

新一代数据管理系统

周龙骧 (中国科学院数学研究所)

摘要

集中式数据库管理系统在十年前已经推出了若干成功的商品。分布式数据库系统经过十年的研究与开发,其技术已经成熟,正处于已经推出和正在推出一些商品的时期。数据库管理系统已经进入了酝酿、研究和开发新一代系统的时刻。

一、引言

衡量一门学科是否已经成熟,其基本问题是否已大部分得到解决,可以看它是否已出了一本比较完整的书。衡量一种计算机软件技术是否已经成熟,可以看它是否已有一种提供市场的产品推出来。集中式的关系数据库系统十年前已经推出了若干种成功的商品,因此可以说关系数据库技术十年前已经成熟了。分布式数据库技术从七十年代下半期开始,投入了巨额的资金和研究努力,经过十年的研究和开发,已有若干原型系统研制成功,并已推出和正在推出一些商品。因此可以说分布式数据库系统技术也已臻于成熟。从八十年代初开始(甚至可以上溯到七十年代末)人们开始探讨数据库系统新的研究方向。新的DBS样式(Shape)(例如数据库机,数据库和人工智能的结合——演绎数据库,知识库,专家系统等,数据库和计算机网络的结合即分布式数据库);数据库的新的应用领域(例如管理信息系统,办公自动化系统,决策支持系统, CAD/CAM/CASE/CIM);

和数据库的新应用密切相关的是数据库作为大系统的一个组成部分的重要作用;新的数据模型;新的数据表示和结构(如图形,图象,声音,自然语言,正文、文档等);新的用户接口以及数据库理论的深化。这些新方向的研究取得了丰硕的成果,它们深化和提高了原有的理论和技术,使原有的系统更为精炼,高效,应用更为广泛,同时开拓了未来,迎来了新一代数据管理系统的研究和开发时代。

二、历史回顾

虽然早期的高级语言大多数是以程序为中心,如ALGOL, FORTRAN等,但数据处理却扮演着越来越重要的角色。回顾起来数据管理已经走过了四个阶段,有了四代数据管理系统。

1. 第一代(1950年代) 这一代DMS基于顺序文件,其硬件基础是磁带,著名的语言是COBOL。数据按记录组织,顺序记在顺序文件(即磁带)上。这一代DMS是面向批处理的。

2. 第二代(1960年代) 由于出现了分

Office, ACM TOOLS 5:4(1987)

[6] E. M. Gerson, S. L. Star, Analyzing Due Process in the Workplace, ACM TOOLS 4:3(1986)

[7] S. F. Ehrlich, Strategies for Encouraging Successful Adoption of Office Communication Systems, ACM TOOLS 5:4 (1987)

时,交互式终端,磁鼓,磁盘,这一代DMS基于顺序文件和直接存取文件。特别是后者发展了索引,杂凑等技术的文件系统。可以说文件系统技术已经很成熟了。

3. 第三代(1960年代中期——1970年代中期) 这一代DMS是真正的DBMS。IDS.1是第一个集成化的DBMS。第三代DMS的特征是基于存取模型(Access Model)的。数据模型允许DBA描述DB的物理结构,存取路径等。用户可以经由存取路径进行导航式存取(Navigator)。有名的事件有DBTG报告。CODASYL 1971年推荐文本是DBTG报告的发展,它标志了以C.W.Bachman为首的第三代DMS达到了巅峰状态。Bachman在1973年获得计算机科学界的最高奖——图灵奖。他的图林奖得奖演讲的名称是:“作为导航员的程序员”。

4. 第四代(1970年代——1980年代中期) 第四代DMS是关系型数据库系统。关系模型由E.F.Codd在1970年提出,引起了一系列的理论研究和实际系统的开发。整个七十年代可说是关系系统从理论到实践全面取得辉煌成就的时代。在理论上确立了完整的关系理论,在实际上开发了成功的关系系统,著名的有IBM圣约瑟研究实验室研制的System R及其产品SQL/DS, DB2, 加利福尼亚大学柏克莱分校研制的INGRES及其商品,关系软件公司的ORACLE等。关系系统使用统一的语言来定义、操纵和检查数据,用户只须指明what而无须指明How,系统将自行决定如何存贮和取得数据。SQL是由IBM设计的类英语的结构化查询语言。美国国家标准协会(ANSI)1986年10月批准的美国国家标准SQL是以E.F.Codd为首的第四代DBMS达到顶峰的标志。关系模型的奠基人E.F.Codd在1981年获得图灵奖。

