

●张学福

数据仓库技术与我国市场决策支持信息资源建设

ABSTRACT Data warehouse is a set of integrated and stable data supporting management decision, oriented to subjects and belonging to different periods of time. It is necessary and possible to use the technique of data warehouse in the development of market decision support information resources in China. 12 refs.

KEY WORDS Data warehouse. Decision support information. Information resources.

CLASS NUMBER G350

1 概述

信息化是当今社会发展的潮流和趋势,其程度的高低标志着一个国家的生产力水平,同时也决定着下个世纪的竞争实力与发展机会。1995年西方七国布鲁塞尔会议把信息基础设施建设和信息资源设施建设作为推进信息化建设的两个组成部分。信息基础设施建设包括数据通信骨干网、相应的数据交换标准以及一个基础软件;信息资源设施建设是指高度组织的信息提供系统,包括信息内容、信息内容处理方法,应用软件以及一套应用服务体系^[1]。在全球信息高速公路的推动下,我国在信息基础设施建设方面取得了长足的进步,限于篇幅,这里仅从信息资源建设角度进行探讨。关于我国传统数据库信息资源建设问题,笔者曾在文献[2]中论述过,这里不再赘述。本文是从另一角度来探讨市场信息资源建设问题,即探讨对公司、企业等经营、决策起重要作用的决策支持信息

资源的建设问题。

许多公司、企业在经过多年的计算机应用和市场积累后,已经保存了大量的原始数据和各种业务数据,其中包括订货单、存货单、交易条款和客户情况等,这些数据真实、具体地反映了公司、企业运作的本质状况。面对激烈的市场竞争,公司、企业的决策者们逐渐意识到公司、企业的数据和由此产生的信息是公司、企业的重要财富。但是,面对堆积如山的数据,决策者们很难从中发现公司、企业运作的本质规律。任何不符合客观事物规律性的决策都会给公司、企业带来损失,甚至失败。因此,他们需要从大量的业务数据中探索业务活动的规律性、市场的运作趋势,并从中得到参与市场竞争所必需的决策信息。

目前,决策者们一般是直接使用联机事务处理系统中的数据进行决策支持。采用该种方式对局部性问题进行决策支持数据分析效果比较好。而面向市场的决策往往是涉及整个公司、企业范畴的数据和信息。如果公司、企业的数据分布在若干个系统中,这样的

* 本文系国家社科基金“九五”规划项目“市场信息资源建设研究”的成果之一。

数据整合过程几乎难以完成。而且,这些数据最多保存过去 30~90 天的数据,这对于决策者,如财务管理者来说是远远不够的,因为他们要观察过去和现在的资金流动情况。除此之外,一个数据库的每一条记录也并不是某项决策都需要的,为此,在信息的分析过程中,分析人员需要从系统中人工下载数据,重新设置数据结构后传送到另一个数据库中,然后进行多维数据分析,才能获得有用的信息。随着市场竞争的日益加剧,这种掺杂着人工过程的信息分析已经难以满足公司、企业等决策用户的需要。能够很好地汇总分析公司、企业的各种数据的、满足决策者决策需求的数据仓库系统随之产生。本文通过对数据仓库技术特点等方面的介绍,探讨了基于数据仓库技术的市场决策支持信息资源建设的有关问题。

2 数据仓库技术

2.1 数据仓库的概念

关于数据仓库,国际上没有严格、统一的定义。目前普遍采用 W. H. Inmon 对数据仓库的定义:“数据仓库是支持管理决策过程的、面向主题的、集成的、稳定的、不同时间的数据集合”^[3]。因此,数据仓库不是一种产品,而是一种技术手段——它把某单位各个不同的数据从传统数据库中进行清理、抽取和转换,并按决策主题的需要进行重新组织,然后放到一个集中的数据仓库中,供决策人员分析、利用。

2.2 数据仓库的特点

从前面的定义可看出,数据仓库不同于传统的数据库,它除具有传统数据库的数据的独立性、共享性等特点外,还具有以下特征^[4,5]:

