

大数据的存储渊源

万 赞

美国休斯敦大学

关键词：关系数据库 非结构化数据库

随着大数据时代的来临，流行长达 20 年的传统关系数据库已经失去了信息系统的标准数据存储方式的地位，数据的存储方式呈现出多元化趋势。

IBM的创新困境

关系数据库原理最早是由埃德加·科德 (Edgar Codd) 在 1970 年提出的。科德在 IBM 公司做过程序员。1963 年他在美国密歇根大学攻读博士学位。科德获得计算机博士学位时已经 42 岁。大器晚成的他于 1967 年又回到 IBM 做研究工作。3 年后，他在《美国计算机学会通讯》(CACM) 上发表了关系数据库的开山之作“大型共享数据库数据的关系模型” (A Relational Model of Data for Large Shared Data Banks)。

在科德发表这篇论文之前，IBM 已经在 20 世纪 50 年代末通过编译器和高级编程语言把软件的编程环境从具体的计算机硬件中分离出来，然后在 60 年代中期通过 S/360 操作系统将软件进一步分成操作系统和应用软

件。科德的关系数据库则是在此基础上将数据操作从具体的计算机软件环境和物理存储模式中独立出来。这样无论是程序员还是普通的数据库使用人员，只需要了解数据间的逻辑关系，就可以通过数据库操作语言来管理和分析数据。

不过，在科德的文章发表之前，IBM 刚好向市场上推出了层次型数据库 IMS 系统。尽管 IMS 系统在设计理念和效率上不及关系数据库系统，但为了收回投资成本，IBM 决定尽量推迟关系数据库的上市，关系数据库只作为未来系统总开发项目的子项目“系统 R”进行开发。科德没能直接管理该项目。在设计“系统 R”的操作语言 SQL 的过程中，IBM 也没采纳科德提出的关系化 Alpha 语言的许多优点。这些决策的失误使得 10 年后 IBM 在关系数据库方面失去了市场先机。

新技术的传播

让 IBM 始料未及的是，科德的那篇论文和“系统 R”的研

发人员陆续发表的相关论文很快引起了一批计算机科研人员的兴趣，其中包括加州大学伯克利分校的尤金·黄 (Eugene Wong) 和迈克尔·斯通布雷克 (Michael Stonebraker)。这两位从美国国家科学基金会 (NSF) 以及美国空军和陆军等军方机构申请到经费，从 1973 年开始基于 DEC 小型机的 UNIX 系统的关系数据库 INGRES 的研发工作。该系统成为第一套全面支持 ACID 特性的关系型数据库。

ACID 是指：传统关系数据库在数据操作过程中必须同时保证操作的原子性 (atomicity)、一致性 (consistency)、隔离性 (isolation) 和持久性 (durability)。这四大特性是关系数据库支持银行转账以及其它不可分割的数据库交易逻辑操作的基本要求。INGRES 通过相对简单的数据库操作语言使得 ACID 在底层自动实现。另外，INGRES 的数据库操作语言 QUEL 采用了科德的 Alpha 设计理念，在数据结构和组织上采用了 B 树和主键等创新，这些都成为后来关系数据库

的标准设计模式。

由于INGRES项目与加州大学伯克利分校的UNIX项目一样,对使用者只收取很少的名义版权费用,所以从1974年第一版上市到1985年项目正式结束的10年间,INGRES软件不但被美国几百所大学和研究机构使用,而且成功衍生出数个商业版关系数据库产品。参与该项目的加州大学伯克利分校的学生中也出现了一大批数据库专家,其中包括赛贝斯(Sybase)公司的创始人之一罗伯特·爱泼斯坦(Robert Epstein)。斯通布雷克和黄两人则在1980年与其他合作者一起创立了RTI(Relational Technology, Incorporated)公司,开始了INGRES数据库的商业推广。

就在斯通布雷克和合作者开发和推广关系型数据库的同时,加州出现了另外一个发现关系数据库潜力的小公司RSI(Relational System Incorporated)。1976年成立的RSI的三个创始人是拉里·埃里森(Larry Ellison)、罗伯特·迈纳(Robert Miner)和爱德华·奥茨(Edward Oates)。他们曾经在加州生产数据存储设备的公司一起工作过。迈纳是一个非常经验丰富的数据库专家,他最先发现了科德的关系数据库论文,并介绍给埃里森和奥茨。

