数据库的发展史

2017年12月20日 09:01:25 qq_41397900 阅读数：5545

数据库系统的研究和开发从20世纪60年代中期开始到现在，几十年过去了，经历三代演变，取得了十分辉煌的成就：造就了C.W. Bachman（巴克曼）、E.F.Codd（考特）和J. Gray（格雷）三位图灵奖得主；发展了以数据建模和数据库管理系统（DBMS）核心技术为主，内容丰富的一门学科；带动了一个巨大的数百亿美元的软件产业。今天，随着计算机系统硬件技术的进步以及互联网技术的发展，数据库系统所管理的数据以及应用环境发生了很大的变化。其表现为数据种类越来越多、越来越复杂、数据量剧增、应用领域越来越广泛，可以说数据管理无处不需无处不在，数据库技术和系统已经成为信息基础设施的核心技术和重要基础。

数据库技术从诞生到现在，在不到半个世纪的时间里，形成了坚实的理论基础、成熟的商业产品和广泛的应用领域，吸引了越来越多的研究者加入。数据库的诞生和发展给计算机信息管理带来了一场巨大的革命。几十年来，国内外已经开发建设了成千上万个数据库，它已成为企业、部门乃至个人日常工作、生产和生活的基础设施。同时，随着应用的扩展与深入，数据库的数量和规模越来越大，数据库的研究领域也已经大大地拓广和深化了。40年间数据库领域获得了三次计算机图灵奖（C.W. Bachman、E.F. Codd和J. Gray），更加充分地说明了数据库是一个充满活力和创新精神的领域。本节沿着历史的轨迹，追溯数据库的发展历程。

背景材料：数据库著名人物

Edgar F. Codd（埃德加·考特）

Edgar Frank Codd（1923 - 2003）是Michigan大学哲学博士，IBM公司研究员，被誉为“关系数据库之父”，并因为在数据库管理系统的理论和实践方面的杰出贡献于1981年获图灵奖。1970年，Codd发表题为“大型共享数据库的关系模型”的论文，首次提出了数据库的关系模型。因为关系模型简单明了以及具有坚实的数学理论基础，所以一经推出就受到了学术界和产业界的高度重视和广泛响应，并很快成为数据库市场的主流。20世纪80年代以来，计算机厂商推出的数据库管理系统几乎都支持关系模型，数据库领域当前的研究工作大都以关系模型为基础。

C.J. Date（戴特）

Date是最早认识到Codd 在关系模型方面所做的开创性贡献的学者之一，他是关系数据库技术领域中非常著名的独立撰稿人、学者和顾问，他使得关系模型的概念普及化。他参与了IBM公司的SQL/DS和DB2两大产品的技术规划和设计。30多年来，Date 一直活跃在数据库领域中，其著作有《数据库系统导论》、《对象关系数据库基础：第三次宣言》（1998）等。

Jim Gray（吉姆·格雷）

Jim Gray毕业于Berkeley大学，先后供职于IBM公司、微软旧金山研究所。Gray曾参与主持过IMS、System R、SQL/DS、DB2等项目的开发。他在事务处理方面取得了突出的贡献，使他成为该技术领域公认的权威，他的研究成果反映在他发表的一系列论文和研究报告之中，最后结晶为一部厚厚的专著Transaction Processing: Concepts and Techniques。Gray“开创性的数据库研究”为数据库系统的应用奠定了坚实基础，并在1998年获得了计算机科学领域的最高奖项——图灵奖。

数据管理的诞生

数据库的历史可以追溯到50多年前，那时的数据管理非常简单。通过大量的分类、比较和表格绘制的机器运行数百万穿孔卡片来进行数据的处理，其运行结果在纸上打印出来或者制成新的穿孔卡片。而数据管理就是对所有这些穿孔卡片进行物理的储存和处理。

数据库系统的萌芽出现于20世纪60年代。当时计算机开始广泛地应用于数据管理，对数据的共享提出了越来越高的要求。传统的文件系统已经不能满足人们的需要。能够统一管理和共享数据的数据库管理系统（DBMS）应运而生。数据模型是数据库系统的核心和基础，各种DBMS 软件都是基于某种数据模型的。所以通常也按照数据模型的特点将传统数据库系统分成网状数据库（Network database）、层次数据库（Hierarchical database）和关系数据库（Relational database）三类。

