第1卷第3期　　空 军 工 程 大 学 学 报(自然科学版)　　Vol. 1 No. 3
2000年8月　　JOURNAL OF AIR FORCE ENGINEERING UNIVERSITY(NATURAL SCIENCE EDITION)　　Aug. 2000

数据仓库技术研究

张水平，　郑飞雁
(空军工程大学 电讯工程学院，陕西 西安　710077)

摘　要：数据仓库是近几年数据库领域研究的热点问题之一，也是决策支持系统(DSS)与数据库技术的结合点。在对传统的数据库与数据仓库的实现技术进行分析的基础上，讨论了数据仓库的实现过程、关键环节及相关技术。

关键词：数据仓库；数据模型；数据挖掘；联机分析处理
中图分类号：TP31　　**文献标识码**：A　　**文章编号**：1009-3516(2000)03-0068-04

在90年代初W. H. Lnmom提出了“数据仓库(DW，Data Warehouse 以下简称DW)”的概念，它的目标是借助于数据仓库全面、大量的数据存储，依靠数据挖掘技术与数据分析工具，达到高效的决策支持。数据仓库被定义为“是支持管理过程的、面向主题 的、集成的、随时间而变的、持久的数据集合。”由于数据仓库是数据库、数据挖掘(DW，Data Mining)、联机分析处理(OLAP，On Line Analytical Processing)等多类技术 的结合体，故在实现过程中仍有许多问题急待解决。但可经预计，下一世纪将会有各类实用的数据仓库应用系统问世。本文在对传统的数据库与数据仓库的实现技术进行分析的基础上，讨论了数据仓库实现过程、关键环节及相关技术。

1　数据仓库与数据库技术的异同

1.1　数据的主要特征

数据仓库与数据库的数据特征如下：

- 面向主题的结构设计——DW 是以最终用户的观点组织和管理数据，而传统的数据库为了提高应用程序查询数据的效率，因而以应用的观点设计库结构。
- 管理大量的信息——由于数据仓库的设计目标是在众多的数据库中获得决策信息，因而它含有大量的历史数据(一般为10GB左右)。而传统的数据库一般为100MB，因为传统的数据库为了提高运行效率，通常会对历史数据进行必要的备份后，将其从运行库中清除。
- 异质的数据源——由于数据据仓库的数据源来自于不同种类的文件(内部与外部数据源)，故数据存储的介质和格式会有很大的不同。因而数据仓库不但要处理不同数据库中的信息，还必须处理不同格式的数据文件。
- 高度概括的信息——传统的数据库存储的信息具体而详细，但不利于用户理解，数据仓库必须从大量具体的数据中进行高度概括，并挖掘出准确信息。

1.2　基本任务

数据仓库基本任务与传统数据库有很大的区别，由于数据仓库的数据源可以来自于不同DBMS的数据库中(内部数据源)，也可以来自于不同格式的文件中(外部数据源)。而这些数据源可看作DW中输送数据的管道。在输送数据的过程中，DW的设计者与传统的数据库设计者相比必须考虑如下额外的工作：

- 将这些数据源的模型转换为通用的描述形式；
- 将同义的数据元素的名称、数据类型、尺寸进行统一的规范——即净化数据元素，这项工作要求严

收稿日期：1999-12-17
作者简介 张水平(1956-)，女，山西新绛县人，副教授，硕士，主要从事计算机网络与数据库技术研究.

格区分所有数据源的同义词和多义词；

- 并非所有数据源中的数据元素都适用于数据仓库，故必须从各数据源中抽取子集，为形成DW的整体模型奠定基础；
- 把相似的数据源集成为统一的资源模型；
- 通过增加时间戳、来源戳、分割、衍生元素，提供扩展的模型用于存储聚集、概括值，从而获得仓库模型；

1.3 数据操作

- DB中支持用户对DB的大量数据更新操作；DW中则主要是查询操作，更新极少。与DB相比，DW中的数据相对稳定。
- DB为用户和开发者提供的是非常庞大和复杂的结果，但DW中要提供给用户的是可视化、易于理解的结果；
- DB中主要保存当前的数据，历史的数据被及时归档后立刻删除，以提高系统运行效率；而DW中则存储了大量的衍生数据，目的是为了节省工作量和提高运行效率，因为对大量的历史数据的处理往往很花费时间。
- DB包含其所需的、支持操作的所有数据细节；而DW只含有价值的概括性数据。由于上述种种原因，DW与DB相比其建模的方法有很大区别。