(1) 面向主题。传统的数据库是面向应用的,它可以满足要求很高的在线事务处理(OLTP)。数据仓库是面向主题的,而主题

是一个在较高层次上将数据归类的标准,每一个主题基本对应于一个宏观的分析领域,它可以综合公司、企业网络不同信息点上数据库中不同时间段的数据,从而能更好地支持分析决策。

(2) 存储详细的历史数据。传统数据库中很少涉及历史数据,一年左右或更短的近期数据即可满足日常分析需求,数据更新较快。数据仓库通常存有有关各个主题的不同时期的综合信息,为便于进行发展趋势和结构变化等各种分析,数据经集成进入数据仓库后是极少更新的,数据时限一般为 5~10 年。

(3) 存储综合性数据。传统数据库数据冗余度一般很小,而且通常情况下也不存储综合性的数据。由于数据仓库支持各种管理分析与决策,因此在存储详细历史数据的基础上,还采用层次结构定义法存储各层次的综合数据,数据冗余度大。这种既存储详细数据又存储综合数据的方式比较适合人们的一般需求,如问题研究往往从综合层次入手,不断细化和深入,逐步找到问题的原因或解决方案。

由以上可知,数据仓库将大量用于事务处理的传统数据库进行清理、抽取和转换,并按决策主题的需要进行重新组织,可为辅助决策集成多个部门不同系统的大量数据,是支持管理决策过程的、面向主题的、集成的、稳定的、不同时间的数据集合,其目标是进行决策支持。结合我国市场信息资源建设的实际情况以及公司、企业等决策用户的实际需求,数据仓库技术比较适合我国市场信息资源建设中决策支持信息资源的建设。

3 在市场决策支持信息资源建设中采用数据仓库技术的必要性

3.1 对市场分析决策的需要

对市场情况的分析结果是公司、企业决

策的重要依据,因此特别强调分析结果的有效性。为保证其有效性,要求提供市场分析的是一种能反映多个层面经济活动的信息及与之相伴的数据结构,这类信息不但能反映市场演变的瞬间状态,而且还能反映它的历史和趋势。这些需求是传统的数据库无法满足的,以面向主题等特征为特征的、适合于该种需求的数据仓库技术也就提到日程上来。

3.2 提高对客户服务质量的需求

在激烈的市场竞争条件下,公司、企业的命运和前途在很大程度上取决于它与客户的联系以及它所提供的服务质量。服务质量的好坏依赖于它所使用信息的质量。传统的数据库无法保证整个公司、企业所有数据更新完全同步,再加上其他因素的影响,无法从根本上改善对客户的服务质量。经由一个单一的数据仓库提供面向“客户”这一主题的信息,是提高对客户服务质量的有效途径。

3.3 实现公司、企业规模优化的需要

为增强公司、企业产品的竞争能力,公司、企业决策者根据公司、企业运行的实际情况,需要及时对业务处理过程进行重组,使自己的资源配置优化,形成规模化经营,进而提高公司、企业的竞争能力。为了很好地完成重组过程,公司、企业的决策者不仅需要具备分析其业务信息的能力,而且需要能对业务运作过程本身的合理性做出判断,也即提供的信息能便于决策者从多个侧面进行分析、评价和判断。这种需求,传统的数据库是满足不了的。

3.4 公司、企业资产/负债管理的需要

资产/负债管理是任何公司、企业都必须面对的问题。传统的数据库无法准确描述采购、库存等的变化和相互影响情况,而数据仓库能够比较准确地反映出诸如采购、库存等对资产/负债状况有影响的要素的变化和相互影响,因而能超前和主动地强化管理。

3.5 决策支持信息支撑环境的需求

目前,多数组织机构需要提供两种不同

的数据。一种是针对日常事务处理需要的数据,另一种是针对决策支持过程的数据。前一种数据大多是实时数据,随时被更新;后一种数据包括经过汇总、推导、整合过的数据。这是两类性质完全不同的数据,如果使用相同的数据环境(如支持 OLTP 应用的传统数据库)同时支持这两类数据,那么性能、处理能力和产生的效益都会大打折扣。而这两类数据的初始数据源都是相同的,但用于决策支持的数据要对原始数据进行转换后得到。数据仓库的主要目的是以最小的代价将前一种数据转换为后一种数据,以满足决策者的各种需求。