最早出现的是网状DBMS，是美国通用电气公司Bachman等人在1961年开发成功的IDS（IntegratedData Store）。1961年通用电气公司的CharlesBachman 成功地开发出世界上第一个网状DBMS也是第一个数据库管理系统——集成数据存储（IntegratedData Store，IDS），奠定了网状数据库的基础，并在当时得到了广泛的发行和应用。

网状数据库模型对于层次和非层次结构的事物都能比较自然的模拟，在关系数据库出现之前网状DBMS要比层次DBMS用得普遍。在数据库发展史上，网状数据库占有重要地位。层次型DBMS是紧随网络型数据库而出现的。最著名最典型的层次数据库系统是IBM 公司在1968 年开发的IMS（InformationManagement System），一种适合其主机的层次数据库。这是IBM公司研制的最早的大型数据库系统程序产品。

关系数据库的由来

网状数据库和层次数据库已经很好地解决了数据的集中和共享问题，但是在数据独立性和抽象级别上仍有很大欠缺。1970年，IBM的研究员E.F.Codd博士在刊物《Communication of the ACM》上发表了一篇名为“ARelational Model of Data for Large Shared Data Banks”的论文，提出了关系模型的概念（如图7.4左图所示），奠定了关系模型的理论基础。这篇论文被普遍认为是数据库系统历史上具有划时代意义的里程碑。Codd的心愿是为数据库建立一个优美的数据模型。后来Codd又陆续发表多篇文章，论述了范式理论和衡量关系系统的12条标准，用数学理论奠定了关系数据库的基础。

关系模型有严格的数学基础，抽象级别比较高，而且简单清晰，便于理解和使用。但是当时也有人认为关系模型是理想化的数据模型，用来实现DBMS是不现实的，尤其担心关系数据库的性能难以接受，更有人视其为当时正在进行中的网状数据库规范化工作的严重威胁。为了促进对问题的理解，1974年ACM牵头组织了一次研讨会，会上开展了一场分别以Codd和Bachman为首的支持和反对关系数据库两派之间的辩论。这次著名的辩论推动了关系数据库的发展，使其最终成为现代数据库产品的主流。

1970年关系模型建立之后，IBM公司在San Jose实验室增加了更多的研究人员研究这个项目，这个项目就是著名的System R。其目标是论证一个全功能关系DBMS的可行性。该项目结束于1979年，完成了第一个实现SQL的DBMS。

同时，1973年加州大学伯克利分校的Michael Stonebraker和EugeneWong利用System R已发布的信息开发自己的关系数据库系统Ingres。他们开发的Ingres项目最后被Oracle公司、Ingres公司以及硅谷的其他厂商所商品化。后来，System R和Ingres系统双双获得ACM的1988年“软件系统奖”。1976年Honeywell（霍尼韦尔）公司开发了第一个商用关系数据库系统–MulticsRelational Data Store（MRDS）。关系型数据库系统以关系代数为坚实的理论基础，经过几十年的发展和实际应用，技术越来越成熟和完善。其代表产品有Oracle、IBM公司的DB2、微软公司的MS SQLServer以及Informix、ADABASD等等。

关系代数

关系代数（Relation Algebra）是一种抽象的查询语言，用对关系的运算来表达查询，作为研究关系数据语言的数学工具。关系代数的运算对象是关系，运算结果亦为关系。关系代数用到的运算符包括四类：集合运算符、专门的关系运算符、算术比较符和逻辑运算符。比较运算符和逻辑运算符是用来辅助专门的关系运算符进行操作的，所以关系代数的运算按照运算符的不同主要分为传统的集合运算和专门的关系运算两类。

传统的集合运算是二目运算，包括并（Union）、交（Intersection Referential Integrity）、差（Difference）、广义笛卡尔积（Extended Cartesian Product）四种运算。专门的关系运算包括选择（Selection）、投影（Projection）、连接（Join）和除（Division）。

