《Architecture of a Database System》

（中文版）

Joseph M. Hellerstein, Michael Stonebraker and James Hamilton

翻译：林子雨

厦门大学数据库实验室

http://dblab.xmu.edu.cn

中文版网址：http://dblab.xmu.edu.cn/node/459

厦门大学计算机科学系教师 林子雨 翻译作品

http://www.cs.xmu.edu.cn/linziyu

2013 年 9 月

前言

本文翻译自经典英文论文《Architecture of a Database System》，原文作者是 Joseph M. Hellerstein, Michael Stonebraker 和 James Hamilton。该论文可以作为中国各大高校数据库实验室研究生的入门读物，帮助学生快速了解数据库的内部运行机制。

本文一共包括 6 章，分别是：第 1 章概述，第 2 章进程模型，第 3 章并行体系结构：进程和内存协调，第 4 章关系查询处理器，第 5 章存储管理，第 6 章事务：并发控制和恢复，第 7 章共享组件，第 8 章结束语。

本文翻译由厦门大学数据库实验室林子雨老师团队合力完成，其中，林子雨老师负责统稿校对，刘颖杰同学负责翻译第 1 章和第 2 章，罗道文同学负责翻译第 3 章和第 4 章，谢荣东同学负责翻译第 5 章、第 6 章、第 7 章和第 8 章。

如果对本文翻译内容有任何疑问，欢迎联系林子雨老师。

林子雨的E-mail是：ziyulin@xmu.edu.cn。

林子雨的个人主页是：http://www.cs.xmu.edu.cn/linziyu。

厦门大学数据库实验室网站是：http://dblab.xmu.edu.cn。

本文中文版的网址是：http://dblab.xmu.edu.cn/node/459。

林子雨于厦门大学海韵园

2013 年 9 月

摘 要

数据库管理系统（DBMS）广泛存在于现代计算机系统中，并且是其重要的组成部分。它是学术界以及工业界数十年研究和发展的成果。在计算机发展史上，数据库属于最早开发的多用户服务系统之一，因此，它的研究也催生了许多为保证系统可拓展性以及稳定性的系统开发技术，这些技术如今被应用于许多其他的领域。虽然许多数据库的相关算法和概念广泛见于教科书中，但关于如何让一个数据库工作的系统设计问题却鲜有资料介绍。本文从体系架构角度探讨数据库设计的一些准则，包括处理模型、并行架构、存储系统设计、事务处理系统、查询处理及优化结构以及具有代表性的共享组件和应用。当业界有多种设计方式可供选择时，我们以当前成功的商业开源软件作为参考标准。

第 1 章 概述

厦门大学计算机科学系教师 林子雨 编著

个人主页：http://www.cs.xmu.edu.cn/linziyu

中文版网址：http://dblab.xmu.edu.cn/node/459

2013 年 9 月

第 1 章 概述

数据库管理系统（DBMS）是一种复杂的、关键任务软件系统。今天的数据库管理系统包含了学术界和工业界数十年的研究以及大量的企业软件开发成果。数据库管理系统属于最早期广泛应用的在线服务系统之一，因此，具备前沿的设计方法，这些设计方法涵盖数据管理、计算机应用、操作系统以及网络服务等方面。早期的数据库管理系统是计算机科学领域最具影响力的软件系统之一，而且，那些因为数据库研究而产生的理念和实现技术也被广泛地借鉴和创新。

由于诸多原因，数据库管理系统架构的相关介绍并没有像它应该的那样被人们广泛地熟知。首先，应用数据库群体较小。由于市场只能支撑几个高水平的竞争者，因此，只有一小部分成功的数据库产品存在。从事数据库设计和应用的人们彼此联系紧密，他们往往来自于同一所学校，研究同样的知名项目，然后合作开发几个相同的产品。另一方面，数据库管理系统的教学领域往往忽视对体系架构问题的讲解。数据库教材一直关注那些易于教学、研究和考试的算法和理论知识点，没有从应用角度对数据库架构有一个全局的讲解。总而言之，关于如何构建一个数据库方面的知识，并不是保密的，科室，它并没有被系统地写下来并供人们讨论交流。

本文中，我们希望通过几个方向的讨论，介绍清楚现代数据库系统架构的主要方面。这些内容部分见于教材中，我们会给出合适的注释。另外有些内容埋藏于用户手册中以及一些数据库相关团体的口头交流中。在适当的情况下，我们使用商业开源软件作为复杂多样的数据库架构的实例。当然，受篇幅所限，在这至少有十年历史的数百万行代码中，它们的特性以及一些好的创新点就不能一一列举了。我们的重点在于整个系统的架构设计，并着重讲解那些没有被教课书着重谈到的、但是却使那些广为人知的算法发挥作用的系统环境。我们希望读者已经熟悉主流的数据库教材，并且对现代操作系统如 UNIX、Linux 以及 Windows 有基本的操作能力。在下一节整体介绍完一个数据库管理系统的架构之后，我们在 1.2 节为每一部分组件提供一些参考资料作为背景阅读。