3.6 深化经济体制改革的需要

我国的公司、企业正处在由计划经济向市场经济转轨时期,要由计划经济体制过渡到市场经济体制,发展社会主义市场经济,无论是宏观决策部门还是微观的经济实体,都需要大量、完整的决策信息作为决策的参考和依据。

4 在市场决策支持信息资源建设中采用数据仓库的可能性

4.1 在线分析处理(OLAP)技术的支持

在线分析处理也称多维分析,它是一种数据分析技术,能够完成基于某种数据存储的数据分析功能,满足用户决策所需的在线分析处理的需求。下面简要介绍一下 OLAP 的特点^[6]:

(1) 响应时间短。OLTP 因其提供的在线查询范围小,问题相对比较简单,响应时间一般都比较短。而用户所需的决策信息量非常大,OLTP 已不能在用户可忍受限度内很好地完成查询。OLAP 以其专门的数据存储格式、大量的事先运算等技术的支持,可在较短的时间内对大量的数据进行分析、处理。

(2) 多维性。为把数据库中大量数据变成有用信息,为决策支持服务,原有的支持

OLTP 的数据库的数据组织、存取方法和实现方法已不能满足需求,如它不能支持多维数据库组织、多维聚集计算等。OLAP 的一个重要特征是多维数据分析,它是通过把一个实体的多项重要的属性定义为多个纬度,使得用户可以方便地汇总数据。简化了数据的分析处理逻辑,并能对不同纬度值的数据进行比较,而纬度则表示了对信息的不同处理角度。应用多维分析可以在一个查询中对不同阶段的数据进行纵向和横向比较,这在决策过程中非常有用。因数据仓库中存储的大量数据是根据多维方式组织的,所以 OLAP 既可以独立使用,也可以与数据仓库结合。现在一般都采用二者结合方式,即 OLAP 的数据来源于数据仓库。

(3) 可分析性。OLTP 系统是以索引为基础来满足用户的各种查询操作,不具备可分析性。而支持 OLAP 的数据库的各种操作不能完全基于索引进行,它能处理与应用有关的任何逻辑分析和统计分析,来满足各种用户的决策需求。

(4) 信息性。信息性是指不论数据量有多大,也不管数据存储在哪里,OLAP 系统能及时获得信息,并且管理大容量的信息。对 OLTP 系统来说,这些是很难做到的。

(5) 可支持任何复杂的查询。传统 OLTP 中的查询是基于索引的,往往比较简单而且固定。OLAP 中的查询一般都比较复杂,查询时经常需要联合搜索很多表,有时一个查询的返回结果可能有几千条记录,需要做进一步的处理,才能获得对决策支持有用的信息。

4.2 数据开采技术的支持

数据开采是从数据库或数据仓库中发现并提取隐藏在其中的信息的一种技术。它建立在数据库,尤其是数据仓库基础之上,面向非专业用户,支持即兴的随机查询。数据开采技术能自动分析数据,对它们进行归纳性推理和联想,寻找数据间内在的某些联系,从

中发掘出潜在的、对决策行为起着十分重要作用的信息,从而建立新的业务模型,以达到帮助决策者制定市场策略、做出正确决策的目的^[7]。如前面介绍 OLTP 在查询过程中,有时一个查询的返回结果可能有几千条记录,用户很难根据这些信息做出决策,这时就需要根据返回的信息做进一步的数据开采,最终获得对决策支持有用的信息。

4.3 我国数据库信息数据量已达到一定规模

截止到 1995 年底,我国自建库数量已达 1038 个,数据库所占空间在 100MB 以上的比例为 24%,有近一半在 10MB~100MB 之间^[8]。另外,我国各种规模的信息系统经过近几年的发展,电子化信息规模达到了一定程度;政府部门在长期对社会经济进行统筹规划和监督调节的过程中,积累了和继续产生着大量对国计民生至关重要的信息资源;特别是公司、企业在长期经营过程中积累了大量的信息。调查研究结果表明,大多数公司、企业并不缺少数据,而是受阻于过量的冗余数据和数据不一致,并且其数据量正以成倍的速度增长。只不过是它们变得越来越难以访问、管理和用于决策支持。

