

数据库管理系统内部结构 及其 C 语言实现

唐常杰 著
张天庆 周 韬 魏志毅

- HBase 全部 C 语言源程序剖析
- 数据库管理系统实现技术
- 集成环境界面开发技术
- 大程序设计技术



电子科技大学出版社

〔川〕新登字 016 号

数据库管理系统内部结构 及其 C 语言实现

唐常杰

张天庆 周 韬 魏志毅 著

*

电子科技大学出版社出版
(成都建设北路二段四号) 邮编 610054

四川峨影印刷厂印刷

新华书店经销

*

开本 787×1092 1/16 印张 24.75 字数 602 千字
版次 1995 年 9 月第一版 印次 1996 年 10 月第二次印刷

印数 5001—6000 册

ISBN 7-81043-257-5/TP·89

定价:28.00 元

内 容 提 要

本书以作者开发的 HBase 为例,详细地介绍了数据库管理系统(DBMS)的内部结构及其 C 语言实现技术。HBase 兼容于 dBase,某些地方优于 dBase 3(例如支持下拉式菜单,鼠标,Alt 热键、支持 C 语言作为 HBase/dBase 的宿主语言等等)。书中列出了用 Borland C++ 3.1 编写的 HBase 1.5 的全部源程序,并有详细的中文注释。对 DBMS 的内部结构进行了多角度的讨论,介绍了主要思想、一般理论、数据结构、算法、调用树和特殊技巧,通过完整的程序清单,本书还附带地介绍了集成环境界面的开发技术和用 C 语言组织大程序的技术。书中列出的通用下拉式选单、鼠标库、编辑器、B-树、计算器、日历、备份例程、以及众多工具函数源程序都可视为用 C 语言表达的知识财富。本书可作为广大计算机工作者、数据库管理员和有关科研人员的参考书,也可作计算机专业的大专生,本科生,研究生学习 DBMS 的实现技术、解释器实现技术、大程序编制技术和集成环境开发技术的教材或实验室用书。

序 言

下一世纪的数据库系统将采用什么模型、它有什么特点?

还在 90 年代初期,勇于创新的理论家们就开始自问自答。上百种理论模型被提出,几十种实验系统被研究。深谋远虑的软件企业家们一方面把技术上成熟的关系数据库系统琢磨成表现公司特色的艺术品,一方面物色新型数据库管理系统的实验原型、搜罗活跃于新方向的人才,准备着一场数据库领域的跨世纪竞争。

在这方面,中国落后了。

在数据库的理论研究、应用研究和系统实现三个方向中,我国在数据库管理系统的系统实现技术方面的差距更大一些。要追赶上去,需要一大批熟悉数据库管理系统内部结构及其实现技术的人才,年轻一代需要培养,中年一代需要再学习。

目前,在数据库应用和理论方面的专著和论文已经远远超出了个人的阅读能力。使学习理论和应用的读者感到困惑的不再是找不到文献,而是如何从书山万海中筛选适读材料。但是,关于数据库管理系统实现技术的书籍还为数不多,由于软件业者的商业利益和技术保密等原因,细致地用程序语言来描述数据库内部结构和实现技术细节的书籍更为鲜见。读者呼唤着这样的书籍。

本书的宗旨正是响应这一呼唤,以作者开发的 HBase 为例,介绍了 DBMS 的内部结构及其 C 语言实现技术。与那些花了数千个人年工作量的 DBMS 庞然大物相比,HBase 只不过是 个小不点,但却具有 DBMS 的主要结构,以个人自学方式能在不太长的时间内得其要略。

HBase 兼容于 dBase,某些方面优于 dBase 3。本书列出了 HBase 1.5 版的全部 C 语言源程序,并附有详细的中文注释。同时,HBase 1.51 版的可执行文件和全部源程序将以软件的形式出版,供读者选购,其中将包含本书付印之后所作的改进、可能出现的勘误和锦上添花的模块。读者可与有关方面联系,以便获得最新的版本。

只要读者对 DBMS 和 dBase/FoxBase 有一定了解,粗通 C 语言,就能通过本书,循序渐进,学习、掌握 DBMS 的主要内部结构和实现技术,同时把 C 语言技术提高,开发出有自己特色的 DBMS。

