

数据仓库——企业信息化建设的方向

Data Warehouse: the Direction of Enterprise Information Development

庄玉良¹ 王 萍²

(中国矿业大学管理学院, 副教授¹; 研究生² 徐州 221008)

当今世界科学技术日新月异、市场环境瞬息万变, 现代企业的生存与发展, 已在很大程度上取决于能否及时、准确、全面地利用信息资源。换言之, 设法从大量原始的数据中析取、提炼出有价值的信息以作为人们行动和决策的依据, 已成为现代企业信息化建设的目标和方向, 因此数据管理技术, 即如何科学地组织和储存数据、如何高效地获取和处理数据, 一直是信息技术面临的关键问题之一。

一、数据仓库有利于解决数据有效利用的诸多问题

数据管理技术随着计算机软硬件技术的发展

和方式等一系列行为。

更深入的探讨还表明, 区分技术预见与技术预测的还有后验评价法, 其中最具直观性的区分方法就是要看研究工作者是如何组织和如何操作此类活动的, 如果在他们所开展的活动过程中将技术的预测和选择等权利完全或大部分委交于科学界的“精英专家”和一部分技术专家的话, 那么, 这种预测和选择就不是技术预见意义上的“预测”和“选择”, 而是技术预测意义上的“预测”和“选择”。虽然在世界各国进行着各式各样的“预见活动”, 且大都采用“德尔菲问卷调查法”, 但如果采用的方法仅仅是向分布在科技界和政府科技管理部门工作的少数技术专家发放问卷, 那么, 这种“预测”和“选择”仍然不是技术预见意义上的“预测”和“选择”, 而是技术预测意义上的“预测”和“选择”。

在我们试图将技术预见与技术预测严格区分开来的同时, 也有不少学者倾向于将二者给统一起来, 即将技术预测称之为技术预见的初级阶段。但我们认为, 无论致力于将技术预见同技术预测区分开来, 还是致力于将技术预见同技术预测统一起来或连贯一致, 都是为了更好地理解什么是技术预见、什么是技术预测, 它们都有哪些特点、有哪些共同点、差别是什么, 等等。无论怎么说, 我们在这里

而不断发展, 已从人工管理、文件系统阶段发展到了数据库系统阶段。企业利用数据库技术建立了信息系统, 较好地实现了数据的存储和管理目标, 企业的信息化建设因此而迈出了一大步, 但是在已存储的大量数据的有效利用上却仍存在着以下诸多问题。

(1) 注重数据的登录和保存 (Record Keeping), 忽视数据的综合处理与信息的系统分析 (Information Analysis)。

(2) 人、财、物、产、供、销等管理环节之间“数据孤岛”现象严重, 同一数据的多源性导致数据的不一致性, 难以提供全面、可靠、科学的决策信息。

(3) 企业即使已消除“数据孤岛”, 但数据处理

所探讨的东西不是定论的东西。我们探讨的结论是未定型的结论, 它的基础是否牢固不在于我们的态度是怎样的, 而在于我们到底掌握了多少有价值的历史资料。

参 考 文 献

- [1] 李健民, 浦根祥. 试论产业拉动因子在技术预见中的角色. 科技导报. 2001(5)
- [2] 李健民, 浦根祥. 技术预期与政府控制. 科学管理研究. 2001(7)
- [3] 李健民, 万劲波. 技术预见与上海市环境科技的发展. 科技导报. 2001(12)
- [4] 李健民, 浦根祥. 技术预见时代的来临. 世界科学, 2002(4)
- [5] Martin, Ben R, Irvine, J. (1989), 'Research Foresight: Priority-Setting in Science', Pinter Publishers, London
- [6] Anderson, J. (1997), "Technology Foresight for Competitive Advantage", Long Range Planning, Vol. 30, No. 5, pp. 665-677
- [7] Godetm, M. (2000), "The Art of Scenarios and Strategic Planning: Tools and Pitfalls," Technology Forecasting and Social Change, No. 65, pp. 3-22
- [8] Martin, B. and Johnston, R. (1999), "Technology Foresight for Wiring up the National Innovation System," Technological Forecasting and Social Change, No. 60, 99. 37-54

(责任编辑 王宏章)

仅面向部门事务(称为事务驱动, Transaction Driven)的应用模式设计, 数据是专用的, 属性设计的系统性差, 企业系统内缺乏一体化的、高度集成的信息综合分析(称为分析驱动, Analysis Driven)模式。