4.4 我国已有数据仓库技术的应用实例

由中国银行广东省分行承担和开发的国家“八五”科技攻关项目“中国银行省、市两级金融管理信息系统示范系统”,就是运用数据仓库等技术建立起来的,该项目已于 1996 年 6 月通过国家鉴定和验收^[9]。为了有效地研究消费者的消费动态、提高竞争优势,上海强生有限公司利用 IBM 公司的可视化数据仓库解决方案 Visual Warehouse 和面向对象的关系数据库 DB2,成功地建立起了数据集市,有效地提高了对信息的利用和分析能力^[10]。这些都为我国利用数据仓库技术进行市场决策支持信息资源建设积累了经验。

4.5 网络通信支撑环境

(1) 中国公用分组交换数据网(CHINA

PAC):该网为不同速率、不同类型终端之间的通信及数据资源共享,提供较高质量的国际和国内数据通信业务。

(2) 中国数字数据网(CHINADDN):它可向用户提供高达 2MB 或更高速率的数据通道,用于大容量高速数据通信。

(3) 金桥工程:它是国家公用经济信息网工程。它以卫星为主,与邮电数据网互为备用,并与各部委和各省市的信息数据专用网互相连通,它能进行数据、语音、图像的传输,是一个综合专业数字网。

4.6 国家政策方面

1984 年 9 月,邓小平同志在为《经济参考报》题词中作出“开发信息资源,服务四化建设”的重要指示。在邓小平同志的率先垂范下,党和国家许多领导人都先后对信息工作做出重要指示,如江泽民总书记指出:“四个现代化,哪一化也离不开信息化”。1995 年 5 月,中共中央、国务院《关于加速科学技术进步的决定》指出:“重视科技信息的有效利用和传播,加强科技图书、资料和数据库的建设,要有计划地建立全国科技信息资源传输的设施,建设连接全国科研机构、高等学校的科教信息网络,实现科技信息共享和交流的现代化。”再加上《中华人民共和国国民经济和社会发展“九五”计划和 2010 年远景目标纲要》等一系列重要文件,都为我国数据库信息资源建设提供了良好的发展环境。

在市场经济条件下,正确及时的决策是公司、企业生存与发展的最重要环节。越来越多的公司、企业认识到,只有充分利用、发掘其现有数据,才能实现更大的效益。数据仓库正是汇总这些数据的基础,进而支持多维数据分析、数据开采等技术和传统的查询及报表功能。这些对于在当今激烈的竞争中保持领先是至关重要的。因此,我国在市场决策支持信息资源建设中采用数据仓库技术不仅非常必要,而且各项条件也已基本具备。

5 在市场决策支持信息资源建设中应用数据仓库技术的几点建议

5.1 加强宏观调控,促进以数据仓库为基础的市场决策支持信息资源建设协调发展

数据仓库技术是一项回报率较高的项目,在由 IDC 为主开展的一项关于数据仓库建设的投资与收益的调查的项目中,发现这些公司的平均投资收益率(ROI)为 401%,有些公司的 ROI 超过 1000%。超大型数据仓库项目的 ROI 为 533%,大型数据仓库项目的 ROI 为 322%^[11]。尽管如此,因其投资规模较大,平均为几百万美元,为避免我国数据库建设中的盲目建库等类似现象出现,防止公司、企业仓促上数据仓库项目(个别公司、企业可能因资金等条件制约,在投资回收期到来之前出现资金等困难,影响公司、企业的正常运转,严重的可能会导致公司、企业倒闭),国家应责成某一统一的管理机构,在调查的基础上进行宏观调控,确定哪些公司、企业可以自己或联合建设数据仓库,这样既可以避免重复建设,又可在不影响公司、企业正常运作中建立自己的数据仓库,增强公司、企业的竞争力。