1.4 数据模型及建模方法

- 数据库的数据模型在传统上有三种：关系型、层次型、网络型。目前流行的数据库主流产品主要是关系型的。
- 数据仓库的数据模型也有以下三种[1]，但与数据库的数据模型不同。它们是：星型模型、雪花模型、混合模型。

2 数据仓库的体系结构与处理过程

2.1 系统体系结构

DW的体系结构如图1所示。

- 数据源——可以是在不同系统环境下建立的数据库文件和不同种类的数据文件。
- 数据仓库管理工具——负责对数据仓库进行实时管理。包括数据仓库中数据的输入（输入过程中须对数据进行数据的净化、转换、概括和聚集等），数据仓库管理工具必须随时监测数据源中的数据。一旦发现改变，应及时处理，并将正确数据输入至数据仓库中。另外，它还负责数据存储组织、数据的分发、数据仓库的例行维护等等。
- 数据仓库——它是通过数据仓库管理工具将来自于异地、异构的数据源加工后而形成的一种数据存储池。最终目的是为高级的决策支持服务。

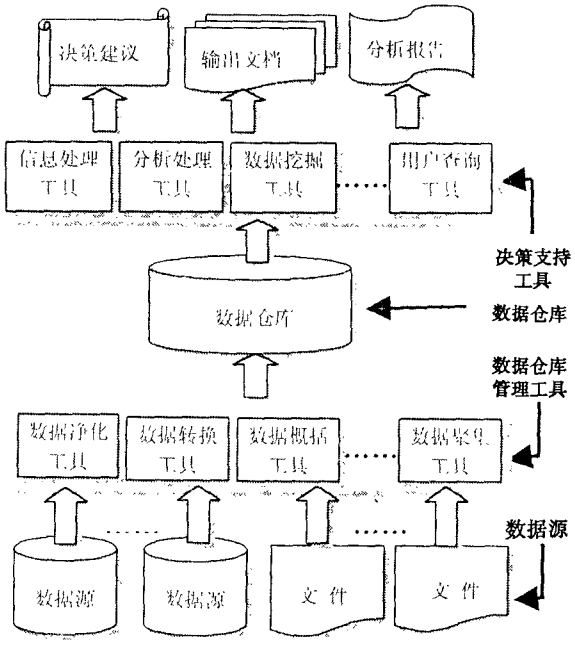

图1　系统体系结构

- 决策支持工具——主要是对数据仓库中的数据进行分析，挖掘数据库和数据仓库中的知识，并将其转换成辅助决策信息。

2.2 处理过程

- 数据仓库获取数据——从数据源中获取数据，包括对异质数据库、不同格式文件中的数据进行补漏、净化、转换（如对字段的同名异义、异名同义、单位不一等的处理）、分析综合，实施各类运算（如求和、求平均量、求统计量等），在必要时加盖时间戳。最终按照数据仓库的结构将数据存入数据仓库中。在这一过程中，应借助触发器技术以保障数据源一旦改变，数据仓库系统能够自动启动数据

获取模块,将新增的部分追加到数据仓库中,从而使数据仓库中的数据全面、准确。

- 数据存储与管理——数据仓库的存储与管理类似于数据库管理系统中的数据存储的组织、数据的维护、数据的日常管理等功能。
- 查询、分析数据——该过程的完成应借助于各种工具,如信息处理、查询工具、数据挖掘工具、分析处理工具等等,得出可视化图表、图形。
- 提交决策信息——根据上一步得出的分析结果,由专业人员进一步分析之后给出解释,数据仓库的最终用户将在这里获取比传统的数据库更直接、更易懂的决策信息。