目前,国际计算机科学的书籍和文献中,流行一种新的表达形式,即用程序或准程序——计算机科学的语言来表达软件理论和技术思想。这反映了时代对跨世纪计算机科学人才的新要求,理论和实践的进一步结合。理论家不再是只画框图或作算法分析,实干家也将

运用更多的理论。

有鉴于此,作为一个尝试,本书采用了夹议夹叙的形式。“议”——对 DBMS 的每一要点,介绍一般原理,评议算法优劣,指出技术要点和设计思想,作为一种探索性的方法,我们还给出函数的调用树;“叙”——通过 C 语言源程序,附之以详细的中文注释,描述 DBMS 的内部结构和实现细节。列出的程序都是实实在在可运行的程序,把程序中详细的中文注释集中在一起,就是详细的算法。按这一方式,读者会发现,学理论时不觉抽象、空泛,学技术时不觉盲目零乱。理论立即付诸于实践。

本书的组织如下:

第一章对数据库管理系统作了一个简要的概述,分析了 DBMS 实现中的要点和难点。在第二章中,讨论了数据库管理系统体系结构,简介了 HBase 的模块及调用关系,给出了 HBase 的主程序清单。第三章介绍了 HBase 的用户界面,着重介绍了用控制与数据分离的方式实现的集成环境界面、鼠标支持和联机求助等等技术。

第四章和第五章是本书中较为重要的部分,分别讨论了存储结构、文件输入输出管理和 B-树索引,值得花较多的时间和精力去掌握这一部分内容。

第六章数据定义语言和第七章数据操纵语言实现了 Create、Modify、Edit、Append、Insert、Delete、Select、Project、Join、Browse 和 Sort 等 HBase、dBase 命令。

第八章用户输入与输出,着重介绍了 List 和 Display 的实现技术。第九章讨论了数学表达式处理与统计计算模块。

第十章是重点和难点,详细地介绍了 HBase 命令解释器,其中包括:解释器的数据结构、解释器专用类型和变量、专用工具过程、命令的默认参数、简单命令的解释执行模块 Cmd_Simp、词法分析器、HBase 过程处理模块以及命令行解释器主体模块 Cmd_C 等等。

第十一章介绍了一个新的语言范畴 C-DB。以 HBase 的模块为基础,可以用 C 语言作为 XBase 的数据操纵语言,11.4 节给出了一个示范程序。

第十二章介绍 HBase 的基本工具模块、通用的下拉式选单和鼠标库。第十三章讨论 HBase 的内外编辑器。这两部分内容不是 DBMS 核心技术,但却是不可缺少的部分。

第十四章讨论 DBMS 的优化和移植,包括常规优化、查询优化的策略和技术以及关于向网络和客户服务器方式移植的若干考虑。

第十五章介绍了 DBMS 安全性和完整性的一些实现方法和措施。

第十六章介绍了系统服务,包括 DOS 文件操作、备份、配置以及日历、计算器等等。

第十七章介绍了随书软件的安装方法。17.2 节给出了一组测试程序:包括学生成绩数据库 Score.DBF、学生成绩数据库应用程序 Score.PRG 以及学生成绩数据库应用程序过程文件 ScorePro.PRG。17.3 节给出了基于 HBase 的实习指南,供教学时参考。17.4 节按字母顺序列出了 HBase 的函数及所在模块。参考文献仅列出了最主要的参考书和参考论文目录。

本书中的若干专业术语的译法参照了较新的《英汉计算机技术辞典》(1992 年 6 月,电子工业出版社出版)。例如:“Menu”译为“选单”、“Default”译为“默认”,等等。

从探索到测试完成,HBase C 语言版本的开发历时一年半。但是,一个可以运行的程序,还不是能成书的程序。成书的程序讲究可读性,读之须如散文,而不要暗伏玄机,讲究程序美,行文要精练,而不要重复冗余;讲究规范性,风格要正,而不要奇。为此,改进优化又一年。其间,吴子华、杨文联、熊明、熊岚、杨文川、段晓冬、于中华、相利民、胡军、练林、汪姝、刘杰、

黄桂钦等先后参加了部分方案论证和程序设计工作,付出了辛勤的劳动,他们和署名作者一起,试图清除探索之路的荆棘,为读者开垦一片由程序组成的沃土。我们欢迎读者有标注地(即在参考文献或有关地方声明)引用我们的程序,允许读者复制程序或程序清单以用于非商业目的。所有这些,凝聚着一个希望:让读者通过自己的辛勤耕耘,学习和掌握:

