浅谈数据仓库

浅谈数据仓库

江婷 谢文阁

摘要：阐述了数据仓库的概念特征和历史背景，并介绍了数据仓库的发展过程及其国内外的研究现状；同时还简要的介绍了数据仓库中所存在的问题，并展望了该技术的发展前景。

关键词：数据仓库 数据挖掘 决策支持 联机分析处理

Abstract：It describes the concept and historical background of data warehouse, and introduces its development process and current study circumstance. Meantime, it also briefly brings forward the problems about data warehouse and its development front ground.

Key Words：Data Warehouse；Data Mining；Decision Support；On-line Analytical Process

随着网络时代的到来，世界经济全球化已经变成一种共识。经济的全球化必然带来信息的全球化，伴随着信息与决策支持系统的发展过程产生了适应决策分析的决策环境—数据仓库，它是近年来兴起的一种新的数据库应用。数据仓库的出现和发展是网络时代的数据特征，也是数据库系统应用到一定阶段的必然产物。

1数据仓库的概念特征

数据仓库之父William H.Inmon 在1993年所写的论著《Building the Data Warehouse》中首先系统地阐述了关于数据仓库的思想、理论，为数据仓库的发展奠定了历史基石。在文中，他将数据仓库定义为：“一个面向主题的、集成的、随时间变化的、非易失性数据集合，用于支持经营管理中的决策制定过程。”其重要特征是：（1）面向主题性（2）数据集成性（3）数据的时变性（4）数据的非易失性。 [1]

数据仓库系统作为决策支持系统包括数据仓库技术、数据挖掘技术和联机分析处理技术。这些技术不仅体现了当今世界最先进的信息技术，而且还提供了能够对企业管理决策提供实际支持的系统。可以毫不夸张的说，数据仓库成功关键在于用户的应用情况，而不是数据仓库开发技术的熟练应用。[2]

2、历史背景

随着计算机技术的飞速发展和企业界不断提出新的需求，传统的数据库技术已满足不了这种需求。近年来，随着计算机应用，特别是数据库应用的广泛普及，人们对数据处理的这种多层次特点有了更清晰的认识。总结起来，当前数据处理可以大致划分为两大类：操作型处理（事务型处理）和分析型处理（信息型处理），这种分离划清了数据处理的分析型环境与操作型环境之间的界限，从而由原来的以单一数据库为中心的数据环境发展为一种新环境：体系化环境。

数据库系统作为数据处理阶段，主要用于事务处理。事务处理环境不适合决策支持系统的原因主要有以下五条：（1）事务处理与分析处理的性能特性不同；（2）数据集成问题；（3）数据动态集成问题；（4）历史数据问题；（5）数据的综合问题。[3]要提高分析和决策的效率和有效性，分析型处理及其数据必须与操作型处理及其数据相分离，并且把分析型数据从事务处理环境中提取出来，按照决策支持系统处理的需要进行重新组织，建立单独的分析型处理环境，正是为了构建这种新的分析处理环境而出现的一种数据存储和组织技术——数据仓库技术应运而生[5] 。

3、研究现状

整个80年代直到90年代初，OLTP一直是数据库应用的主流。然而当OLTP应用到一定阶段后，用户便发现单靠拥有联机事务处理不足以获得市场竞争的优势，他们需要对其自身业务的运作以及整个市场相关行业的情况进行分析，而做出有利的决策。因此，著名的数据仓库专家Ralph Kimball写道：“我们花了二十多年的时间将数据放入数据库，如今是该将他们拿出来的时候了。” [4]因此现在的实际情况是：20年前查询不到数据是因为数据太少了，而今天查询不到数据是因为数据太多了。针对这一问题，人们设想专门为业务的统计分析建立一个数据中心——数据仓库。 [6]

数据仓库在国外的应用已较为普遍，并呈现出应用较早，在电子化数据积累方面比较领先，业务应用较为丰富，有比较完善的管理和实施等特点。从

目前看，世界500强的企业多数都在建设或已经建成系统。国外电信运营商数据仓库的建设起始于20世纪90年代中后期，如：AT&T Wireless。

随着中国市场竞争的加剧和企业信息化的需要，国内的数据仓库建设得到了迅猛发展，如邮政行业引入数据仓库进行基本业务分析等。但整体来讲，由于国内数据仓库的建设和应用起步较晚，与国外相比还有相当的差距。我国数据仓库系统建设存在的问题，主要表现在以下方面：1、中国的信息化基础设备相对不太完善；2、企业的竞争意识和服务意识还不够强；3、数据仓库的价格居高不下；4、管理机制的缺乏；5、技术人才缺乏；6、数据挖掘工具本身不成熟；7、数据积累不充分。[5]

