计算机科学 1999Vol. 26№. 4

数据库技术：回顾与展望

Database Technologies: The Retrospect and Prospect Views

周傲英 邱越峰 田增平 施伯乐
（复旦大学计算机科学系 上海 200433）

Abstract In this paper, the successful 30-year history of database system is retrospected, and the significant achievements in the field of data management are surveyed. Some sophisticated database techniques are listed. In the meantime, the challenges faced by the database system, when entering the new century, are analyzed. Based on the description of the novel database applications, some research directions, which are worthy of paying close attention to, are presented.

Keywords DBMS, Semi-structured data, Transaction processing, Query optimization, Novel database application, Data mining

数据库系统的研究和开发在其三十年的历史中取得了巨大的成功，形成了一个数百亿美元的产业。数据库技术和系统的应用已经遍及各个领域，奠定了数据库系统作为当今社会信息基础设施核心技术的地位。尽管如此，数据库的研究和发展一刻也未曾停止过，传统的研究主要集中在增强和提高数据库管理系统(DBMS)的功能和性能上。但是，DBMS作为管理大容量数据的工具无疑会受到所管理的对象和所应用的环境的影响。目前，数据库系统要管理的对象不再局限于传统数据库所擅长管理的结构化数据，半结构化的数据及各类多媒体数据的管理需要对数据库技术提出了挑战。另一方面，数据库系统的应用环境也发生了变化，不再是一个可控的封闭环境，万维网(WWW)和其他信息服务形式产生了新的应用环境。

正是基于以上考虑，本文在回顾数据库系统的发展历史，综述现有的重要技术的基础上，通过对新的应用举例的讨论，分析了21世纪的数据库技术所面临的挑战，列出了与当前及未来信息系统发展和社会信息化进程密切相关的几个研究方向。

1. 数据库系统的发展

数据库系统实际上是一个集成了大容量数据管理技术的大型复杂软件系统。从数据库技术的起源和发展历史来看，数据库系统也是从数据库管理工具集逐渐演变而来的。数据库技术的诞生是以六十年代IBM推出的数据库管理产品IMS为标志的。IMS是层次型DBMS的代表。其后，数据库又经历了网状型、关系型、面向对象型及对象关系型等发展阶段。

在数据库系统出现以前，各个应用拥有自己的专用数据，通常存放在专用文件中，这些数据和其他文件中的数据有大量的重复。数据库概念的一个重要贡献就是应用系统中的所有数据独立于各个应用而由DBMS统一管理，实现了数据资源的整体管理。IMS系统的推出，使得数据库概念得到了普及，也使得人们认识到数据的价值和统一管理的必要，但由于IMS是将数据组织成层次的形式来管理，有很大的局限性。网状型数据模型就是CODASYL DBTG为试图克服这种局限性而提出的。网状型是基于图来组织数据，对数据的访问和操纵需要遍历数据链来完成。这种有效的实现方式对系统使用者提出了很高的要求，阻碍了系统的推广应用。1970年，IBM研究中心的E. F. Codd博士发表了关于关系数据模型的著名论文，被公认为是数据库发展历史上的一个重要里程碑。在关系模型中，数据是按照数学中关系概念来组织和理解的，对用户而言，数据库中的按照关系形式组织的数据可简单直观地理解为二维表。由于关系模型的简单易理解及其所具有的坚实理论基础，整个七十年代和八十年代的前半期，数据库界集中围绕关系数据库进行了大量的研究和开发工作，对关系数据库概念的实用化投入了

大量的精力。迄今为止，关系模型仍然得到大多数数据库产品厂商的支持。目前的数据库产品市场中，关系数据库系统占据了绝大部分份额（如 DB2、Oracle、Sybase、Infornux、Ingres 等）。

在关系数据库技术和系统取得交口赞誉的同时，对更强功能的信息管理技术的要求也越来越迫切。人们首先注意到的是非商业应用系统（如 CIMS、CASE 等）的需求，这些应用中数据的典型特点是结构复杂、对管理有特殊要求。因此，在八十年代初期，面向对象数据模型应运而生，面向对象数据模型是基于面向对象程序设计原理的，其最初的目的是提供一个可扩充的数据模型，使用户可对各应用专用的数据类型进行抽象。经过六、七年的研究和开发，形成了一批面向对象数据库（OODB）系统。随后，这些厂商自发成立了一个对象数据管理组（ODMG），形成了 ODMG 标准。由于技术惯性等原因，OODB 并未像人们期望的那样取得更大的成功。许多关系数据库厂商也在保持原有产品风格的同时，吸收了对象技术的优点，提出了所谓的对象关系数据库（ORDB）模型，将他们各自的产品升级为对象关系数据库管理环境。