1. DBMS 内部结构及其 C 语言实现技术。这是本书的主旋律。

2. 集成环境界面开发技术。书中给出了控制与数据分离的界面模块,其中包括通用的下拉式选单和鼠标库模块,读者可以通过模仿,改造的方式学习集成环境界面开发技术。

3. 中、大型软件的工程组织技术。HBase 1.5 版源程序共 30 余个模块,420 余个函数、约一兆字节。十几年前,这样一个系统可使一个软件公司立足于软件企业之林。它的总体设计、各模块之间的协调,沁润着若干经验和技巧,其中有许多来是自失败之后的反思。

4. Borland C 的特殊技术和良好的编程风格。

5. 使用新的语言范畴——C-DB,使 C 语言成为 XBase 的宿主语言。

我国在数据库管理系统的实现方面的起步较晚,一代新人将承担起急起直追的重任。面对新世纪的呼唤,面对发达国家不断推陈出新的 DBMS 成果,临渊羡渔,不如退而结网,开展一些踏实的工作。现在抛出版本号仅仅为 1.5 的 HBase,目的是让它作为一个开展讨论的基础。我们对批评与改进意见的渴求更甚于对它的褒扬。以此,作为“退而结网”中的一根平凡的网线。

作 者

1995 年 2 月于四川联合大学东区

目 录

第一章 数据库管理系统发展简史和实现技术要点	(1)
1.1 数据库技术发展简史	(1)
1.2 数据库系统=数据库管理系统+数据+应用程序	(3)
1.3 HBase 简介	(4)
1.4 实现 DBMS 的十个重要环节	(5)
第二章 DBMS 的控制流、数据流和模块结构	(8)
2.1 DBMS 的体系结构	(8)
2.2 HBase 模块间的支承结构	(9)
2.3 HBase 的模块及组织原则	(9)
2.4 HBase 的模块及调用关系	(12)
2.5 HBase 主要模块及依赖关系一览表	(12)
2.6 HBase 的主程序模块	(14)
2.7 HBase 的主程序清单	(15)
2.8 理解 HBase 的内部机制	(17)
第三章 用户界面	(18)
3.1 用户界面的一般原则：实用而质朴	(18)
3.2 HBase 的用户界面外观	(19)
3.3 数据与控制分离的设计思想	(19)
3.4 事件驱动方式和事件代码	(19)
3.5 HBase 的集成环境界面模块 Hdb_Menu	(20)
3.6 集成界面模块中主要控制流	(21)
3.7 集成界面模块程序清单	(22)
3.8 集成环境界面通用控制程序	(33)
3.9 下拉式选单控制模块头文件	(34)
3.10 鼠标驱动模块头文件	(34)
3.11 联机求助	(35)
第四章 数据的存储结构与传输机制	(37)
4.1 按结构存储,按结构理解	(37)

4.2	记录存储方式	(38)
4.3	XBase 库文件元数据存储结构	(39)
4.4	工作区——包容一个关系全部元数据的结构	(39)
4.5	全局类型及变量模块 Glob_Var. C 清单	(40)
4.6	文件输入输出管理	(47)
4.7	HBase 专用高速缓存	(48)
4.8	文件输入输出模块头文件	(49)
4.9	文件输入输出模块程序清单	(49)
4.10	打开与关闭工作区	(56)
4.11	工作区模块头文件程序清单	(57)
4.12	工作区模块程序清单	(57)
4.13	全局宏定义模块	(63)
第五章	索引与 B-树	(67)
5.1	从映射观点看索引	(67)
5.2	B-树结构中的三个层次	(69)
5.3	索引文件中的回收栈	(70)
5.4	B-树的高速缓存 Cache	(71)
5.5	B-树的三大特点	(72)
5.6	B-树的搜索、插入与删除	(72)
5.7	B-树模块头文件	(74)
5.8	B-树模块中的重要调用树	(75)
5.9	HBase 的 B-树模块要点	(77)
5.10	B-树模块 BTree. C 程序清单	(78)
第六章	数据定义语言 DDL	(109)
6.1	DDL 简介	(109)
6.2	Create 和 Modify 调用关系树	(110)
6.3	Create 和 Modify 算法	(111)
6.4	DDL 模块清单	(111)
6.5	用户数据的录入与修改—Edit	(126)
6.6	Edit_Rec 模块中调用关系树	(126)
6.7	Edit_Rec 模块中 Edit 的算法	(126)
6.8	Edit_Rec 模块程序清单	(127)
第七章	数据操纵语言 DML	(134)
7.1	关于查询语言的安全性、完备性简介	(134)
7.2	实用 DML 的附加功能	(134)
7.3	DML 的五要点	(134)

