

数据库系统概论新技术篇

王 珊

中国人民大学信息学院

2017年5月

数据库新技术概述

王 珊

中国人民大学信息学院

2017年5月

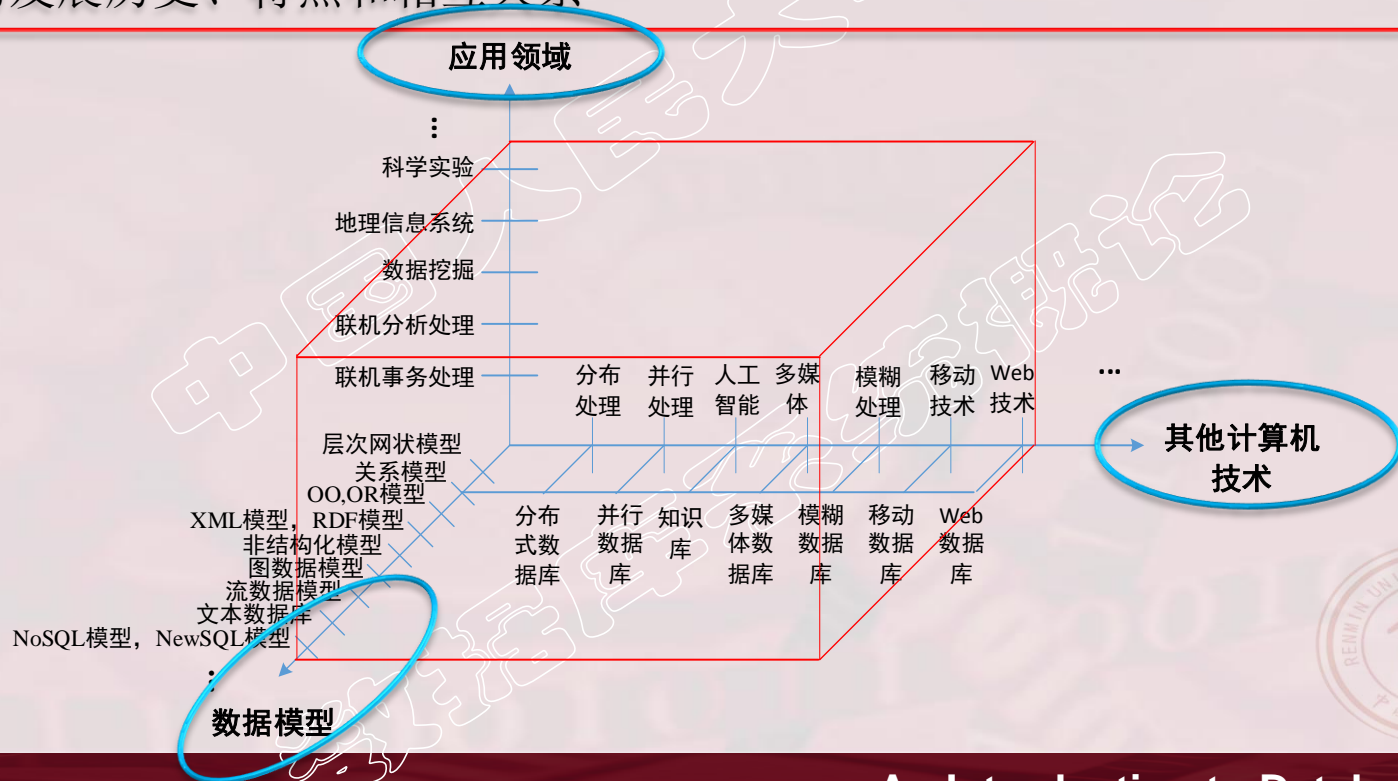
目录

- 数据库技术50年发展回顾与启示
- 当前数据管理与应用面临的挑战
- 大数据时代数据库新技术概述



数据库技术50年发展回顾与启示

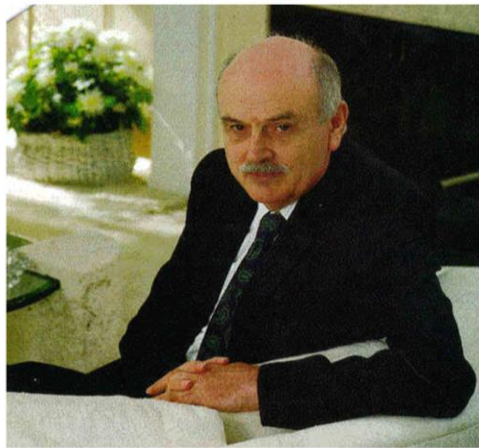
通过一个三维空间的视图从数据模型、相关技术、应用领域3个方面，描述数据库系统的发展历史、特点和相互关系



