计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/CN 11-1777/TP
44(Suppl.): 447～451, 2007

实时主动数据仓库中的变化数据捕捉研究综述

林子雨　杨冬青　宋国杰　王腾蛟
(北京大学信息科学技术学院　北京　100871)
(cainiu@263.net)

Change Data Capture in Real-Time Active Data Warehouses: A Survey

Lin Ziyu, Yang Dongqing, Song Guojie, and Wang Tengjiao
(*School of Electronics Engineering and Computer Science, Peking University, Beijing* 100871)

Abstract Real-time active data warehouse is the most recent stage in the evolution history of data warehouses. It supports both strategic decision and tactic decision, which will bring great benefits to organizations. There are two types of data existing in real-time active data warehouses, i.e., real-time data and non-real-time data. Accordingly, change data capture methods are classified into two kinds, including those supporting real-time change data capture and those not supporting real-time change data capture. Based on extensive research work in this field, those change data capture methods are systematically discussed, which may meet the requirements in real-time active data warehouses.

Key words real-time active data warehouse; change data capture; non-real-time data

摘　要 实时主动数据仓库是数据仓库的最新发展阶段和未来发展趋势,它为企业提供了对战略决策和战术决策的双重支持.实时主动数据仓库中包含两类数据,即实时数据和非实时数据,相应地,需要两种不同类型的变化数据捕捉方法,即支持实时变化数据捕捉的方法和普通的(不支持实时的)变化数据捕捉方法.结合在该领域的研究经验,对实时主动数据仓库中可以使用的多种变化数据捕捉方法进行了系统地论述,并比较各种方法的应用条件、优点、缺点和适用场合.

关键词 实时主动数据仓库;变化数据捕捉;非实时数据

中图法分类号 TP301

实时主动数据仓库是数据仓库的最新发展阶段和未来发展趋势[1].由于具备实时性和主动性等特征,实时主动数据仓库可以为企业提供战略决策和战术决策的双重支持.文献[2-5]包含了该领域目前有代表性的研究成果.文献[6]阐述了该领域面临的挑战及其解决方案,数据集成就是其中之一,而变化数据捕捉技术是实现数据集成的关键.

本文在文献[7]的基础上,结合我们在实时主动数据仓库领域的研究经验,对该领域使用的变化数据捕捉方法进行归类介绍,并比较各种方法的应用条件、优点、缺点和适用场合.其中一些方法已经在传统的数据仓库中得到广泛应用,比如复制、基于表记录方法和数据库快照比较,而另一些方法则是近些年来才出现的,并且是专门为了满足实时应用需求的,比如基于日志的变化数据捕捉.

1　变化数据捕捉方法

实时主动数据仓库中包含两类数据,即实时数据和非实时数据[6].为了满足用户对这两类数据的

收稿日期:2007-07-05
基金项目:国家自然科学基金项目(60473015);国家"八六三"高技术研究发展计划基金项目(2006AA12Z217);HP 中国实验室联合项目

需求,需要采用不同的变化数据捕捉方法,从而把源系统中发生的变化反映到数据仓库中. 由此,变化数据捕捉方法也可划分为两大类:实时变化数据捕捉类和非实时变化数据捕捉类[5].

在各种不同的变化数据捕捉方法中,属于实时变化数据捕捉类的方法可以支持对源系统中发生的变化进行实时地捕捉,而属于非实时变化数据捕捉类的方法,则只能用于允许存在延迟的变化捕捉.

表1给出了可用于实时主动数据仓库的各种变化数据捕捉方法,从中可以看出各种方法是否支持实时的变化数据捕捉.

表1 变化数据捕捉方法

变化数据捕捉方法		是否支持实时变化数据捕捉
基于表记录的方法	记录所有变化	否
	记录最后变化	否
	混合式	否
复制	事务复制	是
	快照复制	否
触发器		是
数据库快照比较	基于批量拷贝	否
	基于快照复制	否
基于日志的变化数据捕捉		是
刷新表		否

2 基于表记录的方法

基于表记录的方法很早就已经应用于数据仓库的变化数据捕捉[8]. 顾名思义,该方法需要在表中记录系统中发生的数据变化. 根据记录变化数据的方式的不同,该方法又可进一步划分为3种[7],即记录所有变化(方法1)、记录最后变化(方法2)和混合式(方法3).