7.4	三个层次的删除	(135)
7.5	三种删除的实现	(135)
7.6	数据删除模块 DELETION 清单	(136)
7.7	选择操作算法的实现	(139)
7.8	选择模块 SELECT 程序清单	(141)
7.9	投影操作的原理与算法	(150)
7.10	投影模块 Project 程序清单	(151)
7.11	连接操作的基本原理和算法	(155)
7.12	连接模块 Join 程序清单	(157)
7.13	数据库浏览命令 BROWSE	(161)
7.14	BROWSE 模块源程序清单	(163)
7.15	排序操作的原理与算法	(183)
7.16	排序模块 Sort.c 程序清单	(184)
第八章	用户输入与输出	(187)
8.1	用户输入	(187)
8.2	用户输出	(188)
8.3	List.C 程序清单	(189)
第九章	数学表达式与统计计算	(196)
9.1	表达式处理	(196)
9.2	Math.C 程序清单	(197)
9.3	统计计算	(203)
9.4	Statist.C 程序清单	(204)
第十章	命令解释器	(210)
10.1	通用性和专用性的矛盾	(210)
10.2	数据库应用程序的执行方式	(210)
10.3	实现解释器的基本数据结构	(211)
10.4	解释器的专用类型和变量模块 Cmd_Var	(213)
10.5	HBase 解释器的调用树	(218)
10.6	解释器专用工具模块的头文件 Cmd_Tool.H	(219)
10.7	解释器工具模块 Cmd_Tool.C 程序清单	(220)
10.8	默认参数设置模块 CmdDefau.C	(227)
10.9	简单命令解释模块 Cmd_Simp.C	(230)
10.10	词法、句法、语义和语用	(250)
10.11	解释器的词法分析	(251)
10.12	词法分析模块 CmdParse.C 程序清单	(252)
10.13	命令行解释器 Cmd.C	(264)

10.14	HBase 的过程管理模块 Procedur. C	(280)
第十一章	C 作为 XBase 的数据操纵语言	(284)
11.1	语言的语义密度	(284)
11.2	混交语言——CDML	(284)
11.3	CDML 的应用	(285)
11.4	CDML 演示程序清单	(285)
第十二章	HBase 的基本工具模块	(290)
12.1	人类与工具	(290)
12.2	软件工程与 HBase	(290)
12.3	Tools 模块的要点与用途	(291)
12.4	一个被广泛使用的工具函数——字符串编辑 EditString	(291)
12.5	TOOLS 源程序清单	(292)
12.6	通用下拉式选单控制模块程序清单	(311)
12.7	鼠标驱动模块程序清单	(326)
第十三章	编辑器	(331)
13.1	DBMS 与编辑器	(331)
13.2	HBase 的内部编辑器	(331)
13.3	HBase 的外部编辑器	(332)
13.4	一个用 Turbo Prolog 实现的外部编辑器	(332)
第十四章	HBase 的优化与移植	(334)
14.1	常规优化	(334)
14.2	查询优化	(338)
14.3	HBase 的移植	(338)
第十五章	安全性和完整性	(340)
15.1	HBase 的安全性	(340)
15.2	存取授权表	(341)
15.3	用局部函数依赖维护 HBase 的完整性	(341)
第十六章	HBase 的系统服务	(343)
16.1	一般系统服务	(343)
16.2	备份服务	(353)
16.3	HBase 的配置服务	(353)

第十七章 软件的安装和测试	(357)
17.1 随书软件的内容及其安装	(357)
17.2 测试程序	(360)
17.3 基于 HBase 的实习指南	(366)
17.4 HBase 的函数及所在模块一览表(字典序)	(366)
参考文献	(380)

第一章

数据库管理系统发展 简史和实现技术要点

1.1 数据库技术发展简史

60年代以前,第一代计算机主要用于数值计算,为大量的科学计算服务。到了60年代,社会大系统中出现了包括人力(Man)、物力(Material)和货币(Money)的巨大物理流,又称为3M流。巨大的物理流伴随着巨大的信息流,用以运载信息的数字、文字和符号,汇成海量的数据流。对数据的搜集、存储、加工和传输工作,既宏大浩繁又琐碎细致。应运而生的新一代计算技术被广泛地应用于信息处理。