4、发展过程

数据仓库最早的概念可以追溯到20世纪70年代的MIT一次研究，该研究致力于开发一种优化的技术架构并提出这些架构的指导性意见。第一次，MIT的研究员将业务系统和分析系统分开，将业务处理和分析处理分成不同的层次，并采用单独的数据存储和完全不同的设计准则。

1988年，为解决全企业集成问题，IBM爱尔兰分公司的Barry Devlin Paul Murphy第一次提出了“信息仓库Information Warehouse”的概念，将其定义为：“一个结构化的环境能支持最终用户管理其全部的业务，并支持信息技术部门保证数据质量”。同时，在1988——1991年期间，一些前沿的公司已经开始建立数据仓库。1991年，Bill Inmon出版了其有关数据仓库技术的第一本书，第一次提供了如何建设数据仓库的指导性意见。后来，Ralph Kimball的第一本书“The Data Warehouse Toolkit”掀起了数据集市的狂潮。

现在，数据仓库应用更是得到了极大的发展，各大厂商纷纷宣布产品支持数据仓库并提出一整套用以建立和使用数据仓库的产品，因此掀起了数据库热。比如Informix公司的数据仓库解决方案，ORACLE公司的数据仓库解决方案等等。这同时也掀起了引起学术界的极大兴趣，国际上许多重要的学术会议，如超大型数据库国际会议（VLDB），数据工程国际会议（data engineering）等，都出现了专门研究数据仓库、OLAP数据挖掘的论文。[6]

5、存在的问题

数据仓库的理论知识的研究已经相当丰富，而且应用领域遍及通信、零售业、金融以及制造业等行业，然而数据仓库的理论知识在一些方面还有待进一步深化和充实，如：（1）目前数据仓库中用到的源数据大都是结构化和半结构化的，对于如何将非结构化的数据抽取、转换、装载到数据仓库中是一个很重要的问题。（2）

数据挖掘技术作为数据仓库中的一部分，它的研究方兴未艾，其研究与开发的总体水平相当于数据库技术在70年代所处的地位，迫切需要类似于关系模式、DBMS系统和SQL查询语言等理论和方法的指导，才能使DMKD的应用得以普遍推广。预计在本世纪，DMKD的研究还会形成更大的高潮，研究焦点可能会集中到以下几个方面：①发现语言的形式化描述；②寻找数据挖掘过程中的可视化方法；③研究网络环境下的数据挖掘技术，特别使在因特网上建立数据挖掘服务器，并且于数据库服务器配合，实现Web Mining；④加强对各种非结构化数据的挖掘；⑤处理的数据将会涉及到更多的数据类型，而且这些数据类型或者比较复杂，或者是结构比较独特。⑥更多的数据挖掘方法。（3）提高数据的质量。（4）数据仓库在可视化技术上的应用。（5）分布式数据仓库。

6、未来的发展趋势

数据仓库是一项基于数据管理和利用的综合型技术和解决方案。随着各种计算机技术，如数据库技术和应用开发技术的不断进步，数据仓库技术也在不断发展。

迄今为止，数据仓库的实用化已走过了近十年的历程，其规模都达到了T级，应用领域遍及通信、零售业、金融以及制造业。数据仓库的规模越来越大，被广泛应用于更高精度的数据分析中。数据仓库的支撑技术也有了新的进展。由此看来，数据仓库技术的发展前景将会十分广阔，在保险、铁路、航空、电信、医疗等行业中得到更加广泛的应用。

参考文献：

[1] W H Inmon. Building the Data Warehouse 2nd Edition. John Wiley & Sons Inc.1996

[2] Pailraj Ponniah.数据仓库基础.北京：电子工业出版社，2004

[3] 石丽，李坚.数据仓库与决策支持.北京：国防工业出版社，2003

[4] 陈京民.数据仓库原理.北京：中国水利水电出版社，2004年4月

[5] 沈云秋，张寅生.浅谈数据仓库技术.计算机应用研究.1999，（1）：11-13

[6] 张云涛，龚珍.数据仓库原理与技术.北京：电子工业出版社，2004