(4) 虽方便了当前数据的查询, 但难以进行历史数据的综合应用(历史数据通常在一个周期结束后经备份即删除), 在一定程度上, 以计算机为手段的信息系统与手工数据处理系统的区别只是数据存储介质的不同, 多周期(如若千年)数据的综合处理存在相当难度。数据量迅速增加与数据分析手段滞后的矛盾逐渐突出, 人们望“数”兴叹。

(5) 数据处理的内容和实现的功能带有明显的手工管理的烙印, 所形成的是操作型数据库(Operational Database), 而数据之间潜在的关系却难以深层次地挖掘, 对于高层次决策者的决策支持程度较低。

在信息系统的数据管理上存在的问题需要借助新的信息技术予以解决。数据仓库(Data warehouse)就是为解决数据有效利用的新型数据管理技术, 它在数据容量、数据集成、数据综合、数据分析、数据查询等方面都得到了显著的改善。从信息系统的前景来看, 数据仓库是对组织的数据管理基础结构的重要改进, 会提高整个组织的报告与分析能力^[1]。数据仓库是数据挖掘(DM, Data Mining)和知识获取(KDD, Knowledge Discovery in Database)技术得以有效应用的基础, 是复杂的客户关系管理(CRM)启动的关键因素^[1], 是企业信息化建设的方向。

二、数据仓库的概念与特点

数据仓库的概念是 90 年代初美国 Prism Solutions 公司副总裁 W. H. Inmon 在《创建数据仓库》(Building the Data Warehouse)一书中首先提出的。他认为^[2]: 数据仓库是面向主题的(Subject-Oriented)、集成的(Integrated)、稳定的(Nonvolatile)、不同时间的(Time Variant)数据集合, 用以支持经营管理中的决策制定过程; 它是 90 年代信息技术架构的新焦点, 可提供集成化的和历史化的数据; 它可以集成种类不同的应用系统; 数据仓库从事物发展和历史的角度来组织和存储数据, 以供信息化和分析处理之用。

斯坦福大学数据仓库研究小组则这样定义数据仓库: “数据仓库是集成信息的存储中心, 这些信息可用于查询或分析”。^[3]

数据仓库与传统数据库在很多方面存在着不同。传统数据库是对于事件发生后的数据描述, 它面向事务, 因此被称为事务型数据库或操作型数据库, 其数据采用关系表格形式存储, 它的处理方式是联机事务处理(OLTP, Online Transaction Process-

ing), 强调瞬间响应。数据仓库用于联机分析处理(OLAP, Online Analysis Processing), 不特别追求瞬间响应。OLAP 采用关系型数据模型和多维数据模型, 使用多维数据库(MDDB, Multidimensional Database)或基准表(Table), 采用星型(Star Schema)、雪花型(Snowflake Schema)、立方体树(Cube Tree)等形式组织数据, 通过切片与切块(Slice and Dice)、钻取(Drill)、旋转(Pivoting)等操作分析数据。操作型数据库与数据仓库的区别^[4]可用表 1 表示。

表 1 操作型数据库与数据仓库的比较

比较内容	操作型数据库	数据仓库
用户	各层次用户	高层管理者、分析人员、客户服务代表
目标	保存记录(用于OLTP)	信息分析(用于OLAP等)、决策支持、数据库营销
更新	联机	成批处理
查询	自动、细节	整合、概括、集成
时间	主要是当前	历史、现在、未来
数据来源	内部	内部、外部
方向	面向实体(产品、账户、客户等)	面向种类(产品种类、账户种类、客户种类)
数据容量	GB	GB/TB
过程	事务驱动	分析驱动
结构	相对静态	动态

数据仓库与操作型数据库在数据上是有联系的。数据仓库中的数据来源于操作型系统和外部信息源, 即从信息源系统中提取(Extract)数据, 并经过清理(Clean)(如检测和修正错误)和转换(Transform)(输入到主题组中或者经过概括)而加载(Load)到数据仓库中^[5], 形成一个面向主题的新型数据结构。因此, 数据仓库不是对操作型数据库的替代, 它们在企业的信息环境中承担不同的任务, 对企业决策分析和日常事务处理发挥着各自的作用。