结构化查询语言

1974年，IBM的Ray Boyce和DonChamberlin将Codd关系数据库的12条准则的数学定义以简单的关键字语法表现出来，里程碑式地提出了SQL（StructuredQuery Language）语言。SQL语言的功能包括查询、操纵、定义和控制，是一个综合的、通用的关系数据库语言，同时又是一种高度非过程化的语言，只要求用户指出做什么而不需要指出怎么做。SQL集成实现了数据库生命周期中的全部操作，提供了与关系数据库进行交互的方法。SQL可以与标准的编程语言一起工作。自产生之日起，SQL语言便成了检验关系数据库的试金石，而SQL语言标准的每一次变更都指导着关系数据库产品的发展方向。1986年，ANSI把SQL作为关系数据库语言的美国标准，同年公布了标准SQL文本。

1976年IBM的Codd发表了一篇里程碑的论文“R系统：数据库关系理论”，介绍了关系数据库理论和查询语言SQL。随后，Oracle的创始人LarryEllison非常仔细地阅读了这篇文章，敏锐意识到在这个研究基础上可以开发商用软件系统，而当时大多数人认为关系数据库不会有商业价值。几个月后，Ellison他们就开发了Oracle 1.0。今天，Oracle数据库的最新版本为11g，包括几乎所有的世界500强企业都在使用Oracle的数据库。

Larry Ellison

Larry Ellison是世界上最大数据库软件公司Oracle（甲骨文）的CEO，2008年仅次于Bill Gates的世界第二富豪。他的产品遍布全世界，似乎谁都无法离开他：当我们从自动提款机上取钱、在航空公司预定航班，或将家中电视接入Internet等等。毫无疑问，这时的你就在和Oracle打交道。

Larry Ellison是美国犹太人，俄罗斯移民，1944年出生在曼哈顿。1977年6月Larry Ellison、Bob Miner和Edward Oates合伙出资2 000美元成立了软件开发研究公司。

Larry一向怀疑所谓“传统的智慧”，不相信权威的观点，特别是那些人云亦云的权威。对他来说，事情必须合理才行。正是这种思考方式在企业经营上非常有价值。他始终相信较早占领大块市场份额是最主要的：“当市场已建立好，你知道百事可乐要花多少钱才能夺得可口可乐1%的市场？非常非常昂贵！”Oracle公司连续12 年销售额每年翻一番，成为世界上第二大软件公司，同时Larry也成了硅谷首富。正像一位硅谷资深人士评论的那样：拥有普通技术和一流市场能力的公司总是能打败拥有一流技术而只有普通市场能力的公司。

面向对象数据库

随着信息技术和市场的发展，人们发现关系型数据库系统虽然技术很成熟，但其局限性也是显而易见的：它能很好地处理所谓的“表格型数据”，却对技术界出现的越来越多的复杂类型的数据无能为力。20世纪90年代以后，技术界一直在研究和寻求新型数据库系统。但在什么是新型数据库系统的发展方向的问题上，产业界一度是相当困惑的。受当时技术风潮的影响，在相当一段时间内，人们把大量的精力花在研究“面向对象的数据库系统（Object OrientedDatabase）”或简称“OO数据库系统”。

然而，数年的发展表明，面向对象的关系型数据库系统产品的市场发展的情况并不理想。理论上的完美性并没有带来

市场的热烈反应。其不成功的主要原因在于，这种数据库产品的主要设计思想是企图用新型数据库系统来取代现有的数据库系统。这对许多已经运用数据库系统多年并积累了大量工作数据的客户，尤其是对大客户来说，是无法承受新旧数据间的转换而带来的巨大工作量及巨额开支的。另外，面向对象的关系型数据库系统使查询语言变得极其复杂，从而使得无论是数据库的开发商家还是应用客户都视其复杂的应用技术为畏途。

数据管理的变革：决策支持系统和数据仓库

20世纪60年代后期出现了一种新型数据库软件：决策支持系统（Decision Support System，DSS），其目的是让管理者在决策过程中更有效地利用数据信息。

决策支持系统是辅助决策者通过数据、模型和知识，以人机交互方式进行半结构化或非结构化决策的计算机应用系统。它是管理信息系统向更高一级发展而产生的先进信息管理系统。它为决策者提供分析问题、建立模型、模拟决策过程和方案的环境，调用各种信息资源和分析工具，帮助决策者提高决策水平和质量。