埃里森等三人经过一番研究,发现关系数据库在设计思想上远远超越当时流行的其它数据库软件,于是他们决定开发自己的关系数据库产品来抢占市

场。恰好此时,美国中央情报局(CIA)正试图寻找一家公司来开发关系数据库。科德的文章一出现,中央情报局就试图说服IBM开发关系数据库软件,这样中央情报局的工作人员就可以通过相对简单的SQL查询语言而不是编写程序来分析情报资料了。但是IBM迟迟不肯满足他们的需求。中央情报局的项目联系人曾经与迈纳合作过,而迈纳等人不但正准备开发关系数据库,而且是在PDP-11小型机平台上开发,这恰好满足中央情报局将小型计算机安装到侦查飞机或者窃听潜艇的需求。于是迈纳等三人成立不久的RSI公司非常幸运地迎来了第一个客户。后来,此项目在中央情报局的代号成为RSI公司的新名字,这就是甲骨文(Oracle)系统。

商业意味浓厚的甲骨文系统和学术背景浓厚的INGRES系统就此成为关系数据库发展历史上的主要竞争对手。

关系数据库时代

进入20世纪80年代,关系数据库为越来越多的用户所重视。RSI公司凭借着超强的市场营销团队和可靠的技术后台在新市场上取得了先机。1983年,RSI公司更名为甲骨文。1984年,其销售额翻了一番,达到1270万美元。而此时,斯通布雷克等人的RTI公司凭借着INGRES项目在业界的名声,其销售额在同

一年提高了3倍,达到900万美元,大有取代甲骨文市场领先地位的势头。

埃里森发现,RTI在技术方面处于领先地位是因为他们拥有加州大学伯克利分校最优秀的计算机人才。于是他也开始花重金从加州理工大学、斯坦福大学等名牌大学招揽人才。这一招果然奏效。公司花费半年时间都没有解决的一个技术难题被新招募来的程序员用了一个周末就解决了。而斯通布雷克除了不善于市场营销外,还犯了一个严重的决策性失误,即当美国国家标准局准备把IBM的SQL作为关系数据库的标准操作语言时,斯通布雷克出于学究体面,没能出面为INGRES的QUEL语言作为竞争标准提出辩护。结果当SQL被选定后,RTI不得不花费4年多时间重新开发基于SQL语言的INGRES新版本。最终甲骨文成功甩掉了RTI的市场追击,后者则被卖给了王嘉廉的国际联合电脑公司(CA)。

在甲骨文等公司一路高歌进军关系数据库市场以后,IBM才开始加速推动“系统R”的开发,在1983年将其命名为“DB2”后推向市场。DB2起初只能在IBM大型机上使用,而甲骨文系统可以在所有小型机、工作站,甚至微软DOS系统上运行。再加上当时分布式计算烽火燎原的势头,IBM已经无暇顾及关系数据库市场。所以,DB2虽然有IBM这一强大的后盾,但在市场

占有率上一直落后于甲骨文。

80年代中期,整个关系数据库市场上有50多家大大小小的公司。由爱泼斯坦和马克·霍夫曼(Mark Hoffman)等人于1984年根据INGRES创立的赛贝思公司一度对甲骨文公司构成威胁。霍夫曼等人根据摩尔定律把新数据库的平台定位在价格更低的升阳(SUN)工作站服务器,这使得赛贝思数据库恰好进入方兴未艾的以客户/服务架构为主的分布式计算市场上。赛贝思数据库根据这一市场特点设计出存储过程(stored procedure)功能,使得复杂的系统逻辑可以设计成调用对象,然后封装存储在服务器端,客户端用户随时可以通过简单的指令使用。此外,该系统还实现了数据和交易操作的参照完整性(referential integrity)和两阶段提交(two-phase commit)等更为灵活的数据操作功能。

由于进入市场的时间较晚,赛贝思的市场占有率一直低于甲骨文。1986年,赛贝思与微软达成协议,以赛贝思的关系数据库为基础帮助微软为IBM的OS/2开发一款用于低端产品市场的关系数据库。后来IBM和微软在OS/2上分道扬镳,赛贝思与微软继续合作开发了这一数据库,这就是后来的微软SQL Server数据库。赛贝思在技术方面的优势一直到1992年甲骨文推出7.0版才宣告结束。

在甲骨文盘踞关系数据库市场的90年代中期,开源软件运

动产生的Linux和MySQL对甲骨文造成了最大的威胁。这两套来自北欧的开源软件成为谷歌、维基百科、亚马逊、脸谱、推特、雅虎等所有采用集群技术的互联网公司的单机操作系统和关系数据库。随着云计算技术的推广,越来越多的公司开始采用Linux和MySQL,这使得MySQL的市场占有率直逼甲骨文。最后甲骨文公司不得不动用美国联邦政府的力量对欧盟施加压力,在2009年通过收购升阳公司将MySQL收入囊中。