综上所述, 可以将数据仓库技术的特征归纳如下^[6]: (1) 存储面向管理应用和综合分析的集成化和综合性的信息, 从历史的角度描述系统结构和状态的变化, 采用能够反映时间维特征的数据结构; (2) 基于传统的面向组织事务的数据库或外部数据源, 经过提炼、加工、汇总和归一化整理(Consolidation), 生成符合数据应用语义规范要求的数据集合; (3) 能够支持多种复杂的数据应用和综合性的管理决策分析。

二、数据仓库系统的开发、实施与管理

数据仓库系统(Data Warehouse System)通常由源数据、仓库管理、数据仓库和分析工具 4 部分组成, 可利用信息包图、星型图等工具, 通过数据仓库结构与环境的定义(即定义元数据)、数据源分析与确定、数据的清理与转换、数据加载与存储、数据查询与分析等步骤和功能建立起来。

与传统的数据库系统相比, 数据仓库系统的建立也要经过概念设计(需求分析)、逻辑设计、物理设计等环节, 但是开发和管理数据仓库系统的技能

表 2 操作型数据库建设和数据仓库建设的任务

阶段	内容	数据库建设	数据仓库建设
计划	未来在规模和使用方面的发展	中等	很大
	数据持续性和整合性	事务处理	集成
	存储要求	较少	多(磁盘阵列、并行性等)
设计	目标	满足业务操作数据要求	满足业务分析信息要求
	方法	主要是E-R图	主要是多维图形
	范围	面向应用	面向企业
实施	方法	整体	渐进
	数据粒度	基于事务的更新	成批更新
	查询方式	报告预定义	分析特别查询
	数据水平	细节	概括与集成
维护	安全	基于视图	基于用户
	安全风险	安全的相对破坏产生影响	安全的相对破坏有深远影响
	备份与恢复	备份花费较少时间	备份花费较多时间; 备份行为比较关键

和考虑的因素要比建立传统数据库系统复杂得多, 开发费用和维护成本极高。因此, 考虑到数据仓库规模、建立速度、开发和维护成本、见效快慢等因

素, 形成了数据仓库系统建设的两种模式, 即渐进开发模式和直接开发模式。渐进开发模式 (Incremental Development Approach) 主张^[4]首先分别面向一个主题或面向一个部门 (Department-Oriented) 建立小型数据仓库, 即数据集市 (Data Mart), 最后将各个数据集市整合在一起, 形成数据仓库, 这是考虑数据仓库规模和复杂性后的最好战略。直接开发模式又叫企业级开发模式 (Enterprise Development Approach)。将若干数据集市整合成一个数据仓库会产生不相容问题, 直接建立企业级数据仓库 (EDW, Enterprise Data Warehouse) 比数据集市的整合要简单。^[7]

笔者认为, 企业信息化建设是一个系统工程, 操作型数据库系统的开发和数据仓库系统的建设是这个工程的两个阶段, 后者建立在前者基础上, 因此, “系统规划, 分步实施”是企业信息化建设的总体策略。操作型数据库系统的开发要面向全局, 考虑未来数据仓库系统建设的需要, 采用基于业务

表 3 数据仓库成长的阶段和基准因素的变化

基准因素	初始阶段	成长阶段	成熟阶段
数据(主题数量、数据模型、存储数量)	单一或几个主题区域的有限数据	多个主题区域的数据	全企业范围的数据, 完好的集成, 多时段
结构(数据集市、数据仓库的结构)	单一数据集市	多数据集市	一个仓库, 并带有相关数据集市
产品环境的稳定性	程序是特殊而逐步发展的	程序还没有很好建立	程序有章可循
仓库员工(阅历、技能、专业化程度)	内部人员无经验, 频繁使用顾问	内部人员已有经验, 对顾问的依赖不严重	内部人员有经验, 分工与职责明确
用户(种类、数量和区域)	通过数据集市服务的业务单位分析员	通过数据集市服务的所有业务单位分析员; 对信息需求和计算机技能的要求不同	整个组织的用户访问数据仓库; 供应商和客户都可以访问仓库数据
用户技能和工作情况的影响	一些用户可能不具备更多分析工作需要的技能或爱好	更多的用户经历了工作所需要的技能的转变	整个组织的用户都需要提高计算机技能以适应工作
应用(利用仓库数据的各种应用)	报告或预定义特殊查询; 逆推已发生了什么	报告和预定义查询; 对事情为什么发生和“如果...就...”的假设进行更多的分析	报告, 预定义特殊查询, DSS, EIS; 数据挖掘提供了预测模型能力; 与操作系统集成
成本和收益(与仓库相关的各种成本和收益)	成本中等; 包括节省时间的收益, 新的和更好的信息, 改善决策	包括节省时间的收益, 新的和更好的信息, 改善决策; 收益首次超过成本	包括节省时间的收益, 新的和更好的信息, 改善决策; 再造业务流程, 支持公司目标; 高投资回收率可能实现
组织影响(仓库对组织运营有多大影响)	对几家业务单位的操作和战术方面有影响	对另一些业务单位的操作和战术方面有影响	在全组织范围内经常有战略及操作和战术方面的影响