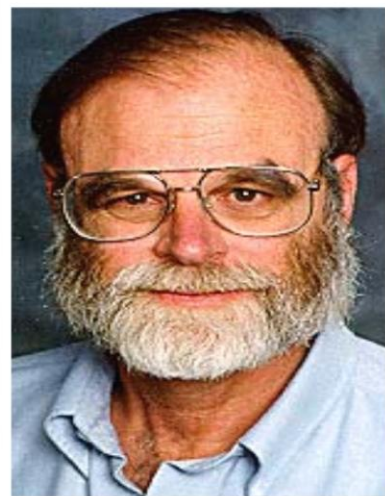
四位图灵奖得主



C.W.Bachman (1973)



E.F.Codd (1981)



James Gray (1998)



**M.R.Stonebraker
(2014)**



第一代数据库系统

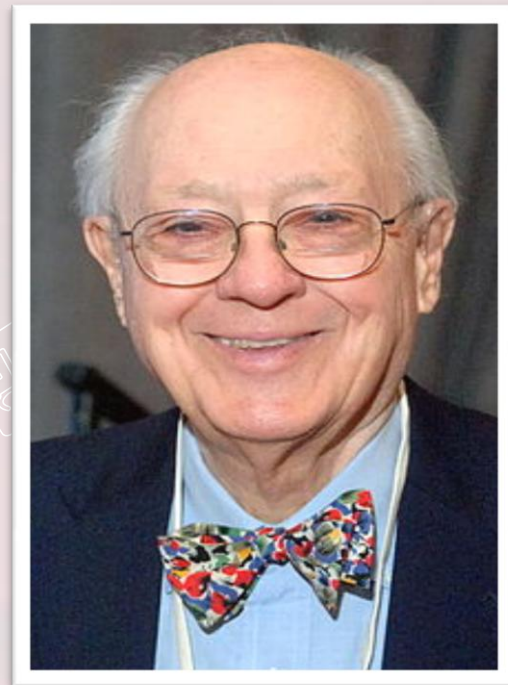
- 第一代数据库系统指层次和网状数据库系统
- 代表系统：
 - IMS（Information Management System）
 - 1969年，IBM公司研制，**层次模型**的数据库管理系统
 - DBTG报告
 - 20世纪60年代末70年代初，数据库任务组DBTG（Data Base Task Group）提出
 - DBTG报告确定并建立了数据库系统的许多概念、方法和技术。
 - DBTG所提议的方法是基于**网状结构**的。但是，提出的概念、方法和技术是普遍适用的。



20世纪60年代的代表成就

- IDS: Integrated data store, 为通用电气公司开发了第一个数据可独立存在、多个应用可同时共享访问的数据库产品。
- 推动与促成了DBTG报告, 提出数据库系统三级模式结构, 确定了数据库系统的基本结构。