60年代中期,数据银行(Data Bank)一词就出现了。当时泛指一个较大的文件管理系统,与现在的数据库管理系统含义相差甚大。60年代末,美国系统发展公司为海军基地研制数据库,借用了Base(基地)一词,Data Base形象地指代在信息的海洋中巡航所必须的基地。

从哲学上看,在客观世界三要素,即物质、能量和信息之中,信息区别于前二者的最大特点是不守恒性。能量和物质服从守恒定律,能量物质库中物质取走一点少一点。而数据库中的信息却可无限复制。如果说,早期的数据银行一词还多少意味着一点数据的独占权,而数据库一词则更多地意味着数据的持久性和共享性,带来了信息社会的曙光。由此可见,把Data Base译为“数据库”要比译为“资讯基”合理。

1969年,IBM公司宣布了网状数据管理系统产品IMS,因此许多专家认为数据库诞生于60年代末。

数据库技术从诞生起,就横跨了理论、应用和系统研究三大领域。

各大计算机厂商和软件公司是早期开发和应用数据库技术的主力。到1978年,美国信息处理学会(ASIS)编辑出版的《数据库目录》中搜集的数据库就有300多个。1982年出版的《联机数据库目录,工业通讯录》中的数据库猛增加到1350个。

系统开发和应用研究的需要促进了理论研究的发展。1979年,E. F. Codd在著名的“大型共享数据库的数据关系模型”一文中提出了关系数据库模型,并开展了理论研究工作。经过一大批计算机科学家十多年的努力,关系数据库以简单灵活、数据独立性高、理论严格等

优点表现了强大的生命力。为此,E. E. Codd 获得了 1981 年的 ACM 图灵奖(计算机科学最高奖)。

数据库设计方法也不断完善。1971 年,美国数据库系统语言会议(Conference on DataBase System Language, CODASYL)的 DBTG 报告提出了三级抽象模式,开始解决数据独立性问题。1979 年 Peter Chen 提出了数据库逻辑设计的实体—联系方法。1978 年新澳尔良的 DBDWB 报告把数据库系统的设计分为四步,即需求分析、信息分析和定义、逻辑设计和物理设计四个阶段。

社会的需要使数据库成了大学里的热门课程,1980 年 J. D. Ullman 写出了“Principles of DataBase Systems”,1981 年 C. J. Data 写出了“An Introduction to DataBase System”,1984 年 David Maier 写出了“Theory of Relational DataBase”,标志着数据库在理论上的成熟。

70、80 年代是数据库厂商创家立业的时代。斯坦福大学的 G. Wiederhold 在他 1983 年出版的《数据库设计》(DataBase Design)一书的附录中列出了 1969~1981 年期间研制的 120 个 DBMS,介绍了这些 DBMS 的研制公司,适用机型和特点,描绘了数据库领域在那个年代群雄并起,百花争艳的繁荣局面;几乎每一所著名大学都有数据库研究队伍,著名公司都推出了 DBMS 产品。有的公司推出了多种 DBMS,而一个优秀的 DBMS 就可使一个公司立足于竞争之林。

经过二十多年的竞争,优胜劣汰。一大批 DBMS 以自己的特色得到用户的认可。表 1.1 列出部分著名的 DBMS。

表 1.1 部分著名的 DBMS

公司	DBMS		公司	DBMS	
Oracle	Oracle	(关系)	Fox/Microsoft	FoxBase	(关系)
Sybase	Sybase	(关系)	Ashton/Borland	dBase	(关系)
Informix	Informix	(关系)	Microsoft	Access	(关系)
Ingres/CA	Ingres	(关系)	Borland	Paradox	(关系)
IBM	IBM-HDBMS	(层次)	IBM	System 2000	(网状)
IBM	IDMS	(网状)	IBM	SQL/DS	(关系)
IBM	DB2	(关系)	Prime	PrimeDBMS	(关系)
DEC	VAX-DBMS	(网状)	DEC V	RDB	(关系)
HP	Image 3000	(关系)	HP	AllBase/SQL	(关系)
Unify	Unify 2000	(关系)	MicroData	KnowledgeMan	(关系)