1988年，为解决企业集成问题，IBM公司的研究员BarryDevlin和Paul Murphy创造性的提出了一个新的术语——数据仓库（DataWarehouse）。之后，IT厂商开始构建实验性的数据仓库。1991年，W. H.Inmon出版了一本“如何构建数据仓库”的书，使得数据仓库真正开始应用。

数据仓库是决策支持系统和联机分析应用数据源的结构化数据环境，是一个面向主题的（SubjectOriented）、集成的（Integrated）、相对稳定的（Non-Volatile）、反映历史变化（TimeVariant）的数据集合，用于支持管理决策（DecisionMaking Support）。

数据挖掘和商务智能

数据仓库和数据挖掘是信息领域中近年来迅速发展起来的数据库方面的新技术和新应用。其目的是充分利用已有的数据资源，把数据转换为信息，从中挖掘出知识，提炼成智慧，最终创造出效益。数据仓库和数据分析、数据挖掘的研究和应用，需要把数据库技术、统计分析技术、人工智能、模式识别、高性能计算、神经网络和数据可视化等技术相结合。

随着数据仓库、联机分析技术的发展和成熟，商务智能的框架基本形成，但真正给商务智能赋予“智能”生命的是它的下一个产业链——数据挖掘。

数据挖掘是指通过分析大量的数据来揭示数据之间隐藏的关系、模式和趋势，从而为决策者提供新的知识。之所以称之为“挖掘”，是比喻在海量数据中寻找知识，就像从沙里淘金一样困难。

数据挖掘是数据量快速增长的直接产物。20世纪80年代，它曾一度被专业人士称之为“基于数据库的知识发现”（Knowledge Discovery in Database，KDD）。数据仓库产生以后，如“巧妇”走进了“米仓”，数据挖掘如虎添翼，在实业界不断产生化腐朽为神奇的故事，其中，最为脍炙人口的当属啤酒和尿布。

话说Wal-Mart（沃尔玛）拥有世界上最大的数据仓库，在一次购物篮分析之后，研究人员发现跟尿布一起搭配购买最多的商品竟是风马牛不相及的啤酒！这是对历史数据进行“挖掘”和深层次分析的结果，反映的数据层面的规律。但这是一个有用的知识吗？沃尔玛的分析人员也不敢妄下结论。经过大量的跟踪调查，终于发现事出有因：在美国，一些年轻的父亲经常要被妻子“派”到超市去购买婴儿尿布，有30％到40％的新生爸爸会顺便买点啤酒犒劳自己。沃尔玛随后

对啤酒和尿布进行了捆绑销售，不出意料之外，销售量双双增加。这种点“数”成金的能力，是商务智能真正的“灵魂”和魅力所在。

1989年，可谓数据挖掘技术兴起的元年。这一年，图灵奖的主办单位计算机协会下属的知识发现和数据挖掘小组（SIGKDD）举办了第一届学术年会，出版了专门期刊。此后，数据挖掘被一直追捧，成为炙手可热的话题，并如火如荼的发展。数据挖掘的内容将在本书第8章详细介绍。

也正是1989年，著名的高德纳IT咨询公司（GartnerGroup）为业界提出了商务智能的概念和定义。商务智能（Business Intelligent，BI）指的是一系列以数据为支持、辅助商业决策的技术和方法。商务智能在这个时候完全破茧而出，不是历史的巧合，因为正是数据挖掘这种新技术的出现，商务智能才真正有了“智能”内涵，这也标志着其完整产业链的形成。

商务智能指利用数据仓库、数据挖掘技术对客户数据进行系统地储存和管理，并通过各种数据统计分析工具对客户数据进行分析，提供各种分析报告，如客户价值评价、客户满意度评价、服务质量评价、营销效果评价、未来市场需求等，为企业的各种经营活动提供决策信息。商务智能也是企业利用现代信息技术收集、管理和分析结构化和非结构化的商务数据和信息，创造和累积商务知识和见解，改善商务决策水平，采取有效的商务行动，完善各种商务流程，提升各方面商务绩效，增强综合竞争力的智慧和能力。