回顾数据库技术和系统的发展历史，数据库领域所取得的成就主要体现在关系数据库、事务管理和查询优化等方面。关系数据库由于其简单性和清晰的概念基础，得到了研究人员、数据库厂商和最终用户的青睐，形成了数据库目前的繁荣局面。事务管理是 DBMS 支持数据共享和多用户操作的关键，是 DBMS 保持数据正确性及简化应用编程人员工作（无须关心其它并发使用同一数据的应用程序的干扰）的基本措施。查询优化是数据库系统性能提高的基础，尤其是在关系数据库系统中，由于系统性能主要由系统自身负责（这是关系系统之所以简单易用的原因之一），查询优化显得更为重要。从某种意义上说，关系数据库得以取代层次和网状型数据库而成为市场主宰，查询优化技术突破是一个重要因素。

2. 现有的数据库技术

传统数据库技术和系统的研究是基于数据存储在访问速度相对较慢的存储设备上且有多个并发用户同时访问的观点而展开的，主要集中研究的问题是数据库系统的性能、正确性、可维护性及可靠性。高性能所涉及到的问题是：数据量远比内存容量来得大，甚至会分布地存放在多台机器上。正确性保证是通过保证数据的完整性约束（如：引用完整性）和可串行化事务来达到。达到系统的可维护性的措施包括将数据库中数据的逻辑和物理结构分离开来，外加 DBMS 提供的一组方便数据库设计和系统性能调节的工具。保证系统可靠性的典型手段是将写前日志和事务结合起来以便在系统软硬件发生故障时维持数据的一致性。

为解决 DBMS 的性能、正确性、可维护性及可靠性等问题，数据库界近三十年来发展了以下技术，并成功地在各种 DBMS 产品中加以实现和广泛应用。

在数据建模方面，对数据模型的理解可简单归结为数据模型由一个定义数据库结构的语言（数据定义语言，DDL）和一个操纵这些结构的语言（数据操纵语言，DML）组成。模式就是用 DDL 描述的数据库的定义。数据库中的所有数据都由模式来描述，这样一来，DBMS 就可将数据的物理存储结构和应用级的抽象（逻辑结构）分离开来，这就是所谓的数据独立性。有了数据独立性，存储结构的改变可以不影响具体应用的运行。关系数据模型是这方面的典范，这也是关系数据库起主导作用的原因所在。

在查询语言方面，数据库界达成的共识是，查询语言是用来描述从数据库中检索数据的高级语言。用它所描述的查询的结构应相对简单、易于理解及便于自动生成和优化。目前流行的 DBMS 大多支持国际标准的 SQL 语言。SQL 能表达要从数据库中返回什么数据而无需涉及存储结构或访问算法。在这个意义上，SQL 是说明性语言，它所表达的查询易于理解。当然，说明性查询语言需要 DBMS 中强有力的查询优化器的支持。查询的请求和具体实现的相互分离也可以保证存储器结构的改变不至于影响查询表达式的有效性。

在查询优化和计值方面，关系数据库之所以成为商业现实，极大程度上是因为关系查询语言优化器的成熟和高效的查询计值算法的开发。基于查询的形式及当前存储结构自动地生成查询执行计划的能力是 DBMS 功能的一个重要组成部分。

在数据管理方面，视图概念的提出简化了用户对数据库的使用，藉之数据库管理员可以用查询语言定义用户眼中的数据库，用户无需了解全貌。另一方面，基于同样的原理，视图可限制用户对整个数据库的访问，使之只能访问视图中可见的数据。此外，索引自动维护和缓冲区管理等数据管理技术也是 DBMS 中的重要功能。

在并发事务管理方面，数据库界提出了事务的

概念，以解决由并发访问和更新所带来的正确性问题。采用了基于原子性的正确性标准，事务的提出简化了应用编程。编程人员无需担心来自其它运行程序的干扰。事务同样也是恢复的基本单元，它与写前日志机制结合可解决数据库系统的可靠性问题。

在分布式系统方面，数据库系统面临的问题是处理数据分布于多台机器时所带来的问题。两阶段提交(2PC)协议是针对这一问题提出的，它既考虑到了分布并发事务的特点，又保留了原子事务的优点。分布查询处理、死锁检测和异质数据集成也是研究的热点，取得了实用的成果。