5.2 优先支持大型公司、企业建设数据仓库

鉴于我国国情,应首先在大型公司、企业,特别是国家重点扶持的、争取进入世界 500 强的公司、企业中优先建设相应数据仓库。在这些公司、企业中建设数据仓库,一是他们拥有雄厚的数据基础和经济基础,二是他们有数据仓库建设所需的硬件基础和人才基础。另外,这些公司、企业的经营者、管理者信息意识好,信息需求发展迅速,因此容易成功,进而增强公司、企业的竞争力,推动公司、企业的发展。在数据仓库的实际建设中,应多借鉴国外的经验,特别是美国在数据仓库建设方面的经验,这样可避免走弯路和不必要的浪费,使我国的公司、企业数据仓库在

比较短的时间内达到世界水平。然后带动其他公司、企业数据仓库的建设。

5.3 数据仓库建设中应贯彻“全盘考虑,逐步完善”的思想

由于数据仓库中的数据是详细的、集成的和历史的,其中的数据量一般都很大,而且随着时间的推移,增长速度也非常快。因此,建立数据仓库时最好在全盘考虑的基础上分步进行,否则建设周期将非常长。这是由于在分步建设过程中每步都能取得阶段性的成果,不需要等到二三年后才能访问数据仓库中的信息。世界上许多成功的1000GB(指用户数据量而非数据库大小)级以上的数据仓库在开始时的规模都不大^[12]。

另外,当数据仓库发展到十分庞大时,按原有模式对其进行处理会使得中央数据仓库的复杂性增加,不易维护,数据冗余度大。一般的做法是采用数据仓库——从属数据集市两级模式,即在原有数据仓库基础上建立一个或几个从属数据集市,对它们进行相应处理,这样既能提高响应速度,又能保证整个系统数据的一致性和易维护性。其中从属数据集市实际上是数据仓库的一个子集市,它存储着供最终用户使用的有关应用或特定用途的公司、企业信息,其数据来源于公司、企业的数据仓库,并依靠后者来维护。

5.4 加强数据仓库技术的培训,提高决策用户的素质

数据仓库技术是为支持其使用者进行决策而建立的,利用数据仓库的人对系统的运作是决定性的因素。只有决策用户成功地利用数据仓库的决策功能才可以证明数据仓库的成功。而决策用户要成功利用数据仓库的

决策功能,就必须对数据仓库的全部情况进行了解,在此基础上掌握相关的内容,掌握的越深入,将来运用的就会越好。因此,数据仓库建成后必须对决策用户进行培训,使数据仓库的决策功能真正发挥作用。

参考文献

- 1 高新民. 南非会议的启示——参加信息社会和发展大会后的报告. 信息经济与技术, 1996(9)
- 2 张学福, 邓平. 我国数据库信息资源建设研究. 中国图书馆学报, 1997(2)
- 3 陈文伟等. 数据仓库与决策支持系统. 计算机世界, 1998-06-15
- 4, 11 李敏强等. 数据仓库系统的结构与设计研究. 决策与决策支持系统, 1997(2)
- 5 郭宜斌. 数据仓库技术的基本概念和发展现状. PC World China, 1996(4)
- 6 陈元, 陈文伟. OLAP的多维数据分析. 计算机世界, 1998-06-15
- 7 杨晓等. 数据仓库、数据集市和数据挖掘. 现代计算机, 1998(1)
- 8 张学福. 我国社科数据库信息资源分析与研究. 情报资料工作, 1997(1)
- 9 温巩, 罗亦文. 运用数据仓库技术建设金融管理信息系统. 国际电子报, 1997-09-01
- 10 IBM可视数据仓库软件助上海强生一臂之力. 国际电子报, 1998-05-18
- 12 王闯舟. 数据集市≠数据仓库. 计算机世界, 1998-07-20

张学福 讲师。出版著作、工具书5部, 发表论文10余篇。通讯地址: 哈尔滨黑龙江大学信息管理系统。邮编 150080。

(来稿时间: 1998-08-18。编发者: 刘喜中)