由于他在在数据库技术的产生、发展与推广应用方面发挥的巨大作用, 1973获图灵奖。



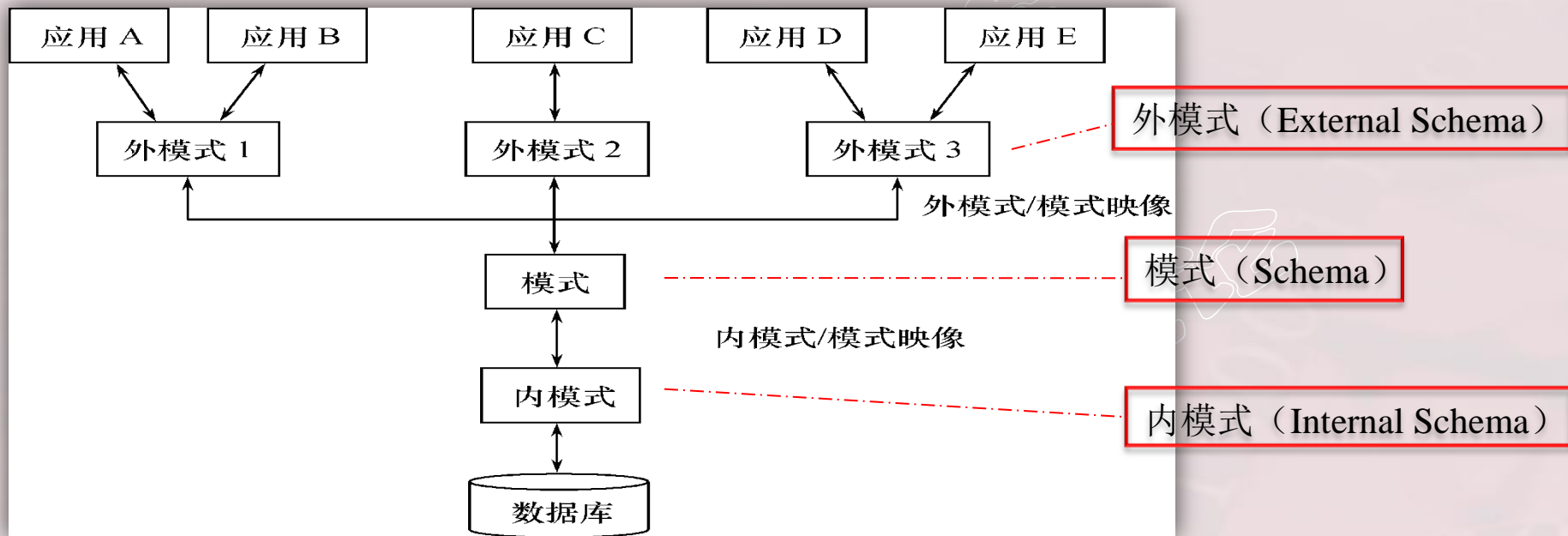
Charles W. Bachman

启示之一

- “数据独立性”思想是个创新！
- 让数据与应用有一定距离（数据与应用分离，数据独立性）
可以增加DBMS系统的适应性和应用系统的稳定性。
- “模式”在三级模式结构中扮演着中心的角色，是应用系统与数据库系统“交流”的桥梁。
- 体现了数据的语义关系。



数据库系统的三级模式结构



数据库系统的三级模式结构

50多年来，数据库应用系统不管多么复杂，基本上都保持了这个结构

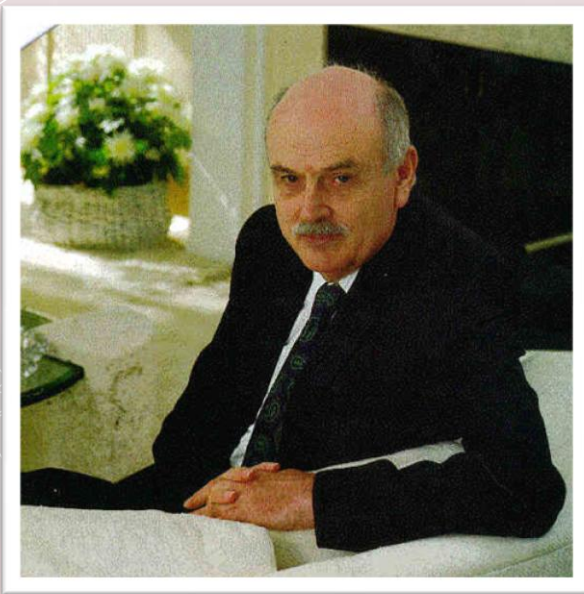
第二代数据库系统

- 1970年，E.F.Codd发表《大型共享数据库数据的关系模型》论文，为关系数据库技术奠定了理论基础
- 典型代表：
 - IBM San Jose研究室开发的System R
 - Berkeley大学研制的INGRES



20世纪70年代的代表性成就

- **关系数据模型的提出**
- 关系数据理论和关系数据库语言
- 关系数据库的研制成功和广泛应用
- **关系数据库的优点**
 - 简单的数学模型
 - SQL非过程化的标准语言
 - 查询优化技术
 - 事务管理技术
- 关系数据库之父，1981年获图灵奖



Edgar F. Codd



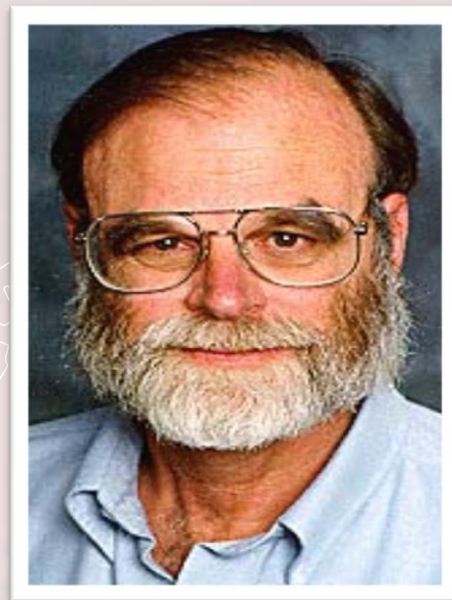
启示之二

- 关系模型是个创新！
- 非过程化语言为提高应用系统开发的生产率提供了很大的空间。
- 系统性能问题，可以通过技术的不断进步来改善。
查询优化技术成为系统研究与开发永恒的主题。



