实时主动数据仓库的概念、问题及应用

宋国杰[1,2] 杨冬青[1] 林子雨[1] 唐世渭[2] 王腾蛟[1] 谢昆青[2]
[1](北京大学信息科学技术学院计算机系，北京 100871)
[2](北京大学视觉听觉信息处理国家重点实验室，北京 100871)
(gjsong@pku.edu.cn)

摘要 近年来，数据仓库技术在学术界和工业界都得到了广泛的关注。实时主动数据仓库（RTADW, Real Time Active Data Warehouse）是数据仓库技术发展的一个新的阶段，是数据库技术的一个新的研究领域，具有十分广阔的应用前景。介绍了实时主动数据仓库的概念和特点，探讨了实时主动数据仓库的研究问题，并列举了一些典型应用。

关键词 实时数据集成；主动决策；数据仓库

中图法分类号 TP301

Concept, Issues and Applications of Real Time Active Data Warehouse

SONG Guojie[1,2], Yang Dongqing[1], TANG Shiwei[2], WANG Tengjiao[1], XIE Kunqing[2]
[1] (Department of Compute Science, School of Electronics Engineering and Computer Science, Peking University, Beijing 100871)
[2] (National Laboratory on Machine Perception, Peking University, Beijing, 100871)

Abstract Recently, data warehouse technology has been paid much attention both in both academia and industries. Real time Active data warehouse (short for **RTADW**) is a new stage during the evolution of data warehouse and is a new research area of database technology and has a wide application future. The concepts and characteristics of RTADW are introduced, and the issues of RTADW are discussed. Some typical applications are also presented.

Keywords Real Time Data Integration; Active Decision; Data Warehouse

随着信息技术的发展，数据仓库技术得到了前所未有的广泛应用，产生了巨大的经济效益。在美国，30%到 40%的公司已经或正在建造数据仓库，其代表有 AT&T 公司、VERIZON 移动通信、沃尔玛百货公司等。据国际权威统计机构 IDC 对欧洲和北美 62 家采用数据仓库技术的企业的调查分析发现，这些企业的 3 年平均投资回报率为 401%，其中 25%的企业的投资回报率超过 600%。

近年来，我国大中型企业也逐步认识到利用数据仓库技术的重要性，并已开始建立自己的数据仓库系统，如中国移动、中国电信、中国联通、上海证券交易所和中国石油等。这些数据仓库系统已经开始在这些企业运营过程中发挥出显著的作用。例如，从 2001 年起，中国移动开始在全国范围内建设数据仓库系统，目前已建成规模庞大的分级式数据仓库，有数万用户在使用这种数据仓库系统，年访问量达数千万人次。仅一项"重入网分析"可以节约成本约数亿元。

但是，随着市场经济步伐的加快和市场竞争的日趋激烈，传统的数据仓库技术已经不能很好地满足当前企业发展和竞争的需要。传统数据仓库仅为企业高层决策者提供战略决策（Strategic Decision），服务于宏观决策和长远规划，如市场细分、产品管理等。然而，随着市场竞争的加剧，企业越来越希望数据仓库在支持战略决策的同时，也能够为市场一线人员提供实时的战术决策（Tactical Decision）服务，如实时营销、个性化服务等。这种既服务于战略决策又服务于战术

收稿日期：2007-07-05
基金项目：国家自然科学基金项目(60473051)；国家高技术研究发展计划 863(2006AA12Z217)；HP 中国实验室联合项目

决策的数据仓库称之为为实时主动数据仓库（RTADW, Real-time Active Data Warehouse）。根据 Gartner 的研究报告，RTADW 已成为数据仓库发展的必然选择，他将进一步提升企业的市场竞争能力。然而，当前对 RTADW 的研究尚不成熟，许多关键技术急需进行深入研究。

自 2002 年起，北京大学数据库研究室与中国移动集团公司在数据仓库的研究和建设方面开始了深入而密切的合作，在数据仓库和数据挖掘技术的研究开发和应用推广方面展开了大量卓有成效的工作，并于 2006 年 5 月在北京大学联合成立了“移动通信数据仓库联合实验室”。以移动通信领域为背景，在北京大学－惠普中国实验室联合项目的支持下，目前我们正在开展面向大规模海量实时主动数据仓库的研究工作。在理论研究和系统开发方面已经取得了丰富的成果。

本文将分别介绍实时主动数据仓库的概念、特点、需要研究的问题以及一些典型的应用。

1. 实时主动数据仓库

实时主动数据仓库的概念

Michaem Haisten 提出了实时主动数据仓库的概念：**RTADW 是一个关系型环境的数据仓库，支持数据的实时更新，快速的响应时间，基于钻取的聚集数据查询能力和动态的交互能力，用于支持不断变化的商业需求**[7]。与传统数据仓库系统相比，实时主动数据仓库系统有许多独有的特点（参见表 1）。