流程再造 (BPR) 的开发战略^[8], 科学地设计业务流程和数据流程, 系统地设计数据属性。数据仓库系统的建设应吸收渐进开发模式和直接开发模式的先进理念, 首先系统地设计数据仓库的结构和数据环境, 然后按照总体蓝图建立数据集市, 最终整合成企业级数据仓库。

表 2 所列是操作型数据库系统和数据仓库系统不同建设阶段的任务^[4], 有助于从事企业信息化建设的人员参考。

从表 2 可以看出, 数据仓库的建设有其特殊的任务和任务; 此外, 有明确的目标和任务、富有成效的业务分析以及获得高级主管的支持是建设数据仓库的有效策略^[9]。

数据仓库的实施与维护符合成长理论 (Growth theory)。Hugh Watson 等学者在 8 位数据仓库专家的共同参与下, 研究确定了数据仓库成长的 3 个阶段及其变化情况^[5]。这 3 个阶段是: 初始阶段(数据仓

库的初始版本)、成长阶段(数据仓库的扩展)、成熟阶段(数据仓库完全融入企业的运营), 可以使用 9 个基准 (Benchmark) 因素描述在不同阶段中的变化。表 3 为数据仓库成长的各个阶段及其基准因素变化。

三、数据仓库的应用

数据仓库的建立为用户综合利用历史数据奠定了基础, 可以大大提高决策的范围、速度和质量。数据仓库中有价值信息的获取需要通过数据挖掘等技术得以实现。数据挖掘是指主要通过概率分布 (Probability Distributions)、相关 (Correlation)、回归 (Regression)、聚类分析 (Cluster Analysis) 和判别分析 (Discriminant Analysis) 等统计模型, 评估假设和结果, 并对结果加以利用; 并通过人工神经网络 (Artificial Neural Networks)、决策树 (Decision Trees)

和遗传算法 (Genetic Algorithms) 等机器学习的人工智能技术, 使用一整套数据训练方法, 学习和发现模型参数^[10]。数据挖掘的主要应用领域有^[10,3]: (1) 零售 通过市场货篮分析(MBA, Market Basket Analysis) 或购物货篮分析(SBA, Shopping Basket Analysis) 了解客户的购买行为和喜好; (2) 银行 鉴别客户的贷款活动, 寻找新客户, 加强客户服务; (3) 信用卡 分析信用卡持有者的购买行为和信誉, 谨防犯罪; (4) 金融服务 为制定投资战略建立交易和风险模型, 用于现金交易、索赔、自动认购、股票选择、信用评估、鉴定犯罪类型、抵押筛选; (5) 直销 根据客户的历史购物记录和人口统计数据将公司的产品分发或邮寄给客户; (6) 航空 根据历史资料发现乘客的旅行模式, 改进或新增航线, 扩展服务范围; (7) 电信 为用户提供其希望购买的各种新型服务, 鉴别短期内可能转移的客户; (8) 制造 控制和规范生产程序, 降低失误, 防止潜在的质量问题出现; (9) 保险 分析决定保险额的主要因素, 预测客户保险模式, 有效预防犯罪的发生; (10) 经营管理 评价客户信息, 评估员工和部门的业绩, 监测子公司或部门的财务舞弊行为。

数据仓库和数据挖掘技术的应用将给用户带来巨大的效益, 有较高的投资回报率。据加拿大国际数据公司所进行的调查, 成功实施数据仓库的平均投资回报率为 401%, 平均期限为 1.7 年^[9]。因此, 数据仓库的建立和实施已成为大多数公司信息技术启动的最高战略目标。据 Palo Alto 管理集团预测, 2002 年数据仓库相关市场 (包括硬件、软件、服务和内部费用) 将增加到 1 135 亿美元^[11]。值得注意