20世纪80年代的代表性成果

- 事务处理理论(ACID)与实现技术使得关系数据库走向极致，是最成功的信息技术之一
- **保证数据库的完整性和一致性**是DBMS的最基本要求
- 在数据库和事务处理研究方面的元创性贡献以及在将**研究原型转化为商业产品**的系统实现方面的技术领袖地位，1998年获图灵奖



Jim Gray

启示之三

- 关系数据库不仅仅是一个产品，而且形成了一个生态：
产品、标准、工具、应用开发、部署、运行平台。

- 产品(Oracle, IBM DB2, MS SQLServer, ...)
- 金仓数据库, 达梦数据库, ...
- 开源数据库, ...)
- 标准(ISO SQL, ANSI SQL, ...)
- 评测(TPC-C, TPC-H, ...)
- 工具(PowerDesigner, ...)
-



21世纪初的代表性成果

新一代数据库系统架构的研究与开发

- 1971年至2000年为第一阶段，关系数据库的体系架构与实现技术研究
- 2001年至2008年为第二阶段，提出“one-size does not fit all”观点，针对不同应用设计开发了一系列新型数据库系统的体系架构
 - 流数据库Auraro
 - 列存储数据仓库C-Store
 - 高性能OLTP系统H-Store
 - 科学数据库SciDB，等等
- 2009年至今为第三阶段，大数据系统的体系架构设计与实践



Michael Stonebraker

现代主流数据库系统架构的奠基人，2014年获图灵奖

启示之四

- **One Size Does Not Fit All**
- 按照应用需求与系统功能分，可以有各种不同类型的系统
 - OLTP (Oracle)
 - OLAP (TeraData)
 - Search (Hadoop)
 - 大数据丰富多彩的应用
 - ...



启示之四

- 传统的**数据库核心技术**是大数据管理和处理技术的基础
- 大数据不等于，也不能取代传统的数据库技术
- 大数据与其他数据处理技术相结合，通过细分的技术和市场，满足用户不同层面的需求。



目录

- 数据库技术50年发展回顾与启示
- 当前数据管理与应用面临的挑战
- 大数据时代数据库新技术概述



数据管理与应用面临的挑战

- 推动数据库技术发展的3个主要动力
 - 数据
 - 应用需求
 - 计算机相关技术



数据管理与应用面临的挑战

一、数据的变化

数据获取手段的自动化、多样化、智能化，导致数据量越来越巨大

— 数据量越来越巨大

对于海量数据的存储和管理，要求系统具有高度的可扩展性和可伸缩性。传统的分布式数据库和并行数据库在可扩展性和可伸缩性方面明显不足。

— 数据类型越来越多样和异构

从结构化数据扩展到文本、图形图像、音频、视频等多媒体数据，HTML、XML、网页等半结构化/非结构化数据，流数据、队列数据和程序数据等。





Big Data时代到来了！



数据管理与应用面临的挑战

二、应用和需求的变化

• 数据处理和应用领域的扩展

- 从OLTP扩展到→ OLAP→SERACH等;
- 对数据仓库中结构化的海量历史数据的多维分析→对海量非结构化数据的复杂分析和深度挖掘;

• 大数据分析挖掘成为大数据应用中的关键

数据的海量异构，形式繁杂，高速增长，价值密度低等问题阻碍了数据价值的创造。现有的分析挖掘算法

- 缺乏可扩展性
- 缺乏对复杂异构数据的高效分析算法
- 缺乏大规模知识库的支持和应用
- 缺乏能被非技术领域专家理解的分析结果表达方法

对数据的组织、管理、检索和分析都是基础性的挑战。

数据管理与应用面临的挑战

三、相关技术的发展

1. 计算机硬件体系结构的发展十分迅速
 - 处理器已全面进入多核时代；
 - 内存容量越来越大，成本越来越低；
 - 非易失性内存、闪存等技术日益成熟。
2. 以集群为特征的云存储成为大型应用的架构
**数据处理平台由单处理器平台向
多核、大内存，集群，云计算平台转移。**

大数据给数据管理、数据处理和数据分析提出了全面挑战。