关系数据库技术大量普及的直接结果就是大量交易数据的积累。1988年,IBM的研究人员第一次提出了数据仓库(information warehouse)的概念,并预测在不久的将来终端用户需要方便有效地分析企业积累的大量交易数据,该预测在90年代成为现实。

1992年,Prism Solutions咨询公司的创始人比尔·恩门(Bill Inmon)出版了《建立数据仓库》一书,提出以第三范式为基础的搭建在关系数据库之上的“企业信息工厂”概念,被业界称为由上向下的数据仓库开发模式。与此同时,前施乐公司帕洛阿托研究员、红砖系统公司创始人拉尔夫·金博尔(Ralph Kimball)则提出了抛开关系数据库模式,用“事实表”加“维度表”的星型模式搭建企业各个部门需要的数据超市,然后通过合并相同的维度表,形成维度表总线矩阵,由下向上形成企业数据仓库的

开发方式。因为金博尔的架构更有利于用户理解和分析数据,所以更受欢迎。

数据仓库的流行使得传统关系数据库以行处理为基础的存储索引结构逐渐受到质疑。因为事实表动辄上百列数据与用户使用事实表时有时只需要读取一行中的几个相关列数据产生了读取效率上的矛盾。于是,以列存储为基础组织数据库架构的思想和相关软件开始出现。此时已在美国麻省理工学院的斯通布雷克再次成为这一新趋势的先锋,他主导的C-Store项目是最早的开源列式数据库之一。斯通布雷克于2005年成立了Vertica公司进行商业化推广。赛贝思公司也推出了相应的IQ列式数据库。德国企业资源计划(enterprise resource planning, ERP)软件巨头思爱普(SAP)则通过整合其收购的各种技术和开源软件推出了列式数据库HANA系统,并在2012年收购了赛贝思,成为列式数据库技术的主导公司。

在数据仓库推动列式关系数据库产生的同期,在线联机事务处理(OLTP)催生了内存数据库的出现。随着90年代电子商务的发展,网络交易量飞速增长,企业需要对这些不断增长的实时交易数据进行高速存取。传统的磁盘存取方式造成高达75%的效率损耗。而随着内存成本的不断下降,通过内存来存储整个数据库成为现实。这一趋势推动了内存数据库的发展。

向非结构化发展

如果说 20 世纪 90 年代数据仓库的出现“破坏”了传统关系数据库的皮毛,互联网的出现则动摇了其根基。

1996 年 3 月,谷歌的两个创始人拉里·佩奇和谢尔盖·布林在斯坦福大学的宿舍里将收集到的 1500 多万条互联网网页的信息通过自己设计的文件格式进行了存储、索引和压缩,从而满足有效读取和分析的需要。由于缺少经费,他们购买了很多廉价的二手硬盘作为存储设备,并为文件系统设计了很大的冗余和很强的自调修复功能,以应对这些硬盘的高损耗。这一系统演变成后来的谷歌文件系统(GFS)以及大表(bigtable)非结构化数据库。谷歌公司成立后,经费充足,他们用内存来存储谷歌的索引数据库,大大提高了系统的反应速度。谷歌的这些创新开启了大数据时代数据存储的非结构化趋势。

非结构化数据库与传统关系数据库最重要的区别是前者减弱了数据库的 ACID 特性。虽然强调 ACID 能够保证数据库的可靠性,但是维持 ACID 特性会使海量数据存储不必要的运营成本越来越高,尤其像一条微博消息的转发数目这样的非关键数据。另外,在很多情况下,在分布式计算中很难保证 ACID 特性。比如谷歌文件系统通过冗余备份的方式来补偿硬盘出错导致数据丢失的问题。当系统需要对分布

在不同地理区域的备份进行写操作时,除非采取强制锁定,否则无法保证所有的备份都同时更新,而频繁的锁定方式显然严重影响系统效率。针对这一现象,在 2000 年分布式计算原则研讨会上,加州大学伯克利分校的里克·布鲁尔(Eric Brewer)提出了布鲁尔猜想,即分布式数据库系统无法同时满足数据的一致性、可用性和分割容忍性(partition tolerance)的要求。用户只能选择其中两项来同时实现。2002 年,这一猜想被麻省理工学院的两名计算机科学家证明,成为帽子定理(CAP Theorem)。根据帽子定理,许多非结构化数据库采取了满足可用性和分割容忍性,然后实现最终一致性(eventual consistency)的分布式存储方式。

促进非结构化的一个更深层的因素是,以摩尔定律为发展方向的英特尔架构在微处理器单核方面已经难以突破 3.5 兆赫运行频率的物理极限,所以只有通过水平扩展内核数量和企业的分布式集群发展来进一步提高数据处理速度。多核微处理器和集群技术使得数据库的分割存储成为必然。当分割容忍成为先决条件,可用性成为用户的基本期待时,帽子定理决定了传统数据库的 ACID 一致性必然会被最终一致性所取代。