1.1 关系数据库：查询的命脉

迄今为止，发展成熟并且得到最广泛应用的数据库类型是关系数据库（RDBMS）。它们广泛应用于许多应用系统中，比如电子商务、医疗文件、人力资源、工资系统、客户联系管理以及供应链管理等等。网络社区的出现也使得它得到更广泛的应用。关系数据库几乎被用作所有的网络交易及在线内容管理系统的数据存储方式，如博客、维基百科、社交网络等等。关系型数据库不仅是重要的软件设施，它也是帮助我们理解未来可能发生的数据库变革的很好的切入点。因此，本文将始终以关系型数据库为例。

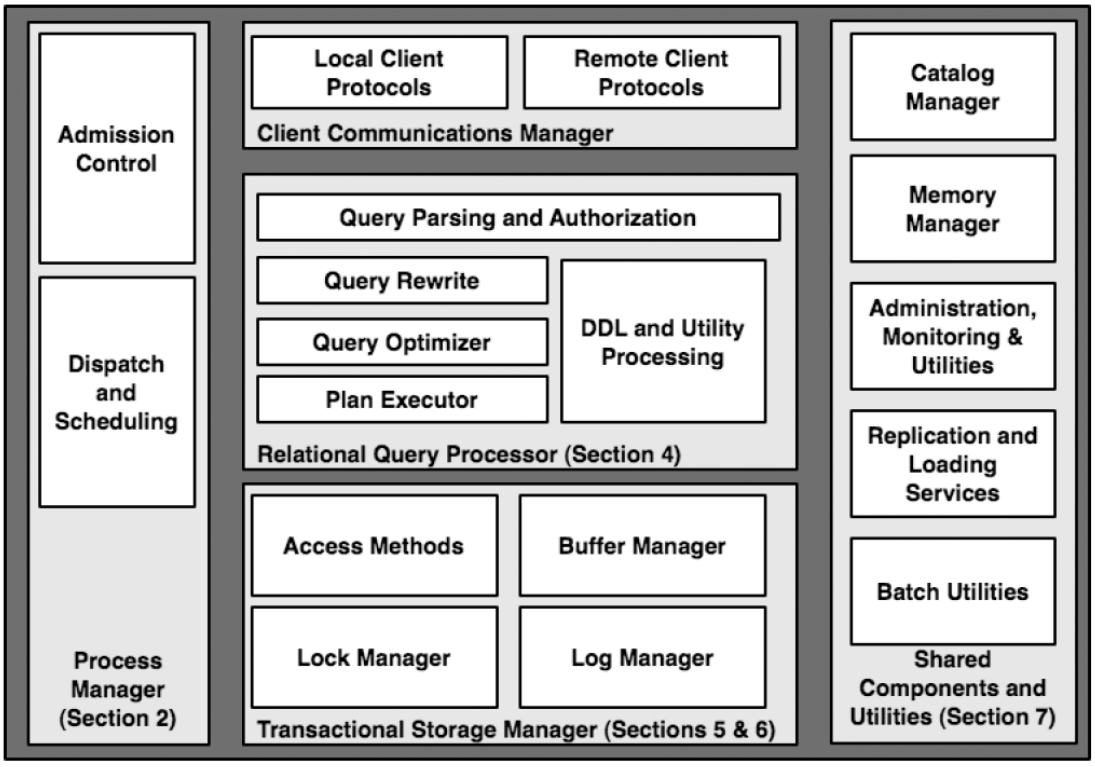

图 1-1 一个 DBMS 的主要组件

如图 1-1 所示，关系数据库主要由五部分组成。为了分别介绍这五个部分的内容以及它们之间如何相互合作，我们先来看看一个查询语句在数据库系统是如何被处理的。这同时也作为本文内容的概览。

让我们来考虑这样一个简单却很典型的数据库在机场的应用实例：查询某次航班的所有旅客名单。这个操作所引发的的查询请求大致按如下方式被处理：

机场登机口的 PC 机（客户端）调用 API 与 DBMS 的客户端通信管理器（Client Communications Manager）建立网络连接。在一些情况下，客户端直接通过 ODBC 或 JDBC 连接协议与数据库服务器建立这种连接。这种处理方式被称为“two-tier”或者

“client-server”。还有一些情形中，客户端与“middle-tier server”建立连接，如 web 服务器、事务处理系统等，它们通过协议代理原本直接建立在客户端与数据库服务器之间的连接，这被称为“three-tier”模式。在一些网络应用中，还会再多一个应用服务器架设于网络服务器和 DBMS 之间，这被称为“four-tier”。为适用于如此多的环境，一个 DBMS 需要兼容许多不同客户端以及中间系统的多种连接协议。不过实际上，DBMS 中负责多种协议的管理器基本上是相同的：为调用者（客户端或者中间件）建立连接并记录其连接地址，对客户端的 sql 语句做出回应，并在适当的时候返回数据以及控制信息。在本文例子中，通信管理器还将为客户端建立安全证书，为新的连接细节以及客户端 sql 命令分配空间，并将客户端的请求传送到 DBMS 更底层进行处理。