表 1. 实时主动数据仓库与传统数据仓库的比较

传统数据仓库	实时主动数据仓库
仅支持战略决策	支持战略决策和战术决策
实时性要求不高	要求结果实时返回
数据传输是单向的	数据传输是双向的
返回很难测量的指标	返回日常运营的指标
以天、周以及月为周期获取数据，并做预先聚合计算	只包含明细数据，可以以分钟为周期获取明细数据
中等规模用户数	多用户的并发访问
仅得到高度限制的报表，适用预处理的聚合表或数据集市	灵活的即席查询、数据挖掘
高级用户、分析员和内部用户	操作雇员、呼叫中心和外部用户

在 RTADW 发展的过程中，ODS(Operational Data Store)是一个过渡阶段。ODS 分三类：（1）实时 ODS。它通过消息中间件实施数据的同步转换和刷新，但是业务系统不能太多，转换数据量不能太大；（2）准实时 ODS。它基本实现数据同步，以 1～2 小时为周期，系统负担较小，具有较好的灵活性；（3）传统 ODS，代价最小，目前在传统数据仓库中常见。

实时主动数据仓库的特点和挑战

实时数据的连续集成

为支持实时的战术决策服务，源系统（或称生产系统）产生的实时数据必须在最小化对源系统入侵程度，并保证实时数据一致性和完整性的情况下，被实时高效地集成到数据仓库中。挑战问题是：(1)在保证源系统性能不降低的情况下，对实时数据在源系统的任何变化进行实时的捕获；(2)保证被连续分发数据间次序的一致性和自身的完整性；(3)在保证数据质量要求的前提下，完成实时、高效的数据加载。

实时数据和历史数据的组织与管理

提供 RTADW 中的实时数据和历史数据的有效的组织与管理策略，使之高效地工作在一种混合的工作负载环境（战略决策和战术决策）中。所要研究的挑战问题：(1)对实时数据和历史数据（指传统数据仓库中存储的数据）进行统一建模，从而对外提供统一的访问视图；(2)研究对实时数据查询所产生的“查询冲突”和“查询不一致性”问题，保证查询处理过程的无阻塞性的和查询结果的一致性；(3)研究实时数据和历史数据的及时信息合并技术，对提交的 RTADW 的任何查询提供“透明”的一体化服务（4）对负载的管理，使得 RTADW 系统高效的运行。

主动的服务决策机制

研究 RTADW 的主动决策服务机制，提供对实时事件进行主动分析和处理的能力。挑战问题包括：(1)研究实时事件的主动捕获机制，具备对外界请求的实时响应能力；(2)研究分析决策过程的自动执行机制，使 RTADW 系统拥有主动服务的能力。

实时主动数据仓库的性能评价

1. 数据的新颖性

实时数据必须被及时的加载到系统当中，从而支持战术查询分析。

2. 时间的一致性

在连续数据集成的环境中，会出现数据时序的不一致性现象，需建立时序模型进行时间一致性管理。

3. 查询结果的一致性

由于数据的动态到达和查询的持续性，会出现同一查询请求但不同时刻不同查询状态的影响，出现查询结果的不一致性，需要查询过程的一致性管理。

4. 主动决策的及时性

RTADW 系统能够实时捕获各种决策规则限定的动作，并做出实时的反映。

5. 可扩展性

用户数目和性能需求随着 RTADW 系统分析应

用的部署增加而增加。

2. 实时主动数据仓库的研究问题

在介绍研究问题之前，我们首先介绍一个RTADW系统的参考结构，如图1所示。

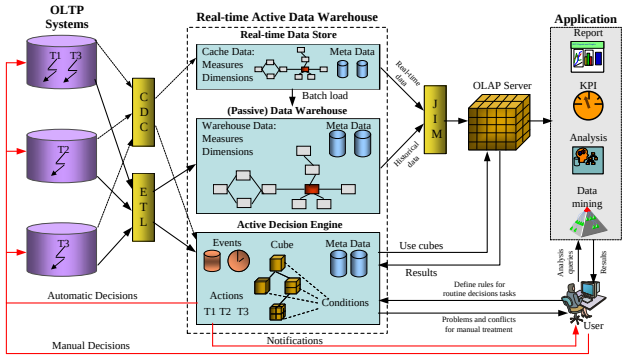

图1. 实时数据仓库的参考结构

如图所示，一个RTADW系统主要包含四个组成部分：数据源、数据抽取、数据仓库、主动决策部分和前端应用。

数据源除了包含传统的静态部分之外，还包含实时的数据源部分（如数据流等）；数据抽取部分包含传统的ETL抽取和实时数据抽取两部分；数据仓库除了存储传统的静态数据之外，还存储实时的数据部分，以及他们之间的周期性的转换和数据的实时合并；主动决策部分主要基于触发器的基础上，利用主动分析规则完成主动的决策分析；前端的展现除了传统展示方法外，还包括一些实时的监控部件（如Dashboard等）。