非结构化趋势和谷歌架构的广泛采用也有着密切的关系。2004 年,谷歌的研发人员将谷歌文件系统和映射化简(MapRe-

duce)的理论架构以论文的形式发表后,引起了很多人的关注,其中包括维基百科的数据库程序员道格·卡丁(Doug Cutting)。卡丁通过与雅虎的合作,用了两年多时间,按图索骥,开发出小象系统(Hadoop)。2008 年雅虎成功地利用小象系统来有效管理雅虎搜索引擎的一万多台 Linux 集群,使得后者以之前 33 倍的速度搜索和索引所有互联网网页。雅虎在小象系统上的成功和以开源软件方式将其推出的举动引起了一大批互联网公司的兴趣。之后不久,亚马逊也采用了小象系统作为其弹性云的管理系统。微软则放弃了已经购买的搜索技术平台,将小象系统应用到它的搜索引擎必应(Bing)上。随后包括脸谱、推特、易贝等在内的一大批互联网公司纷纷采用了小象平台,将其应用到不同的计算服务需求中。

小象系统在众多互联网公司的应用推动了非结构化数据库 NoSQL 的发展。NoSQL 数据库通常具有支持数据库的分割、提供最终一致性支持、不需要用户提前确定各种数据表架构等特点。这些特点为网络数据存取和用户快速搜索并分析数据提供了有效的支持。比如当管理数万台服务器的公司需要保存所有访问过这些服务器的 IP 地址时,显然无法预知每一台服务器可能需要存放 IP 地址的列数,所以也无法提前设定传统的关系数据表行列架构。但是 NoSQL 数据库

通过列式存储、数据压缩和分布式分割存储的方式可以有效地处理这类信息。2005 年到 2010 年间出现的绝大多数 NoSQL 数据库都是以开源软件形式安装在小象系统或者谷歌文件平台上的,例如源自脸谱邮箱查询的分布式数据库 Cassandra,仿照谷歌大表设计的用来进行海量自然语言搜索的 HBase,为开发类似谷歌的平台即服务而产生的全面支持分割的 MongoDB,与 MongoDB

类似的 CouchDB,以及强调内存存储特色的 Redis 等等。NoSQL 运动方兴未艾,尽管有包括斯通布雷克在内的不少专家对其前景表示怀疑,但这类数据库在大数据时代已经扮演了重要角色。

我们有理由相信未来数据库的发展将呈现多元化趋势, NoSQL 不会全面取代传统关系数据库。业界会进一步推动 NoSQL 标准的统一和类似于 SQL 的标准非结构化数据操作语言的发

展,脸谱开发的 HiveQL 正在朝这一方向努力。从目前来看,在大数据时代,不同数据库结构的最优组合将是一个企业最好的数据存储方案。■



万 赞

美国休斯敦大学维多利亚校区副教授。主要研究方向为电子商务和互联网应用。

wany@uhv.edu

CCF绵阳会员活动中心成立

杜子德主持成立 赵强任主席

4月15日,CCF绵阳会员活动中心成立,秘书长杜子德主持了成立仪式。绵阳市副秘书长赵德钧、绵阳市经信委副主任勾承宽、绵阳市科技局副局长李海、绵阳市信息化专家委员会主任吴志杰以及来自中国工程物理研究院、空气动力研究与发展中心、西南科技大学、绵阳师范学院等相关单位的领导、CCF会员及相关从业人员共80余人出席了大会。

会上,杜子德向CCF绵阳首任主席、中国物理工程研究院信息化总师赵强研究员授旗,向CCF绵阳执行委员会成员颁发了聘书。赵德钧、勾承宽、李海对CCF绵阳的成立表示祝贺,希望CCF投入更多的资源支持绵阳市计算机产业的发展,同时介绍了绵阳市软件产业园的建设与发展前景及优惠政策。

CCF绵阳将利用自身资源,积极营造和建立开放交流与资源共享机制,广泛开展学术交流与合作,努力为绵阳地区及绵阳市广大计算机科技工作者提供学术互通的平台,为提高计算机技术水平、促进产业化发展起到积极的推动作用。

成立大会后,CCF绵阳举办了第一次会员活动,赵强和CCF绵阳副主席、西南科技大学计算机学院院长韩永国分别作了有关“国防军工信息化发展现状”和“西南科大计算机学院情况介绍”的报告,并与大家讨论了今后的活动计划。



杜子德(左一)与CCF绵阳会员活动中心执委成员合影