在收到客户端的第一个请求之后，DBMS 必须为之分配一个计算线程。系统必须确保该线程的数据以及控制输出是通过通信管理器与客户端连接的。这些工作交由 DBMS 的进程管理器（Process Manager）来管理（图 1-1 左）。在这一部分中，DBMS 所做的主要工作是准入控制，即系统是否应该立即处理该查询，或是等待系统有足够资源时再处理该查询。我们将在第二章详细介绍进程管理器。

在分配控制进程之后，登机口的查询便可以被处理了。处理工作借助于关系查询处理器（Relational Query Processor，图 1-1 中间部分）中的代码来实现。这些模块检查用户是否有权进行该查询，然后将用户的 sql 查询语句编译为中间查询计划。在编译之后，结果查询计划被交给查询执行器。查询执行器包含一系列处理查询的操作（关系型算法实现）。典型的处理查询任务的操作包括：连接、选择、投影、聚集、排序等等，当然也包括从底层读取需要的数据。在我们的例子中，包括优化查询的操作在内，调用了一个操作集合来解决用户的查询问题。在第四章我们将讨论查询处理问题。

在登机口代理的查询计划的底层，由若干操作从数据库请求数据。这些操作通过调用(call)来从 DBMS 的存储管理器（transactional storage manager，图 1-1 底部）中收集数据；存储管理器负责所有的数据接口（读）和操作调用（建立、更新、删除）。存储系统包括为管理磁盘数据的基本算法和数据结构，比如基本的表和索引。它还包括一个缓冲管理器，用来控制内存缓冲区和磁盘之间的数据传输。回到我们的例子中，在获取数据的过程中，登机门客户端的查询必须调用传输控制代码来保证“ACID”性质（将在第 5.1 节讨论）。在获取数据之前，需要通过锁管理器来确保并发情况下运行的正确性。如果登机口客户端的查询包含对数据库的更新操作，那么，它需要与日志系统进行交互，来确保更新操作的持久性以及撤销操作的完整性。在第 5 章，我们会讨论存储与缓冲管理的更多细节，弟 6 章介绍业务一

致性结构。

在查询的这一时期，查询操作已经开始获取数据并准备好用他们来为客户端结算结果。这一步通过展开我们之前提到的所有操作的堆栈来完成。访问方法把控制控制权交给查询处理器，查询处理器将数据库的数据组织成结果元组；结果元组生成后被放入客户通信管理器的缓冲区中，然后该通信管理器将结果发送给调用者。对于较大的结果结合，客户端会发送更多的请求来获取更多的数据，这也导致了通信管理器、查询处理器和存储管理器的循环操作。在我们的例子中，查询操作的最后传输结束，连接关闭；传输管理器中的结果被清空，进程管理器释放无用的数据结构，通信管理器将连接状态清空。

我们通过这个查询的例子讨论了 RDBMS 的许多关键组件，但还有一些没有涉及到。图 1-1 右边一侧有许多共享组件和工具，它们对于一个功能完整的 DBMS 而言，同样是十分重要的。目录和存储管理器在传输数据时被作为工具来调用，在我们的例子中也是这样。在认证、分解以及查询优化过程中，查询处理器都会用到目录。同样，存储管理器也广泛应用于整个 DBMS 运行过程中动态分配和释放内存的场合。在图 1-1 中最右边列出的其余组件，独立运行于任何查询，它们使数据库保持稳定性和整体性。我们将在第 7 章讨论这些共享的组件和工具。

1.2 本文内容介绍

本文的大部分篇幅着重介绍支撑数据库核心功能实现的架构原理。我们没有过多涉及数据库算法，在文档中你可以很方便地找到它们。对于现代 DBMS 系统的衍生品，我们也很少讨论，它们虽然提供了一些不同于数据库核心部分的特征，但是，这对于数据库架构的改变并不大。然而，在本文的很多章节中，我们感兴趣的内容有些超出了文章应该涉及的范围，我们将尽可能得为这些额外的知识提供参考资料。

我们开篇谈到了数据库系统的整体架构。在所有服务器架构中第一个问题便是服务器整个的进程架构，我们探索出各种切实可行的替代品，首先是单处理机，然后是多种并行处理架构。关于核心服务器系统架构的讨论，是适合于许多其他系统的，但是，在很大程度上是在 DBMS 设计中最早得到应用的。随后，我们更注重于 DBMS 的专有组件。我们从一个简单查询的角度开始，重点关注关系数据库查询处理器。在这之后，我们研究存储架构和数据传输存储控制。最终，我们介绍一些大部分 DBMS 都会有的共享组件和工具，这一点在教材中很少被提及。

附录 1:译者介绍

林子雨(1978－),男,博士,厦门大学计算机科学系助理教授,主要研究领域为数据库,数据仓库,数据挖掘.
主讲课程：《大数据技术基础》
办公地点：厦门大学海韵园科研 2 号楼
E-mail: ziyulin@xmu.edu.cn
个人网页：http://www.cs.xmu.edu.cn/linziyu