2.1 记录所有变化和记录最后变化

记录所有变化:需要创建一张与原表关联的新表,用这张新表记录原表发生的所有变化. 这张新表记录的信息包括:1) 操作类型. 针对某条记录所执行的操作的类型,包括插入、删除和修改. 2) 操作时间. 操作发生的时间,如果是更新操作,则需要记录对哪个字段进行了更新,以及该字段在更新前后的值.

记录最后变化:需要在原表中创建一个新的列,来记录发生的变化和变化发生的时间.

表2给出了两种基于表记录的变化数据捕捉方法的比较,比较内容包括应用条件、优点、缺点和适用场合.

表2 两种基于表记录的变化数据捕捉方法的比较

比较内容	方法1:记录所有变化	方法2:记录最后变化
应用条件	① 数据源中的数据应该存储在一个DBMS中; ② 该DBMS应该支持触发器机制; ③ 应该允许使用触发器.	① 数据源中的数据应该存储在一个DBMS中; ② 该DBMS应该支持触发器机制; ③ 应该允许使用触发器.
优点	① 保证存储原表中发生的所有变化; ② 加快了数据抽取速度.	① 和方法1相比,减小了对操作型系统的负担; ② 和方法1相比,减小了存储空间开销; ③ 加快了数据抽取速度.
缺点	① 对操作型系统带来额外的负担,因为需要对每个插入、更新和删除操作都进行记录; ② 增加了存储开销,因为每次数据抽取完成时,都需要把这些已经被抽取过的数据从日志表中删除.	① 不会记录原表中发生的所有变化,而只记录最后一个变化; ② 创建新的字段增加了存储开销; ③ 删除操作必须是逻辑的; ④ 当每次抽取操作完成以后,需要把那些标记为逻辑删除的记录从原表中物理删除.
适用场合	有必要对数据源中发生的所有变化进行记录	没有必要对数据源中发生的所有变化都进行记录,只需要记录最后一次变化

2.2 混合式

在原表中创建新的字段,对于插入操作和删除操作进行记录;对于更新操作,需要创建一张新表,记录所有变化,包括:1) 更新执行的时间;2) 哪个字段被更新;3) 更新的字段在更新前后的值.

使用该方法时,需要满足以下几个条件:1) 数据源中的数据应该存储在一个DBMS中;2) 该DBMS应该支持触发器机制;3) 应该允许使用触发器.

该方法的优点是:1) 保证针对原表的所有更新操作都被记录;2) 加快了数据抽取速度;3) 和方法1相比,减小了对操作型系统的负载;4) 和方法1相比,减小了存储空间开销.

该方法的缺点是:1) 在更新操作时,具有和方法1同样的缺点;2) 在插入和删除操作时,具有同方法2同样的缺点.

当有必要对所有的更新操作进行记录,并且只需要对最后一次插入和删除操作进行记录时,可以使用该方法.

3 复　制

复制是在数据库之间对数据和数据库对象进行复制和分发,并进行同步以确保其一致性的一组技术. 复制的体系结构(如图 1 所示)包括 3 个组成部分:发布服务器、分发服务器和订阅服务器[9].

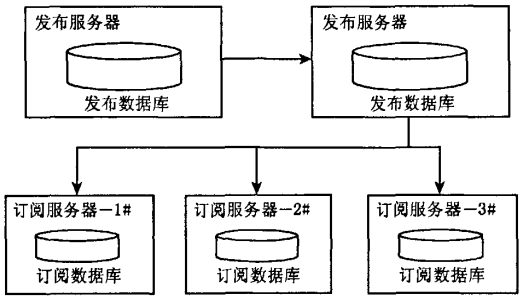

图 1　复制的体系结构

复制包括 3 种类型:事务复制、快照复制和合并复制. 在实时主动数据仓库中,可以使用事务复制来进行实时的变化捕捉,使用快照复制来进行非实时的变化捕捉. 图 2 显示了复制在实时主动数据仓库中的应用.

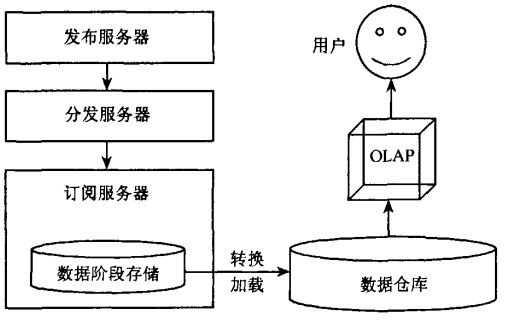

图 2　复制在实时主动数据仓库中的应用

使用该方法时,需要满足的条件是:在数据源和数据目标的 DBMS 必须具备必要的功能来对复制过程进行控制.