表 1.2 部分较新的 DBMS

公司	DBMS	备 注
Polyhot	PolyBase	多媒体数据库管理系统
EmPress	Empress	科学与工程数据库系统
DataBasix	Ease	图形数据库管理系统
Leading Technology	TechBase	工程管理数据库系统

90年代是分布数据库管理系统产品走向成熟、富于远见的理论家们考虑跨世纪 DBMS 的时代。几十种新的数据库模型被提出来。语义模型先荣而后衰,演绎数据库、时态数据库、图形数据库、工程数据库、主动数据库、主存数据库、多媒体数据库等等,百花齐放,面向对象数据库后来居上,显示了巨大的活力和光明的前途。新的 DBMS,包括试用品原型和试制品原型不断被推出。表 1.2 列出部分较新的 DBMS 产品。

我国在数据库方面起步较晚,十多年来在应用、理论和系统方面都取得了长足的进步。特别是在数据库应用系统和理论研究的某些方面,取得了令人瞩目的成果。近年来,在系统实现方面也取得了长足的进步。表 1.3 列出了部分有我国自主知识产权的 DBMS。

表 1.3 部分有我国自主知识产权的 DBMS

DBMS	开发单位	备 注
COBase	北大,人大,中软公司,华中理工大学	关系型,多用户,支持客户-服务器方式。
iTBase	清华大学	X86 微机,关系型
CDB	中科院数学所	X86 微机,关系型
GWBase	中科院成都计算所	X86 微机,关系型
HBase	四川大学	与 XBase 兼容,含时态处理机制,目前唯一开放源程序的 DBMS

1.2 数据库系统=数据库管理系统+数据+应用程序

在微型计算机上用 XBase(dBase, FoxBase, HBase 等等)建立一个学生成绩管理系统,其组成如图 1.1(a),一般而言,在通用的数据库管理系统(DBMS——Data Base Management System)上建立的数据库系统如图 1.1(b)。图中描述了数据库系统各部分支持与被支持的关系。

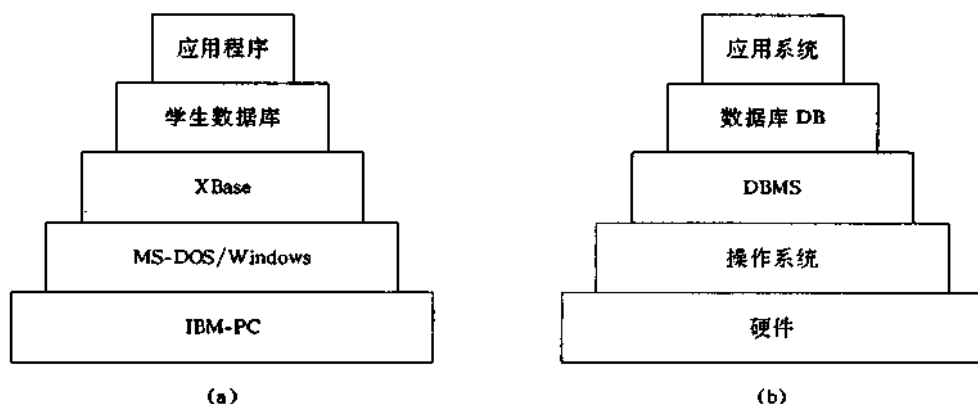


图 1.1 带数据库的计算机系统

数据库的主体是代表信息的字节流集合（用户数据）以及用以识别字节流的模式（属于元数据，称为数据库模式）。用 DBMS 的数据定义语言 DDL(Data Definition Language) 建立了数据模式，再按照模式，填入数据，即组成数据库。

应用程序是用 DBMS 的数据操纵语言 DML (Data Manipulation Language) 和宿主语言 (Cobol, Pascal, Fortran, C 等等) 编写的以应用为目的的，对数据进行查询、修改和存取的工具性程序。

如果把数据比喻为银行中的资金，则硬件是银行建筑物，DBMS 是银行的全套管理机制，而应用程序好比是自动存取款机。

本书重点讨论 DBMS 的实现技术，以主要篇幅介绍了 HBase 的实现技术。在附录中和磁盘上给出了一个学生成绩数据库 SCORE.DBF，应用程序 SCORE.PRГ，以及应用程序过程文件 SCOREPRO.PRГ（能在 dBase, FoxBase 和 HBase 上正确运行）。分别对应于图 1.1 中的 DBMS, DB 和应用程序三个层次。