3. 新的数据库应用举例

从以上对数据库领域现有技术的讨论可以看出，传统数据库所擅长的是进行固定环境(如：公司、机关部门)下大容量结构化数据的管理。但是，随着社会信息化程度的提高和计算机技术在各个应用领域的推广应用，许多新的应用环境也提出了迫切的数据管理方面的需求。本节列出一些需要数据库支持的新的应用领域。这些领域有些已经成功地应用或扩充了现有数据库系统的功能，有些只采用了折衷或变通的方式进行数据管理，如利用中间件技术和系统。

计算机集成制造(CIM)环境中集成的思想最重要的是体现在数据的集成中。众所周知，CIM 系统是当前面临激烈的国际市场竞争形势下，企业赖以生存的关键，被认为是现代企业发展的推进器。在 CIM 环境中，各设计部门、生产部门、行政管理部门都有各自的信息子系统，并且这些部门甚至都是属于完全不同的管理机构或业主，它们通常是为完成某一项制造建设任务而临时聚集在一起的。它们的信息子系统之间的信息交换、协同管理、共享和协调是实现敏捷制造的关键。这些系统所涉及的信息结构复杂、形式多样，对管理有许多特殊要求(如安全保密、协同设计等)。可以想象，如果各自为政，实现系统集成何等困难。为此，国际标准化组织制定了产品数据交换标准 STEP/EXPRESS。但目前很多企业只遵循标准规范了各自的数据交换格式，数据的存放和交换仍是利用文件的方式。针对这种环境，数据库技术和系统的进一步研究和开发会有力地推动 CIM 系统的发展。

万维网(WWW)环境中缺乏数据库技术支持。WWW 环境就是一个管理极大量非标准数据的环境，这其中缺少了数据库所应扮演的角色，是整个数据库界的一块心病。一个很有说服力的例子就是现在许多数据库厂商都在试图在它们的 DBMS 产品中扩充 WEB 连接能力，目的是提供更好的 WEB 服务器。这些努力只是朝着管理 WEB 上极大量非标准数据方向上迈出的一小步，并且值得怀疑的是成千上万的 WEB 站点的建设者会转移来使用一个 DBMS 构造其站点。

个人信息系统广泛使用是未来信息社会发展的必然趋势，数据库技术的应用是必不可少的。个人信息系统就是根据每个使用者的需求对信息进行裁剪并直接提供给本人。使用的主要设备是可移动的无线个人信息设备，如：个人数字助理(PDA)、手持 PC、膝上机等。个人信息系统强调的是个性化信息服务，理想的系统应是根据使用者的工作性质、行为习惯、兴趣爱好等及时提供信息服务。这些服务通常是与地理位置相关的，因而个人信息设备应配有 GPS 卡。在这样的系统中，个人信息设备必须不断地访问远程数据库和监视广播信息。这对当今客户/服务器的性能、可伸缩性、可靠性都提出了挑战。服务器发布信息的方式不再局限于拉(Pull)，推(Push)的方式也同样重要。另一问题是负载平衡的问题，包括多级服务器上及服务器与个人信息设备上数据和任务的分配。

在新的数据应用中，EOSDIS(地球观测系统数据及信息系统)是另一个典型的例子。EOS 是美国 NASA 发射的一组卫星，其目的是收集信息以支持科学家研究大气层、海洋和陆地的长期运动趋势。这组卫星每年发回地球 PB(10^{15}，即千万亿字节)级的信息。这些数据还要与来自其他数据源的数据和信息集成。它不仅要满足科学家的研究需要，还要满足科学业余爱好者的信息需求。数据存储在 EOSDIS 中。EOSDIS 面临的问题是：提供 PB 级的规模空前的数据库的存储管理，支持成千上万信息消费者的各种信息请求以及提供有效的数据浏览和搜索机制。

其他新型数据库应用也各有其特点。电子商务应用中，异质信息流必须有效地集成，对分布验证和资金流动需要不同寻常的安全措施。在保健信息系统中，一个病人的医疗记录可能存在多个医院、医务办公室或保险公司的系统中，病人的病史信息必须从这些不同的地方搜索。其它的信息，如医疗过程药物、诊断工具及其它辅助治疗信息则可以从另外的系统和数据库中得到。其中所涉及到的问题是异质信息的集成、医疗信息的保密性保证及适合于保健