该方法的优点是:1) 可在多个站点保存相同数据的副本;2) 从大量读取数据的应用程序(如数据库仓库和数据集市)中分离 OLTP 应用程序;3) 允许更大的独立性;4) 加速数据抽取过程,因为整个工作都是由 DBMS 自己来完成的.

该方法的缺点是:1) 数据源和数据目标的 DBMS 必须相同,如果不相同,就必须有支持异构复制的第三方工具;2) 根据 DBMS 实施数据复制的具体方式的不同,有可能无法获得数据源中发生的所有变化;比如,使用事务复制可以获得数据源中发生的所有变化,而使用快照复制则不能达到这个目的;3) 需要 DBA 控制和管理在复制过程中可能出现的冲突.

4 触发器

可以通过在数据源中设置触发器,来对数据源中发生的变化进行实时地捕捉[7].

使用该方法时,需要满足以下几个条件:1) 数据源和数据目标的 DBMS 必须有针对插入、更新和删除操作的触发器;2) 数据源和数据目标中的 DBMS 必须相同;3) 数据目标的 DBMS 应该被允许通过触发器的方式进行数据的插入、更新和删除操作.

该方法的优点是,加速数据抽取过程,因为整个工作都是由 DBMS 自己来完成的.

该方法的缺点是:1) 对操作型系统有比较大的负担;2) 数据源和数据目标的 DBMS 必须相同;3) 数据源和数据目标的 DBMS 必须处于工作状态,否则数据源中的变化数据无法更新到数据目标的 DBMS 中;4) 为了获得数据源中发生的所有变化,需要在数据目标的表中采取方法 1 或方法 3 作为补充.

5 基于日志的变化数据捕捉

基于日志的变化数据捕捉是当前比较流行的变化数据捕捉方法,可以满足实时主动数据仓库对实时数据的需求,同时不会增加对源系统的负担[5]. 如图 3 所示,该方法以 DBMS 的事务日志为基础,并对其进行实时监控,一旦源系统发生数据变化,就进行实时捕捉,并根据预先设置的过滤规则,丢弃那些不需要的数据,只对需要实时集成的数据进行分发.

使用该方法时需要满足以下几个条件:1) 数据源中的数据必须存储在 DBMS 中;2) 该 DBMS 必须具备事务日志控制机制;3) DBMS 事务日志中的记录必须能够被正确理解.

基于日志的变化数据捕捉需要访问 DBMS 的事务日志. 有些 DBMS 事务日志细节可以通过该 DBMS 的产品文档来获得,当 DBMS 产品没有提供事务日志(REDO LOG)的细节时,我们就无法理解事务日志的内容,这时,就需要采取其他措施来对事务日志进行解析. 例如,由于 Oracle 公司没有公开 REDO LOG 的技术细节,这就必须要对 REDO LOG 二进制文件进行分析. 文献[10]给出了针对 REDO LOG 二进制文件的分析,给出了 REDO LOG 文件头、文件体块结构,并给出了 REDO LOG 文件的 C 语言描述以及 REDO LOG 事务控制机制,为捕获对数据库的数据更改和操作,实现基于 REDO LOG 的变化数据捕捉奠定了坚实的基础.

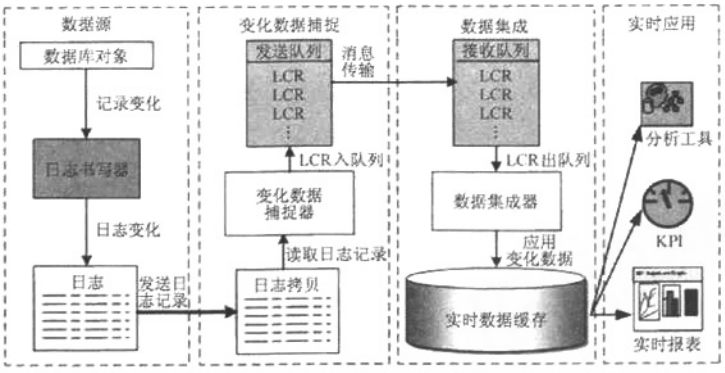

图 3 基于日志的变化数据捕捉

基于日志的变化数据捕捉的优点是:1) 对操作型系统没有额外负担;2) 减小了存储空间和处理的开销;3) 加速了数据抽取.

