- 11- 11.

[7]张俊，张忠能. 实时数据仓库体系架构的研究[J].计算机工程，2004,(30) : 180- 182.

[8]王翀. 商业智能中实时数据仓库的数据挖掘研究[D]. 硕士,武汉大学,2004.5.

[9]Bjarne Berg, Real- Time Data Warehousing Merges with Operational Reporting: How Will You Manage? [EB/OL] http://www.dmreview.com 2006- 02.

[10]许有志，沈洁，唐小燕. 基于多 Agent 的主动数据仓库的研究[J].计算机工程与设计. 2005,26(4) : 947- 950.

An Analysis of a New Architecture of Active Real- time Data Warehouse

Zhang Lei & Wang Beizhan

(Xiamen University, Fujian 361005, P.R. China)

Abstract: The paper discusses the difference between traditional data warehouse and active real- time data warehouse, and presents a new architecture of active real- time data warehouse.

Key Words: traditional data warehouse; real- time data warehouse; active real- time; architecture