人员使用的用户界面设计，在协同设计应用中，并发和共享机制、工作流管理及多版本管理则是最基本的要求。

4. 数据库技术面临的挑战

根据对上述新型数据库应用的分析可知，传统的数据库技术和系统具有显而易见的不适应性，这对传统的数据库技术和系统的研究开发工作提出了挑战。为应付这些挑战，数据库界有两条途径来提供解决办法，一是反省先前的研究开发思路，将现有的思想和技术进行扩充、推广和转移来解决所面临的问题。在上述新型数据库应用的分析中，许多现有的数据管理思想和数据库技术都可稍作改动即可应用。另一条途径是拓宽研究思路，研究全新的技术，提出新的数据管理概念。这两方面结合起来，可为21世纪数据库技术的研究开发开辟新的局面。

总之，数据库技术所面临的挑战主要体现在以下几个方面：一是环境的变化，数据库系统的应用环境由可控制的环境变成多变的异质信息集成环境和WWW环境；二是数据类型的变化，数据库中的数据的类型由结构化扩大至半结构化和多媒体数据类型；三是数据来源的变化，大量数据来源于实时动态的传感器或监测设备，数据量也因此骤增；四是数据管理要求的变化，许多新型应用需支持协同设计和工作流管理，需采用push方式进行信息发布。

传统数据库系统对新型应用不适应的部分原因是由于这些应用提出了传统研究和开发没有覆盖的功能和性能要求，也有一部分原因是因为传统的技术和系统发展过程忽略了DBMS的开放性和易用性等而造成的。例如，现有的DBMS都采用整块式结构，即每个DBMS中各功能模块构成一个难以分割的整体，没有可分离的组件或模块，无法满足许多应用对构建瘦(Lean)型或轻便(light-weight)型数据库系统的要求，用户无法选择使用DBMS提供的功能。如果要用，就得使用DBMS的全部功能，由此带来的问题是，系统运行开销大(加琐、日志、备份等)，系统的安装和维护需要专门经验。这阻碍了数据库技术的广泛应用。

5. 新的研究方向

为克服传统数据库技术和系统的局限性，适应新的应用，数据库技术的研究和发展不应仅局限于增强和提高传统DBMS的功能和性能。本节罗列出一些值得探索的研究课题。

5.1 易用性

尽管DBMS的易安装、易使用和易管理性有了很大的改善，但更多的用户仍更喜欢用文件系统。这说明DBMS的管理还是需要专门技术人员来负责，而大多数数据库用户并未经过数据库技术的专门训练，对这些用户来说，连接到数据库，找到正确的目录或数据库名空间，编写数据库的查询和更新语言是有相当难度的。显而易见，要想让数据库技术象电子表格和字处理软件那样深入大众生活，更好的数据库界面设计是先决条件。首先，数据库研究界不能指望最终用户自己来写SQL程序，用户的所有请求应有简单易懂的界面的支持。其次，系统应提供软件工具，将数据库领域的理论概念变成实际可用的技术，用户不必是理论专家，也可直观地进行数据库设计、完整性检查、系统性能调整等工作。对数据库系统管理人员这样的专家，类似的工具也是必要的。

5.2 可扩充性和组件化

前面已提到，DBMS的整块式体系结构不利于瘦型或轻便型数据库应用系统的开发，解决的办法是采用数据库组件的方式。用户可按需要选择不同功能的组件构成一些新型应用的轻便数据库解决方案。数据库组件有利于实现DBMS的模块化构建，从而提供良好的可扩充性。因为即使在全功能的DBMS中，也有可能要扩充一些模块或组件以支持特定应用所专用的功能。在这方面值得研究的内容有：研究DBMS对外部数据类型的完全支持；研究DBMS的开放体系结构以便按应用要求加入新的数据库功能；研究DBMS组件与操作系统、程序设计语言和网络基础设施等非DBMS组件的协作或集成。

5.3 数据质量和非精确查询

在广域网或因特网环境中，不同信息源的数据的质量各不相同。数据质量包括数据的时效性、完整性和一致性。如何在获得数据源的同时捕获和处理与质量有关的元数据在未来数据库应用系统中将是一个必须解决的问题。这涉及数据质量的度量及其在数据处理中的使用。举例来说，对两个数据质量差异很大的数据施行连接操作显然没有多大的意义。另一个相关的问题是非精确查询。当今的DBMS管理的是可控的封闭环境，查询所得到的结果也是精确完备的。然而，在WEB或其他大型信息源中无法也没有必要保证绝对精确。相似性查询等技术就是典型的非精确查询，但相似性技术是与具体的数据类型(如文本、图象等)密切相关的。目前尚没有将它