基于日志的变化数据捕捉的缺点是:1) 为了防止事务日志的丢失,DBA 必须控制事务日志记录区域的尺寸;2) 日志被妥善备份以前,由于物理或是人为失误,可能造成日志的丢失.

当有必要对数据源中发生的所有变化进行记录,并且愿意承担 DBMS 事务日志丢失的风险时,可以使用该方法.

目前已经有一些比较成熟的、基于日志的变化数据捕捉产品,比如 Oracle Streams[11], Informatica 公司的 PowerCenter 8 Real Time Option, DataMirror 公司的 Transformation Server5.3 和 Attunity 公司的 Attunity Integration Suite 等等.

6 数据库快照比较

在数据库快照比较方法中,需要拷贝一份原表到工作区域,然后把该表命名为:"〈name of table〉_current";当进行一次新的数据抽取时,刚才的表就会被赋予一个新的名字:"〈name of table〉_previous",本次新拷贝过来的表将被命名为:"〈name of table〉_current";对这两张表进行比较,从而确定变化的内容[7].

使用该方法时,需要满足两个条件:1) 数据源中的数据必须存储在 DBMS 中;2) 需要工作区域有比较大的存储空间来存放用于比较的 current 表和 previous 表.

该方法的缺点是:1) 浪费处理和记录的时间,因为需要把数据从数据源拷贝到工作区域;2) 浪费处理时间,因为需要对两张表进行比较从而找到变化的部分;3) 需要大量的存储空间,因需要在工作区域存储两张表:current 表和 previous 表.

可以使用该方法的情形包括:1) 数据源的 DBMS 不支持触发器机制;2) DBMS 的触发器机制被禁止使用;3) DBMS 没有事务日志记录;4) 无法理解 DBMS 的事务日志记录.

拷贝一份原表到工作区域可以采用不同的方式实现,根据实现方式的不同,数据库快照比较又包含两类,即基于快照复制的数据库快照比较和基于批量拷贝的数据库快照比较. 表 3 对两种不同的数据库快照比较方法的进行了对比.

表 3 两种不同的数据库快照比较方法的对比

比较内容	基于快照复制	基于批量拷贝
应用条件	在数据源和工作区域的 DBMS 必须具备必要的功能来对复制过程进行控制.	在工作区域的 DBMS 必须可以执行批量拷贝方法.
优点	加速数据抽取过程,因为整个工作都是由 DBMS 自己来完成的.	① 对操作型系统的日常操作没有额外负担,因为所有活动都是在工作区域进行; ② 加速数据抽取,因为在拷贝数据时不需要任何约束和控制.
缺点	① 数据源和工作区域的 DBMS 必须相同,如果不相同,就必须有支持异构复制的第三方工具; ② 无法获得数据源中发生的所有变化; ③ DBA 要控制和管理在复制过程中可能出现的冲突.	没有办法记录数据源中发生的所有变化,只能得到在数据抽取时刻的变化.

7 刷新表

在该方法中,需要把事实表和维表中的所有数据删除掉,重新开始全部的数据抽取和加载过程[7].

该方法的优点是,不需要在数据源执行任何数据变化捕捉工作.

该方法的缺点是:1) 每次新的抽取、转换和加载过程都需要耗费大量的处理时间,因为需要全部重新来做;2) 无法记录发生的变化;3) 丢失了数据仓库的历史信息.

在以下情况下可以考虑使用该方法:1) 没有必要对数据源中发生的所有变化进行记录;2) 没有必要保留数据仓库的历史信息;3) 没有时间限制,因为重新创建整个数据仓库很耗费时间.

在实时主动数据仓库中,通常只有在重新建设数据仓库时才使用刷新表的方法.

8 结束语

实时主动数据仓库是数据仓库的最新发展阶段,也是数据仓库的未来发展方向.本文介绍了实时主动数据仓库环境下的多种变化数据捕捉方法,并对这些方法进行了归类,比较了不同方法的应用条件、优点、缺点和适用场合.出于实时主动数据仓库建设的实际需要,可以酌情选择不同的变化数据捕捉方法,选择的标准应该着重参考以下几个方面的因素:数据质量、频率、可接受的延迟、转换需求和处理开销.

随着实时主动数据仓库的应用和推广(比如在移动通信和金融保险领域),其巨大的商业价值将不断得到体现.我们会在该领域继续开展更多的研究,努力推动实时主动数据仓库的发展.