数据库领域的国际会议

ACM SIGMOD /PODS

SIGMOD和另外两大数据库会议VLDB、ICDE构成了数据库领域的三个顶级会议。相对而言，SIGMOD比另外两个会议的含金量更高，被录取的难度更大。

ACM SIGMOD数据管理国际会议是由美国计算机协会（ACM）数据管理专业委员会（SIGMOD）发起、在数据库领域具有最高学术地位的国际性学术会议。SIGMOD的前身是SIGFIDET，SIGFIDET成立于1969年，而在1970年的9月，它转变为了SIG。4年后，于1974年，SIG决定改名为SIGMOD（Special Interest Group on Management of Data）。

SIGMOD会议于1974年在美国Michigan首次举办，PODS会议于1982年在美国Los Angeles首次召开。这两个会议于1991年在美国Denver首次联合召开，这次联合举办会议的尝试取得了巨大成功，并大大鼓舞了整个数据库界理论和系统的结合。之后SIGMOD会议和PODS会议都是同时举行。由于ACM SIGMOD会议是由SIGMOD/PODS两部分组成的，所以它的主体同时包含了SIGMOD 和PODS的内容：Research Session、Industrial Sessions、Tutorials、Demonstrations、Keynote Talks、Panel Sessions、Research Plenary Sessions、workshops以及PODS Sessions。它覆盖了当前数据库和信息系统研究中存在的前言问题，而数据库仍是 21 世纪的新兴应用技术的基石之一。SIGMOD会议已经成为了数据库管理领域最杰出的研究和发展成果的实时传播场所。

几十年来，ACM SIGMOD/PODS会议已经向世人证明了它是世界顶级的数据管理会议。会议的目的是在全球范围内为数据库领域的研究者、开发者以及用户提供一个探索最新学术思想和研究方法、交流开发技巧、工具以及经验的平台，引导和促进数据库学科的发展。SIGMOD会议每年都在不同的地方召开，但是几乎都在北美国家，而在这几十年中仅有两次是在北美以外的国家召开，其中2007年6月11日至6月14日，第26届ACM SIGMOD国际数据管理学术会议在北京国际会议中心举行。这次会议受到国家自然科学基金委员会国际合作交流项目资助，由ACM SIGMOD主办、清华大学承办。这是该会议第一次在亚洲举行、也是第二次在北美以外的国家举行，这也从侧面反映出了中国科研水平在不断提高。

VLDB

国际超大型数据库会议（International Conference on Very Large Data Bases，VLDB）是一个专门从事超大规模数据库管理理论、方法和应用研究的专业性学术机构，它涉及的内容也很丰富，包括研究及应用的诸多方面，基本上能够较全面地反映当前数据库研究的前沿方向、工业界的最新技术以及各国的研发水平。1975年，以美籍华裔科学家肖开美教授（Dave Hsiao）为首的一批数据库学者发起

组织了第一届VLDB会议。此后每年召开一次，已成为是数据库领域中最主要、规模最大的国际学术会议之一。

ICDE

数据工程国际学术会议（International Conference on Data Engineering ICDE）是由IEEE计算机数据工程技术学会（TCDE）主办的数据库领域的最高级别的国际性会议之一。会议产生出版季刊数据工程通报（Data Engineering Bulletin）。TCDE致力于研究数据在信息系统的设计、实现与管理中的作用，面向的主要问题包括数据库设计、数据处理、数据库存储与操纵语言、数据采集的策略与机制、数据库的安全性与完整性控制、数据库的工程应用以及分布式系统。

NDBC

NDBC（National Database Conference）即全国数据库学术会议，始于1977年，至今已成功举办了二十多届。 中国计算机学会数据库专业委员会（China Computer Federation Database Society，CCFDBS）是中国计算机学会领导下的数据库学术组织，于1999 年8月24日在兰州大学召开的第十六届全国数据库学术会议上正式成立。自CCFDBS成立以来，数据库专委会致力于办好NDBC这一传统的数据库盛会，为中国内地、香港、台湾、澳门和海外华裔数据库研究者、开发者和用户提供一个中华数据库论坛。每届NDBC会议都会邀请国内外的著名专家到会作特邀报告，还有主流数据库厂商到会展示他们的最新技术。与此同时，NDBC将逐步与国际接轨，成为亚太地区乃至世界性的有影响力的学术会议。