的是, 数据仓库项目不同于传统意义上的信息技术项目, 它的商业潜力大, 但投资大、实施要求高, 也有较高的失败率。我国企业应根据自身实际, 提出切实可行的信息化建设策略, 按照科学的方法和步骤, 逐步达到成功实施数据仓库这一信息化建设的目标。

参考文献

- [1] Hugh J. Watson. Current Practices in Data Warehousing. Information System Management, Winter 2001
- [2] Inmon William H. Building the Data Warehouse. New York: John Wiley & Sons, 1992
- [3] 胡侃, 夏绍玮. 基于大型数据仓库的数据采掘的研究综述, 软件学报, 1998(1)
- [4] Alan Benander. Data Warehouse Administration and Management. Information Systems Management, Winter 2000
- [5] Hugh Watson. Data Warehousing Stages of Growth. Information Systems Management, Summer 2001
- [6] 李敏强, 潘振江, 寇纪淞. 基于数据仓库技术的决策支持系统的研究与应用. 系统工程理论与实践, 1998(3)
- [7] Gardner Stephen R. Building the Data Warehouse. Communications of the ACM, Sept 1998
- [8] 庄玉良. 企业信息化建设新思路: 基于 BPR 的 MIS 开发战略. 中国软科学, 1999(4)
- [9] Wendy T. Marks. Building Customer Data Warehouses for A Marketing and Service Environment: A Case Study. Information Systems Management, Summer 2001
- [10] Lei-da Chen, Toru Sakaguchi. Data Mining Methods, Application, and Tools. Information Systems Management, Winter 2000

(责任编辑 王宏章)

(上接第 37 页)

1/5 左右。

2. 灌溉成本的比较 引洮工程的真实成本为 700~870 元/亩。集雨节灌工程, 因不引用黄河水, 因此没有水资源的机会成本, 只有维修折旧 (按投资的 5%) 和投资利息 (6%), 其值为 $(0.05 + 0.06) \times 1500 = 165$ 元/亩, 不到引洮工程的 1/5~1/4。引洮灌溉由于定额大, 单产即使高一些, 其边际效益仍远远抵不上边际费用。

3. 受益范围的比较 引洮工程因受地形、地块等因素的限制, 在涉及的 900 多万亩耕地范围内, 只能在地形有利的地方灌溉 100 多万亩, 受益面积和人口甚为有限。而集蓄雨水节灌则能适应较复杂的地形, 国家用同样的投资, 能发展更多的灌溉面积, 能使更多的人受益。

4. 占地数量的比较 引洮工程的渠系、渠边道路、地埂占灌溉面积的比重达 15%~25% 左右。集雨节灌工程主要是集雨场占地, 每亩集水场可灌 4 亩多地, 占地比为 25% 左右, 但大多占用坡度在 15° 以上农用价值很低的非农地。

5. 集蓄雨工程在向工业和城市生活供水方面也优于引洮工程 引洮工程单位引水量的投资为 17.3 元/ m^3 , 向工业区和城市供水的自来水厂, 输水损失按 25% 计, 单位净供水量的投资为 23.1 元/ m^3 。集蓄雨工程若修在工厂、市区边缘地带, 蓄水工程以大水池为主, 则其向工业和城市供水的单位水量的投资为 25.3 元/ m^3 , 虽然比引洮工程的投资略高, 但实际工程可随着用水的增长而分期修建, 不积压投资。因此, 今后的工业和城市生活供水, 亦应由集蓄雨工程来替代。

6. 对黄河水资源的影响 引洮工程引水后势将加剧黄河水资源的进一步紧缺; 而集蓄雨节灌和向城市供水, 是开发利用雨水, 基本上不用黄河水源, 因此不会加剧黄河水资源的紧缺。

7. 集蓄雨工程的缺陷 在枯水年份, 集水量将剧减, 这是集蓄雨工程的一大不足。届时可减小灌溉定额和灌溉面积, 作为应急对策。

综上所述, 比较利弊后, 我们建议, 今后对于类似引洮灌溉这类传统工程, 应及早考虑用节水灌溉和集蓄雨工程来替代。

(责任编辑 蔡德诚)