5. 第五代(1980年代——) 如上所述第四代DBMS或称为RDBMS业已商品化,意味着RDBMS的研究已经基本完成,我们当前处于第四代DBMS的末期,进入了第五代

DBMS的研究和开发的时期。第五代或新一代DMS是尚不存在的系统,它应该是什么样?它的体系结构如何?它应具有些什么特征?这些都是没有确定答案的问题。但正如引言中所述经过八十年代以来的研究,我们已经可以对这些问题作一根线条的回答。

三、新一代DMS的特征

迄今为止市场上提供的DBMS涉及的应用领域通常是商务领域,其特点是所处理的事务一般比较小,诸如存款取款,购票订票,财务会计,人事管理,公司管理等。随着应用领域的拓广,传统的DBMS已不足以应付新的应用需要。例如CAD/CAM等工程数据库应用中,其模式变化很频繁。一般是长事务,要求事务嵌套,能部分地COMMIT,要求COMMIT之后能够UNDO等等。又如地理,测绘,城市管理,军事应用等数据库,要求处理图象和图形,大型情报检索应用,要求处理大量的正文,办公自动化系统,要求处理格式极其复杂多样的文档,专家系统中要求理解自然语言以及处理语音等。日本第五代机计划要求其最底层的超大规模集成电路层之上是数据库管理系统层。作为第五代机的组成部分,这里的DBMS有其特定的要求。在涉及人工智能领域时要求新一代DMS具有演绎,推理等功能。这些要求综合起来我们可以看出新一代DMS特征的轮廓,即:

1. 应用范围广泛 CAD/CAM等工程数据库的应用,适应动态变化的模式定义和修改;长事务,嵌套事务,事务可部分地COMMIT,事务COMMIT之后可以UNDO;交互式反复设计修改应保证DB的一致性;分布式设计及分层设计;大范围检索和预先集中提取信息。

地理,测绘,遥感,城市管理服务,军事应用等要求加工管理图象,图形等数据。

情报检索中的大正文加工处理,多索引检索,同义异形检索等。

计算机辅助软件工程(CASE)中的各种

软件信息管理,软件可重用技术,软件子系统集成等。

计算机集成制造生产(CIMS),管理生产的全过程,各种不同类型的对象的表示和管理。

决策支持系统(DSS),管理信息系统(MIS)中的应用,作为DSS, MIS的支持系统。

第五代机中的基础子系统。

演绎数据库系统,知识库管理系统等的系统成份,支持演绎,推理功能,自然语言理解,语音识别等。

办公自动化系统中各种形式文档的识别,表示和管理。

2. 处理对象复杂多样

- 简单的事实可表为关系的元组。
- 复杂的对象有多种,如正文,图形图象(Graphical image),象素结构图象(pixel Structured image)等。
- 由事实和对象的有序集组成通用文档。

3. 具有演绎和推理能力 支持定义规则,支持从所存贮的对象去推导出用户所需的导出对象,支持对象从一种形式到另一种形式的转换,如从关系到正文,从关系到图形图象等以及相反。

4. 用户接口友好方便 各种层次的对用户友好的各种接口。

在用户接口语言方面有高度抽象的数字化逻辑语言(如关系代数),有英语化的语言(如SQL),有图表化或形式化(form)语言(如QBE),有自然语言等以及各种宿主语言如COBOL, FORTRAN, PASCAL, ADA, 还有image导航语言及语音识别设施。

各种方便的报告生成(包括分页,分段,加标题,加脚注,各种字型选择),图表生成(包括直方图,饼图,折线图,曲线图,散点图等),表格加工(包括排序,分页,小计,合计,横向统计,报表格式重组等)等设施。

四、新一代DMS的体系结构

七十年代弄清楚的关于DMS多层抽象的体系结构对于构思新一代DMS的体系结构是大有裨益的。DMS的上端面向各种应用,提供适应各种应用的用户接口, DMS的下端面向OS和物理的存贮设备, DMS则完成从最终用户的要求到具体物理存贮设备中的数据之间的转换。对于新一代DMS如前所述其所面向的对象更加多样和复杂,其所依存的物理设备则更为先进,速度更高,容量更大,并可并行地进行存取。相应地, DMS的转换功能也更为复杂。