数据集成研究问题

RTADW要集成的数据包括实时数据和历史数据两部分。历史数据采用传统的批处理方法进行集成，而实时数据部分则需要进行实时的连续集成。重点研究问题包括：

实时数据的主动变化捕捉

在传统数据仓库系统中，由源系统按预先约定的加载时间和数据格式，定期把需要抽取的数据放到预先约定的接口中，然后由ETL引擎把这部分数据加载到数据仓库。但是，对RTADW而言，要求实时数据一旦由源系统产生就立即加载到数据仓库中，以便支持实时战术分析的需要。因此，RTADW系统需要能够对新产生的实时数据变化（插入、更新等）进行实时捕获，从而及时进行数据加载。

要研究快速数据的变化捕捉方法，实现对源系统中的数据变化进行有选择性的定位和捕捉（即仅捕获实时部分的数据变化），满足零延迟的要求，最小化对源系统的入侵程度（即对源系统性能的影响），降低源系统的负载，确保源系统性能不下降、不当机。

支持数据一致性和完整性的实时数据分发

数据分发是指数据从源系统到数据仓库的传播过程。在传统数据仓库中，数据分发采用批量拷贝的方式，数据间的时间依赖性和事务依赖性在数据批量转移的过程中不受影响，可以保持数据的一致性和完整性。但在RTADW中，捕捉到的每个数据变化都是以消息的形式进行分发，同一事务中包含多个数据变化，也就包含了多条消息，这些消息在网络中进行独立传输。因而，如何保证消息在传输过程中的完整性，以及如何保持多个消息之间的正确顺序，从而有效地维护数据的事务一致性和不同事务间的依赖性。

研究高效的数据分发机制，使得每个捕捉到的数据变化被封装成消息后放入消息队列，由消息队列完成数据的分发，保证消息传输的一致性和完整性，同时有效地维护数据的事务依赖性和时间依赖性。

实时、高效的连续数据加载

在RTADW中，接收到的消息中是未经处理的数据，如果对这些数据进行复杂地清洗和转换操作，将无法满足外部查询对数据实时性的要求；反之，所包含的脏数据会严重影响数据的质量。如何在保证数据质量的前提下实现实时、高效的数据加载是一个需要深入研究的问题。

研究连续高效的数据加载技术实现实时、高效的连续数据加载。实现对数据的清洗和转换过程所包含的内部子环节进行合理有效地组织，从而提高数据处理的速率和并发度。同时根据用户对数据质量的不同需求，对即时加载的数据进行区别对待，合理分配系统资源，提高数据加载性能。

数据的组织与管理

研究RTADW系统中实时数据与历史数据的数据特性，建立有效的数据存储、组织与访问策略，为高效的战略决策和战术决策的执行提供数据平台支撑。

RTADW中的数据建模

传统数据仓库中的数据一般在空闲时（如夜间）以批处理的方式进行更新。由于更新时不对外提供数据查询服务，因而对更新的代价不做过高要求。但是，由于RTADW中的实时数据是以7*24的工作方式对外提供服务，而且要求数据的实时更新和查询结果的实时反馈，所以传统静态数据的组织与管理方式不适合于实时数据。

研究把实时数据和历史数据的有效建模问题，使得对于查询工具而言，只有一个统一的逻辑视图，避

免查询工具和终端用户进行多表连接操作的问题。

实时数据的查询一致性维护

在 RTADW 环境中，数据仓库中数据是实时更新、不断变化的。在这种“动态”的数据环境中使用 OLAP 分析和查询工具，会使查询所涉及的数据在读取过程中不断发生变化，从而导致查询结果的不一致性。

研究在实时数据环境中的查询一致性问题，防止数据在查询过程中被修改，从而保证查询的一致性。同时保证以后到达的查询得到的是更新以后的数据，保证了数据的实时性。

实时数据的查询冲突解决

在 RTADW 环境下，由于数据的查询和更新是同时进行，会导致在某个时刻，对于事实表中的某些记录，查询操作和更新操作会发生读写冲突。当源系统的数据变化过于频繁，数据仓库中的查询数量比较多时，这种冲突将更加突显，甚至可能使系统发生阻塞，无法对外提供服务。