们关联起来研究，因此有必要研究通用的非精确性理论。

5.4 无模式数据库

众所周知，模式在传统数据库中起着举足轻重的作用，不幸的是，在许多新型应用环境中数据不再按预先定义的模式产生。例如，在 WEB 中，数据的结构是动态进化的，难以套用固定的模式。此外，随着新的数据的不断加入，人们也会发现原先设计的模式也是不完全或不一致的，无法接纳外来数据，因此，有必要研究模式管理设施，其中包括精密的数据映射设施。这些映射工具应该是说明性的且能和查询语言结合。另一研究方向是扩充现有的数据库技术用以对非结构化数据进行查询和转换。

5.5 新型事务模型

事务是支持并发用户的关键概念，新型应用环境较传统数据库环境在并发用户的支持上有很大的不同。一是每个并发事务可能长了很多，二是并发事务数可能很多，三是并发用户地理上分布很广。因此，新的事务模型应允许用户介入事务管理，根据具体应用背景定义正确性概念，允许事务嵌套。典型的做法是将正确性和隔离性分别对待，放松对正确性的要求，通过补偿/回退机制来保证数据正确，从而在事务模型中支持部分回退。这样既保证数据一致性，又保证已做的工作不会前功尽弃。在广域网或因特网环境中，2PC 会导致一个站点不能提交事务而使得系统中另一服务站点受到顾客的抱怨，这是不合理的。因此，克服传统的 2PC 协议带来的阻塞问题也是新型事务模型应解决的，提交/补偿是可行的方案之一。正确性需求与事务调度之间的关系也是一值得研究的问题。

5.6 查询优化

在未来的环境中，数据库中的数据类型会非常复杂、优化所考虑的因素会很多，从而使得传统的优化器不敷使用。首先、应针对新的数据类型进行优化，设计相应的索引方法和查询处理策略。其次，优化的标准不再仅限于降低磁盘访问次数和缩短响应时间，还需综合考虑精确性、完整性和信息成本等因素。此外，在移动、无线条件下，查询优化还需考虑带宽及电源使用等因素。

5.7 数据迁移

在分布环境中，数据迁移的成本非常高，因此，通讯线路和中间节点上高速缓存的优化使用是影响性能的重要因素。尽管数据迁移是与分布式查询优化密切相关的，但应考虑系统的整体访问模式而非单个请求的处理，另外，还必须考虑低带宽通讯线路和高负荷服务器的不对称性。

5.8 安全性

在 WEB 的开放环境中，文档的不统一性、相关信息的物理分布使得安全性保护更为困难，传统数据库系统的安全机制很大程度上依赖于模式，而新型应用中数据常是无模式的。因此，需研究新的授权模式的设计、分布式情形下授权模式的扩充、不同安全策略间的互操作性及基于证件的访问控制策略等。

5.9 数据挖掘（又称知识发现）

数据挖掘是目前发展极为迅速的一个研究领域，它综合了机器学习、统计分析和数据库技术，是为数据库中数据的决策型使用服务的。知识发现包括关联规则生成、分类、聚类、序列分析等。这些知识发现任务也可看作是数据库上的即库查询，这些查询计值涉及到在大型数据库上运行归纳机器学习算法或者统计算法。如何扩充数据库系统的功能，使之包括数据挖掘能力，是当前数据库界的一个研究方向，具体说来，就是研究简单的查询原语和新一代查询优化技术。

结束语 本文简要回顾了数据库系统的发展历史，综述了现有的数据库技术。通过对新的数据库应用举例的讨论和传统数据库技术面临的挑战的分析，列举了几个当前或 21 世纪值得注意的几个研究课题。

参 考 文 献

1 Silberschatz A, et al. Strategic directions in database systems--breaking our of the box. Computing Surveys, 1996, 28(4): 764～778

2 Silberschatz A, et al. Database systems: achievements and opportunities into 21st century. Available at: http://www.cs.stanford.edu/pub/papers/lagii.ps, 1995

3 Silberschatz, A, et al. Database systems: achievements and opportunities. Communication of ACM, 1991, 34(10): 110～120

4 Ullman J. Database theory: past and future. In: Proc. of 6th-Symposium on Principles of Database Systems, 1987. 1～10