根据近几年的研究探索,通常认为新一代DMS应当分为三层,即面向应用层,对象转换层和存贮/检索层。

1. 面向应用层 (Application Oriented Layer) 如第三节所述面向应用层应能应付复杂而多样化的对象,对每种类型的应用提供特定的功能,对最终用户提供非常高级非常友好的接口。提供灵活的视图功能(这在前几代DBMS中实现得很不好或根本未予实现)。既能表示和检索图象,编辑,加工和检索正文,管理和使用form,表示和组合复杂对象的文档,以及管理和查找文档;提供自然语言接口,言语识别和综合;专家系统的接口;演绎DB和KB的接口; CAD/CAM/CASE/CIMS的接口等等。

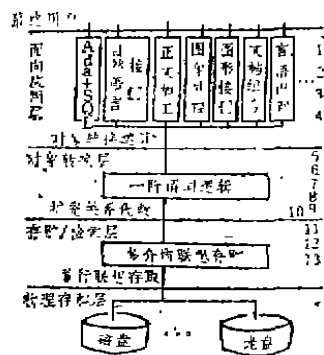
2. 对象转换层 (Object transformation Layer) 将面向应用层的对象及各种复杂请求转换为一系列下层能执行的动作例如扩充的关系代数。执行部分推理(为提高性能)。进行完整性检查,授权检查,事务管理(广义的事务)等。

3. 存贮/检索层 (Multi-media associative memory Layer) 存贮介质是多样的,如磁盘,光盘等。存贮是联想存贮即按内容存贮而非按地址存贮。存取是高速并行的而非串行的。因而检索也是按内容高度并行的检索。

于是我们有如下的新一代DBMS体系结构:

五、讨论

以上给出的新一代DBMS的体系结构是近年来研究探索的某种结果,它仅给出了一



注：1——各种视图；2——各种表示；3——部分推理；4——转换演算；5——通用对象的组合变换；6——复杂请求分解为一系列扩充关系代数运算；7——部分推理(提高性能)；8——完整性检查；9——事务管理(广义事务)；10——授权检查；11——联想存贮；12——联想检索；13——并行存取。

种指导性的轮廓。很明显，很多问题尚未解决，还有很长的路要走。问题的清单可以提一大串，例如：

1. 知识表示的转换问题，如何从一种form转换为另一种form？
2. 自然语言理解问题。

3. 新的查询语言的定义。
4. 如何组合文档？如何存贮和检索它们？
5. 图象的表示和检索问题。
6. 规则库的组织管理问题。
7. 对象的表示和转换问题。
8. 如何高效地实行推理？
9. 广义事务的定义和管理。
10. 优化的准则和优化技术。
11. 事实，正文，图象，对象的存贮方法。
12. 并行存取问题。
13. 存取路径的建立和利用。

如上所述的通用的新一代DMS其体系结构过于复杂，也许可行的办法是设计和实现范围专门一些的系统。例如专用于CAD/CAM/CASE/CIM的系统，专用于ES的系统，专用于DSS，MIS的系统，专用于五代机的系统，专用于图象处理的系统，专用于正文处理的系统等等。将所有功能集成到一个通用的系统中，其可行性和性能还有许多值得研究的课题，期望不久将会有若干原型系统研制出来，它们将提供非常宝贵的经验和教训。

(上接封四)

- [10] N. Adam, Rin Betting, On- 4GLs Or COBOL, CW July 9 1986
- [11] Robert Morison, 4GLs vs COBOL, CW Aug.12 1985
- [12] Richard Watts, Fourth-generation language how to choose the right one, data processing Vol.27 No.9 1985
- [13] Al Hendrickson, Performance Benchmarks For 4GLs, CW July 9 1986
- [14] John Desmond, Language barrier, The fourth generation at work, CW Nov. 11 1985
- [15] Keith Nelson, Technical requirements of fourth-generation languages data Processing Nov. 1985
- [16] 4GLs drawing bead on PCs, CW No.20 1986 P19
- [17] Richard Cobb, Fourth-generation language, from backwater to mainstream, CW No.41 1985
- [18] Dan TagKer, In Search of Fourth Generation Data, DATAMATION July1, 1987
- [19] Rebecca Hurst, DBMS and 4GLs, Splitting the Difference, CW No.1A 1987
- [20] E.F. Codd, Is your DBMS really relational?, CW No.41 1985