研究实时数据的查询冲突问题，利用不同用户查询对数据实时性的不同要求，有效分流不同类型的查询负载，防止系统因查询冲突而发生阻塞，同时又能满足不同类型查询的需求。

实时数据与历史数据的“无缝”集成

在 RTADW 环境下，为了最大程度地减小查询冲突给系统带来的负面影响，保证数据仓库正常高效地运行，实时数据与历史数据通常分开存储。为了最小化对查询工具的影响，不需要查询工具了解获取不同类型数据的方法，而是一旦提出查询请求，就可以得到“无缝集成”后的数据。

研究高效的集成技术，实现实时数据与历史数据的“无缝”集成。能够自动分析查询语句，从而确定数据需求，并从 RTADW 的不同部分提取所需的数据，合并后供查询工具使用。可以自动分析所需数据中实时部分和历史部分的比例，从而更好地选择数据的迁移策略，减少数据传输，改善服务性能。

主动决策服务

研究 RTADW 的主动决策机制，从而支持对实时事件的主动探查，并根据事件的特征进行处理判断，从而触发相应的分析规则。

RTADW 中事件的主动探查

在 RTADW 系统中，为了支持实时主动的决策分析，从而满足系统实时响应的需求，就需要系统具备对各类事件的主动探查机制，从而及时能够实时发现各类异常事件，并进行相应的处理。

研究事件主动探查机制，与 RTADW 的应用需求相结合，提高事件的探查速度，保证 RTADW 的实时性。研究将事件的组织和存储方式，使得在探查到事件后，能够迅速将事件与相应触发的分析规则相匹配，避免了简单查找所产生的巨大代价，进一步提高事件的匹配效率。

支持主动决策的分析规则技术

在 RTADW 的应用中，常常有大量的事件同时发生，导致很多分析规则同时触发，并且各个规则间通常有一定的联系，一些规则的发生会导致其它规则的触发。这就使规则的并发控制以及匹配的效率成为影响 RTADW 性能的重要问题。目前，已有的方法并不能完全解决大量事件并发时带来的效率问题，已有系统不能满足大量用户并发情况下的性能需求。

研究高效得分析规则组织方法，使分析规则之间的关系明确，规则的匹配及规则触发因果关系的查找更加迅速，提高分析规则匹配的效率。研究有效的分析规则所产生的操作组织方式，使得分析规则触发时能够迅速、主动地找到与之对应的具体操作，从而能够迅速对不同的源系统进行相应的操作。

3. 实时主动数据仓库的典型应用

下面列举一些典型的实时主动数据仓库应用.

在移动通信领域的应用

实时监控移动呼叫数据，防止欺诈行为发生。据国际数据公司 IDC 统计，每年全球电信领域因欺诈而造成的损失占电信服务总收入的 5%-10%之多，对电信运营商的信誉和正常业务运营产生了严重的冲击。传统数据仓库技术因其数据延迟过大，不能实时捕获欺诈信息，所以不能很好地解决这一难题。利用 RTADW 技术，把当前的实时呼叫信息和数据仓库中的历史行为信息结合，借助通讯企业提供的企业间共享的欺诈人群信息，可以有效地判定当前客户是否具有发生欺诈行为的可能性，从而进行实时地预防和监控，减少企业因欺诈而造成的损失。

在电子商务领域的应用

根据用户提交信息，实时给客户灵活的定价和折扣。借助于网络技术而蓬勃发展起来的电子商务应用已日趋普及。但是，目前的电子商务一般不具有个性化服务的能力，只能在既定规则下开展不具有针对性的服务，不利于提升客户服务的能力。RTADW 技术可以很好地解决这一难题。对客户提交的实时购物订单，可以根据客户的当前行为信息，借助于在数据仓库中存储的该客户的历史消费信息来判断客户的价值，进行实时又个性

化的定价和折扣，从而提升客户满意度，增强企业的竞争力。

4. 相关研究

NCR 给出了主动数据仓库[1]的概念，但其本质上也是 ODS 和传统数据仓库的结合。其它一些国际数据库研究机构（如 IBM，Oracle，Sybase）也纷纷提出了自己的解决方案，共同点是都采用了 ODS 方案来存储实时数据，用传统数据仓库存放历史数据，从而是 ODS 之上解决对实时数据查询的问题。但是，采用 ODS 解决方案面临的最大问题就是所有的数据必须进行实时的抽取，否则无法满足实时性的需求。但是，根据国际权威机构 Garter 的研究报告指出，真实业务对实时数据的需求量仅占所有抽取数据量的25%左右。因此，对所有的数据都采用这种代价高昂的实时抽取处理方式，必然带来具体的资源（时间、空间等）的浪费