1.3 HBase 简介

本书以作者开发的在 DOS/Windows 平台上运行的关系数据库管理系统——HBase 为例，通过剖析它的完整的 C 语言源程序，来详细地讨论用 C 语言实现 DBMS 的技术和相关理论。

HBase 是国家自然科学基金资助项目。H 即 History(历史)。HBase 的最终目标是开发一个能处理历史性数据和知识的时态数据管理系统，即 Temporal DBMS。英语中，Temporal 有“临时”和“时态”两个意义，本书中的 Temporal 及其缩写 Tempo 均指时态。为了便于交流，HBase 兼容了 XBase 的数据文件格式、DDL 和 DML。

HBase 具有一个实用的界面：兼有下拉式选单和命令行驱动的综合环境。支持鼠标、Alt 热键、上下文相关的联机求助、附有计算器、Ascii 表、DosShell 以及文件管理例程。支持交互方式和命令文件 (PRГ) 解释执行方式。支持 B-树，有内外全屏幕编辑器。

在实用而质朴的界面之下，追求较深层次的探索：关系 DBMS 专用的高速缓存数据结构、带有时态数据处理能力的 DDL 和 DML 的实现、B-树 Cache 中具有低频先见废弃旧节点处理、用框架结构实现解释器的尝试，DBMS 开发工程的组织技术，等等。

HBase 的 1.0 版和 1.5 版分别用 Turbo Pascal 6.0 和 Borland C 3.1 开发,效果和外观基本一致。从读者面、可重用性、科研参考价值,以及篇幅限制诸方面考虑,同时也是为了突出 DBMS 的实现技术这一重点,本书中列出并剖析了用 Borland C++ 3.1 开发的 1.5 版本。对通用 DBMS 部分做了完全的描述,对时态数据处理的专用机制,介绍了基本部分,略去一些过分专深的内容,而只在适当的地方简要地介绍了时态查询语言的实现。

HBase 的源程序在磁盘上共约 1 兆。在下列条件下,即采用预编译头文件,且精心设计工程文件中模块顺序,使头文件不重复编译,根据 BC 3.1 编译器的报告,有 24000 余行。

用 Borland C 2.0 编译时,至少要两兆内存,当用 Borland C 3.1 版、4.0 版或 4.5 版本编译,机器内存为 4 兆,CPU 为 386/DX40 时,在工程文件 HBase.Prj 中所有条目都是 *.C 文件的条件下,用 Build 选项,编译全部模块一次,大约要 4 分钟。

在默认方式下 HBase 按普通数据库管理系统的方式工作。当在 Config.HDB 文件中设置 TempoDb=On,或用命令 Set Tempo=On 后,它才按时态数据库方式工作。

在本书出版的同时,作为电子出版物的磁盘上将列出 1.51 版或更高版本全部源程序,包括可能的勘误和改进。HBase for Windows 是 HBase 的升级版本,考虑到 Windows 编程目前尚未为广大读者掌握,为了不冲击 DBMS 实现技术这一重点,不在本书讨论,而将在以后发行的磁盘上列出。读者可根据书末附录,查阅有关信息,与作者联系,以获得最新的版本。

1.4 实现 DBMS 的十个重要环节

研制和实现大型 DBMS 是一项旷日持久,浩繁而细致的系统工程。网状数据库管理系统 IMS 花了几千个人年,System R 用了 120 个专家年和几千个程序员年。难怪亲经者说,一个风华正茂的年轻人,需献出大部分青春才能把一个大型 DBMS 的内部机制的实现技术吃透。微机 DBMS 相对规模较小,具备了 DBMS 最主要的数据结构和控制机理,初学者在不太长的时间内,可以通过微机 DBMS 了解数据库管理系统实现的主要思想和技术。

一个通用 DBMS 实现过程中至少有下列要点(同时也是难点):

1. 选择数据模型。十年前,DBMS 研制者在确立项目之前要花很大的力量来论证网状、层次和关系模型的优缺点,可行性。现在人们要在数据库的特殊技术上,例如面向对象、主动数据库机制、时态处理机制、多媒体处理机制上权衡论证。突出新的 DBMS 的特色。

2. 与 OS 的接口。主要体现为文件管理系统。数据的持久性和可操纵性曾经被用来作为数据库定义的基本点。持久性通过存储介质(磁盘、磁带和光盘等)体现,可操纵性则要求在内存开辟工作区保存当前数据。工作区与介质之间的数据传递是由文件系统完成的。文件管理系统有三种实现方案,即:

(1) 直接方案:DBMS $\leftrightarrow$ 文件管理系统 $\leftrightarrow$ 存储设备。

(2) 间接方案:DBMS $\leftrightarrow$ 文件管理系统 $\leftrightarrow$ OS $\leftrightarrow$ 存储设备。

(3) 折衷方案:介于前两者之间。一些开发平台为用户提供了可调用的库函数,例如,比较新的开放式客户—服务器系统上提供了在客户方和服务方存取数据的结构和函数,直接调用这些函数,可使开发工作简化且性能随着平台的升级而升级。

方案 1 完全抛开 OS,而开发 DBMS 独立的 I/O 机制,在 DBMS 内部建立数据和磁盘地址(柱面、道、扇区)的对照表,优点是效率高,但工程量大。不易利用 OS 升级的技术利益。方

案 2 中 DBMS 只完成了 DBMS 到 OS 的映射,对硬件的独立性高且易实现,著名的 SYSTEM R 采用了这一方案。方案 3 对常规数据的处理同方案 2,而对要求高速的声像数据和实时控制数据的处理同方案 1。著名的层次数据管理系统 IMS 采用了这一方案。

3. 工作区结构。在数据库系统中,每在内存中打开一个库文件,就需保存并及时跟踪一系列控制性的数据。如文件名、索引的表达式、记录总数、当前记录号、当前记录值、B-树的当前结点和根结点,等等。在 HBase 中,这些数据都包含在工作区结构中。新型 DBMS 的各种特色(例如面向对象机制、主动数据库机制、多媒体处理机制以及本书中 HBase 的时态处理机制等等)主要实现结构在工作区中。

4. 数据定义语言 DDL。通用 DBMS 的通用性是以 DDL 为基础的。用户使用 DDL,能够随意创建和修改数据模式。在 dBase, HBase, Ingres, SQL 语言中一致采用了命令“Create”和“Modify”来实现这一功能。

5. 数据操纵语言 DML。在关系数据库模型中,对数据的操纵分为多个层次。例如,在关系一级的操作有选择、投影和连接。在元组(记录)一级的操作有插入、修改和查询,其中又有按关键字查询、查下一记录,上一记录,第一记录,最后记录等等。此外还有计算平均值、最大最小值和计数等统计功能。DBMS 的各种特色的操作(主动数据库机制、多媒体处理机制、以及本书中 HBase 的时态处理机制,等等)在 DDL 和 DML 中实现。

6. 索引管理系统。“检索”一词形象地刻画了 DBMS 回答用户查询时的“检查”和“索取”两个动作,要点如下:

输入: 用户要求(例如关键字)

输出: 用户所需信息。

步骤: 寻找、检查(匹配关键字),定位。一旦定位,则调用文件管理系统功能;索取数据,进而输出结果。

搜索的效率依赖于被搜索空间的大小和被搜索对象的组织(是否排序)。索引的基本思想是以较小的排序集合{ (关键字,记录地址) } 代替较大的主数据集合,作为被搜索空间。实现了“以小有序管大无序”,从而提高了检索效率。

目前,所有的商品化的 DBMS 都采用了 B-树(或其变种)作结构化索引。

7. 命令解释器,编译器或伪编译器。DBMS 提供的 DDL 和 DML 应能按两种方式运行,即

(1). 交互式。从数据约束(Binding)的观点看,通用数据库管理的数据对象都有迟约束特性(Late Binding)。DBMS 的设计者不可能预知用户将要管理的数据细节。而交互式操作能让学生根据上一操作的结果和反馈决定下一操作的取向。用人的智能干预和导航 DBMS。

(2). 批处理方式。当用户把数据库模式约束为具体对象时,批处理具有可重复性,并获得高效率。批处理是由解释器,编译器或伪编译器支持的。其难点在于:

- ① 循环语句,在一定条件下循环以及跳出循环的操作,如 dBase/HBase 的 Do While, Exit 和 Loop。
- ② 条件语句,If... Else... EndIf 等。
- ③ Case 语句,它实质上是对数据迟约束的一种处理策略。根据运行时,数据被约束状况进行分流。
- ④ 嵌套。包括命令文件的嵌套,过程的嵌套,上述语句的嵌套。