参 考 文 献

[1] S Brobst, J Rarey. The five stages of an active data warehouse evolution. Teradata Magazine, 2001, 3(1): 38-44

[2] Michael Schrefl, Thomas Thalhammer. On making data warehouses active. DaWaK2000, London, UK, 2000

[3] T Thalhammer, M Schrefl. Mohania active data warehouses complementing OLAP with analysis rules. Data & Knowledge Engineering, 2001, 39(3): 241-269

[4] M N Tho, A M Tjoa. Zero-latency data warehousing for heterogeneous data sources and continuous data streams. iiWAS2003, Jakarta, Indonesia, 2003

[5] I Ankorion. Change data capture-efficient ETL for real-time BI. DM Review Magazine, 2005, 16(1): 23-27

[6] J Langseth. Real-time data warehousing: Challenges and solutions. http://dssresources.com/papers/features/langseth/langseth02082004.html, 2004

[7] Rosana L Brito Assayag Rocha, Leonardo Figueiredo Cardoso, Jano Moreira de Souza. Performance tests in data warehousing ETLM process for detection of changes in data origin. DaWaK2003, Prague, Czech Repblic, 2003

[8] A Silberschatz, F K Henry, S Sudarshan. Database System Concepts. 3rd ed. New York: McGraw-Hill, 1997

[9] 杨正洪. 中文 SQL Server 2000 关系数据库系统管理和开发指南. 北京: 机械工业出版社, 2001

[10] 李伟明. ORACLE REDO LOG 文件分析及 C 语言描述. 小型微型计算机系统, 2003, 24(7): 1243-1246

[11] Oracle Streams Replication Administrator's Guide. 10 Release 2 (10.2). http://www.oracle.com/pls/db102/to_doc?partno=b14228, 2006

林子雨 男,1978 年生,博士研究生,主要研究方向为数据库、实时主动数据仓库、数据挖掘.

杨冬青 女,1945 年生,教授,博士生导师,主要研究方向为数据库、数据仓库、Web 数据集成、移动数据挖掘.

宋国杰 男,1975 年生,讲师,主要研究方向为数据仓库、数据挖掘.

王腾蛟 男,1973 年生,副教授,主要研究方向为数据库、数据仓库、Web 数据集成、数据挖掘.

实时主动数据仓库中的变化数据捕捉研究综述

作者： 林子雨， 杨冬青， 宋国杰， 王腾蛟， Lin Ziyu， Yang Dongqing， Song Guojie， Wang Tengjiao

作者单位： 北京大学信息科学技术学院,北京,100871

刊名： 计算机研究与发展 ISTIC EI PKU

英文刊名： JOURNAL OF COMPUTER RESEARCH AND DEVELOPMENT

年，卷(期)： 2007，44(z3)

参考文献(11条)

1. S Brobst. J Rarey The five stages of an active data warehouse evolution 2001(01)

2. Michael Schrefl. Thomas Thalhammer On making data warehouses active 2000

3. T Thalhammer. M Schrefl Mohania active data warehouses complementing OLAP with analysis rules 2001(03)

4. M N Tho. A M Tjoa Zero-latency data warehousing for heterogeneous data sources and continuous data streams 2003

5. I Ankorion Change data capture-efficient ETL for real-time BI 2005(01)

6. J Langseth Real-time data warehousing:Challenges and solutions 2004

7. Rosana L Brito Assayag Rocha. Leonardo Figueiredo Cardoso. Jano Moreira de Souza Performance tests in data warehousing ETLM process for detection of changes in data origin 2003

8. A Silberschatz. F K Henry. S Sudarshan Database System Concepts 1997

9. 杨正洪 中文SQL Server 2000关系数据库系统管理和开发指南 2001

10. 李伟明 ORACLE REDO LOG文件分析及C语言描述[期刊论文]-小型微型计算机系统 2003(07)

11. Oracle Streams Replication Administrator's Guide.10 Release 2 (10.2) 2006

相似文献(1条)

1. 会议论文 林子雨. 杨冬青. 宋国杰. 王腾蛟 实时主动数据仓库中的变化数据捕捉研究综述 2007

实时主动数据仓库是数据仓库的最新发展阶段和未来发展趋势,它为企业提供了对战略决策和战术决策的双重支持.实时主动数据仓库中包含两类数据,即实时数据和非实时数据,相应地,需要两种不同类型的变化数据捕捉方法,即支持实时变化数据捕捉的方法和普通的(不支持实时的)变化数据捕捉方法.结合在该领域的研究经验,对实时主动数据仓库中可以使用的多种变化数据捕捉方法进行了系统地论述,并比较各种方法的应用条件、优点、缺点和适用场合.