3 数据仓库的关键环节及相关技术

3.1 关键环节

- 数据仓库模型的设计——要研制一个实用的数据仓库系统,数据仓库模型的设计非常关键。由于数据仓库是面向主题而不是面向应用的。因此,设计者必须在了解用户要求的基础上,仔细分析分散在各地的数据源的结构及数据本身,并进行语义的同化、结构的集成等复杂的处理,最后确定所采用的模型。
- 数据的输入——从异地、异质数据源中获取数据,并进行净化、集成、综合处理。这是数据仓库建设中最基础、最关键的一步。
- 数据的输出——将数据仓库中的数据按需要分发到分设的 DW 服务器上或经数据挖掘(DM)工具及联机分析处理(OLAP)工具处理,将结果提交给最终用户。这是数据仓库建设中最复杂、难度最大的一步。
- 数据仓库的维护和管理——由于数据仓库包含着来自于不同数据源的大量历史数据及概括性的数据,其数据量随着时间的推移会成倍增加。因此,对其进行管理是数据仓库的重要任务。由于数据量的增加会影响数据的查询速度。因此数据仓库的数据量也不是无限制的,所以对数据仓库也存在着清除和维护的问题。

3.2 相关的技术

- 海量、高速存储设备:由于数据仓库存储了大量的历史数据。其时限约为 5 年至 10 年,其数据量一般为 10GB,甚至达到 TB 级。数据量愈大,愈影响查询速度。因而海量、高速存储设备是数据仓库存在的基础。
- 数据库技术:数据库是数据仓库的基础,是数据仓库的数据源。
- 计算机网络技术:计算机网络是数据仓库赖以生存的硬件环境。
- 并行处理技术:由于数据仓库处理的数据量大且分散在异地,加之要进行复杂的处理及运算。因此,要提高系统的整体速度必须运用并行处理技术。
- (ODBC)技术:数据仓库的数据可能来自于异质数据库中。因此开放的数据库互连(ODBC)技术是必不可少的技术之一。
- C/S 体系结构:由于 C/S 体系结构的优越性,数据仓库的结构平台通常以 C/S 体系结构为基础。
- 决策支持工具(DSS):决策工具是数据仓库应用程序及工具的统称,用这些程序和工具可以检索、操作和分析数据仓库中的数据。然后将结果提交给用户。一个数据仓库应用系统的设计目的就是最终为用户提供决策支持信息。因此,一个实用的数据仓库系统应具有决策支持的功能。决策支持工具一般分以下三种:

信息处理方法:信息处理包含数据分析、基本的统计分析、查询和服务等技术。其结果以报表和图表的形式给出。

数据挖掘(DM)方法:DM 方法主要从大量的、具体的、细节数据中发掘出深层次的内容。它除了要用到统计分析工具以外,还要用到知识发现技术。上述两种方法集中在处理发生了什么。而 DM 方法在发现了所发生的现象以后还可以发掘出原因,以便于预测应采取的措施,同时还给出相应的置信度因子。

分析处理方法:分析处理主要是对大数量的历史数据进行多角度、多方位的分析,即称为多维分析,也称为在线分析处理技术(OLAP)。OLAP 技术通过交互式查询、多层次的概括和聚集、大量的商业转换和数据计算以及借助于模型进行预测、趋势分析和统计分析。并以二维或三维表格和图表、图形的形式给出结果。数

据仓库的数据需要进行纵向、横向多维的综合分析处理，才能得出准确结果。

参 考 文 献

[1] Lnmon W H. Building the Data Warehouse[M]. John WiLey/QED Ny,1996.
[2] [美]Harjinde S GILL. 数据仓库——客户/服务器计算指南[M]. 北京:清华大学出版社,1997.
[3] [美]Joyce Bischoff Ted Alexander. 数据仓库技术[M]. 北京:电子工业出版社,1998.
[4] 江放,李海刚,高国安. 基于数据仓库的数据挖掘及其在决策系统中的应用[J]. 现代计算机,1999,(1):32-35.

Studying on Data Warehouse Technology

ZHANG Sui-ping, ZHENG Fei-yan
(The Telecommunicatıon Engineerıng Institute,AFEU. ,Xi'an 710077,China)

Abstract: Data Warehouse ıs one of the focuses in Database analysis these uyears,and it is the contact between the DSS and the Database technology. On the base of studying ın the realizatıon techonlogy of Data Warehouse and traditionary Database,this paper discuses the realization process of Data Warehouse,key parts and interrelated technology.

Key words: data warehouse; data model; data mining; OLAP