RTADW 的实时数据连续集成主要有如下方法：（1）脚本方法。该方法使用灵活且比较经济，很容易着手开发和进行修改，而且几乎任何操作系统和绝大部分 DBMSs 都可以使用脚本。但是，该方法的实施耗费开发者的大量的时间和精力，而且不易于管理和操作以及不能满足服务水平协议；（2）ETL 方法。该方法是实现大规模数据初步加载的理想解决方案，提供了高级的转换能力。但是该方法通常都是在“维护时间窗口”进行，在 ETL 任务执行期间，数据源默认不会发生变化，从而不能满足实时数据集成；（3）EAI 方法[5]。该方法与 ETL 解决方案并存，并增强了 ETL 的功能，能够支持在源系统和目标系统之间进行连续的数据分发，并提供高级的工作流支持和基本的数据转换。但是该方法受到数据量的限制，不适合数据量较大的环境。

当前，在实时数据和历史数据的组织与管理研究方面，主要有以下一些学术观点[6]：（1）无实时数据存储：该方法把从源系统产生连续加载到数据仓库，可以直接在数据库仓库事实表中插入或更新数据，也可以把数据插入到实时分区当中的单独的事实表中。但是，该方法的缺点是可扩展性不好 ，复杂查询和连续插入及更新混在一起进行会严重影响数据库的性能；（2）阶段存储表存储实时数据：该方法把数据连续地注入到阶段存储表，其结构和数据仓库表的结构相同，其内容会和事实表周期性地进行交换，采用视图集成完成实时数据与历史数据集成。但是，该方法的缺点是在处理数据交换的时候，必须暂时停止对外提供实时查询服务；(3)实时数据缓存：该方法可以彻底避免对数据仓库性能的影响，不用对现有的数据仓库做出修改，可以是另一个专用的数据库服务器，也可以是一个大的数据库系统的单独的实例，把所有那些需要实时数据的查询定向到实时数据缓存，或者把某个查询所需要的实时数据临时地无缝隙地整合到传统的数据仓库中。该方法的缺点要安装和维护一个额外的单独的数据库。

主动规则(Active Rule)[2,3]作为实现主动决策的手段已经被广泛接受。在处理一系列复杂的任务和对数据的自动管理（包括完整性约束检验、转换、安全等）过程中，使用户的参与度降低到最小。主动的信息管理具有基于“事件-条件-动作”ECA（Event-Condition-Action）的规则处理特性，复杂的商业决策构造和处理不需要深入到应用程序和底层数据库。主动数据仓库(ADW)中使用 ECA 规则或者其它事件驱动机制，是为了在传统数据仓库环境中自动执行日常决策任务。Thalhammer[4]采用 ECA 规则来模仿分析人员的工作，故称之为分析规则(Analysis Rules)。它结合了 ADB、传统数据仓库和 OLAP 良好的决策标准从而满足决策过程的自动化。但是，它的数据集成过程是基于传统的批处理方式工作的，这与实时的数据集成不符合，会在迟到的数据和基于不完整信息的决策制定中产生问题。

参 考 文 献

[1] TERADATA CORP., *Teradata Online Document*, 2002.

[2] ABITEBOUL, S.;CLUET,S.; MIGENT,L.; AMANN, B.;MILO,T.; EYAL,A,. *Active views for electronic commerce*. In VLDB’99, Proceedings. of 25th International Conference on Very Large Data Bases, September 7-10,. 1999, Edinburgh, Scotland, UK, pages 138–149. Morgan Kaufmann, 1999.

[3] James Bailey, George Papamarkos, Alexandra Poulovassilis, Peter T. Wood: *An Event-Condition-Action Language for XML*. Web Dynamics 2004: 223-248. 2003.

[4] Thalhammer T,Schrefl M. *Realizing Active Data Warehouses With Off-the-shelf Database Technology*[J].Softw.Pract.Expert,2002,32:1193～1222.

[5] Colin White. *Data Integration: Using ETL, EAI, and EII Tools to Create an Integrated Enterprise*. TDWI Report,November,2005.

[6] Langseth, J., Real-Time Data Warehousing: Challenges and Solutions, DSSResources.COM, 02/08/2004.

[7] Xiaofeng He, Gang Wang, Jiancang Zhao. *Research on the SCADA /EMS System Data Warehouse Technology*. 2005 IEEE/PES Transmission and Distribution Conference & Exhibition: Asia and Pacific Dalian, China.