浅谈数据仓库的实时性

文庆福*
清华大学软件学院 软件 11 班 100084

摘要：数据仓库是一个面向主题的、集成的、时变的、非易失的数据集合，支持管理者的决策过程 [1]。传统的数据仓库只能用于分析已有的历史数据，难以实现数据的实时提取、转变、载入等处理，满足实时分析的要求。本文主要分析目前实时数据仓库所面临的的一些瓶颈以及可能的解决方法。

关键词：数据仓库；实时性；ETL

A Brief probe into Data Warehouse

Qingfu Wen
School of Software Tsinghua University, Class 11, 100084

Abstract: A data warehouse is a subject-oriented, integrated, time-variant , nonvolatile collection of data in support of decision-making processes of managers [1]. Traditional data warehouse can only be used to analyze historical data which can not achieve real-time data extraction, transformation, loading and other processing to meet the requirements of real-time analysis. This paper mainly analyzes some bottlenecks currently faced by the real-time data warehousing and possible solutions.

Keywords: Data Warehouse; Real-time; Extract-Transform-Load (ETL)

1 引言

随着“大数据”概念的提出，数据越来越受到人们的关注，尤其的企业的关注，数据技术的发展对数据仓库的需求也日益增长。传统数据仓库的数据更新一般是每天、每周或是每个月一次，甚至更长，这使得我们在访问数据仓库时所获取的数据并非最新的。从 OLTP 系统更新的记录保存是不包括数据区的，这说明最近的业务记录没有纳入数据区，然而对于电子商业，股票经济，在线通讯等行业的数据信息或是一些突发事件的数据信息，需要及时发送给依赖它们的知识工作者或决策者，他们可能需要实时地获取这些最新的数据信息或是依赖实时数据信息分析结果，来辅助作出判断和决策。

2 简介

实时数据仓库（Real-time Data Warehouse）[2] 可以看作实时行为与数据仓库的结合。以超市商品的购买行为为例，一旦完成购买，与之相关的购买记录，如购买的物品、价格、数量等信息便可实时的存储到数据仓库之中。换句话说，实时数据仓库是这样一个系统，只要行为发生，数据变得可用时，知识工作者就可以实时获取信息，从而做出决策。

3 实时数据仓库的瓶颈 [3]

3.1 数据载入方法

实时数据仓库的面临的第一个挑战就是数据载入的问题。一个实时的数据仓库，不能像传统数据仓库采用缓存方式，最终按照定期地批处理方式进行延期加载，必须以实时地连续加载数据量。这样，实时获取的数据量不能过大。因此如何改变载入方式以扩大实时载入的数据量就成为了一大问题。

*thssvince@163.com

3.2 查询会话

实时数据仓库不仅允许用户在历史数据上进行分析处理，同时也能在实时数据上进行操作。查询会话就是影响实时数据仓库架构的一个因素。实时数据仓库要求能够做到持久在线，即意味着要在进行数据载入更新的同时进行查询处理。二者在发生的过程中可能会相互影响，如何平衡二者以实现一个高效的数据仓库是我们不得不面对的挑战。

3.3 索引技术

索引技术可谓是加速查询过程的关键。对数据进行增加、删除、修改、查找等操作都会与索引结构密切相关，因此，索引结构的好坏直接关系到了数据仓库的效率，实时数据仓库则更会受到索引结构好坏的影响。两种最常见的索引方法就是 B-tree 索引和位图索引。相比而言，在数据仓库中通常会采用位图索引，因为它可以实现更有利的压缩。

3.4 外键约束

外键约束是用来做引用完整性的检查的，例如在我们执行插入操作时，我们需要检查外键属性是否出现在引用的表格中。这个过程同样也会影响到数据载入的性能。为了加速这一个过程，我们可能可以将其移到数据转变的过程中。

4 实时数据库的解决方法

4.1 数据载入 [4]

文中提出了实现数据连续载入的方法，主要是修改数据仓库的模式、优化 ETL 的加载过程以及 OLAP 的查询优化等方式。

4.2 state-of-the-art ETL 工具 [5]

实时数据仓库需要解决的一个主要问题就是数据的更新周期，传统更新周期都是以天来计算，如果更新周期能够降低到秒的级别，则可以解决实时性的问题。降低更新周期可能会带来数据的不一致，或者是数据紊乱。那么如果在降低周期的同时保证数据的一致性和正确性。文中提出了一种 state-of-the-art ETL 工具的方式来解决这一问题。

5 总结

在传统数据仓库中引入实时性，实现数据的实时获取与处理，是数据仓库技术的又一大进步。实时数据仓库的实现需要我们对传统数据仓库体系结构进行优化，需要在实践之中去检验他的合理性。如何做到有条件实时获取数据，如何做到在获取数据同时实时处理请求，如果扩大实时数据的容量都将是我们亟待解决的问题。

参考文献

[1] Jiawei Han, Micheline Kamber, Jian Pei. 数据挖掘概念与技术 [M]. 范明, 孟小峰译, 北京: 机械工业出版社, 2013: 82-117.

[2] Inmon WH. Building the Data Warehoue. John Wiley & Sons Inc, 2003.

[3] Nickerson Ferreira, Pedro Furtado. Near Real-Time withTraditionalData Warehouse Architectures: Factors and How-to. IDEAS '13 Proceedings of the 17th International Database Engineering & Applications Symposium Pages 68-75.

[4] Ricardo Jorge Santos, Jorge Bernardino. Real-Time Data Warehouse Loading Methodology. IDEAS '08 Proceedings of the 2008 international symposium on Database engineering & applications Pages 49-58.

[5] Jorg T., Dessloch S.: Near Real-time data warehousing using state-of-the-art ETL tools. Enabling Real-Time Business Intelligence, Lecture Notes in Business Information Processing, Volume 41. ISBN 978-3-642-14558-2. Springer-Verlag Heidelberg (2010).