数据库技术-绪论

赵亚伟

zhaoyw@ucas.ac.cn

中国科学院大学 大数据分析技术实验室

2017.11.19

绪论—目录

- 数据库系统的发展
- 数据库系统的目标
- 数据库系统的特征
- 研究领域
- 课程采用的例子

从发展历程理解数据库

两个主要问题一个宗旨

- 数据库系统源于计算机系统要解决的两个主要问题
  - 数据的存储
  - 数据的处理
- 数据的存储和处理要遵循一个宗旨
  - 高效（时间和空间的高效，目的：满足实际需求）
- 我们就从这两个问题及一个宗旨来看数据库系统的发展

三个发展阶段

- 按照时间发展，计算机用于数据存储及处理可以大致划分为三个阶段：
  - 人工管理阶段（**50**年代中期以前）
  - 文件系统阶段（**50**年代后期到**60**年代中期）
  - 数据库系统阶段（**60**年代至今）

人工管理阶段

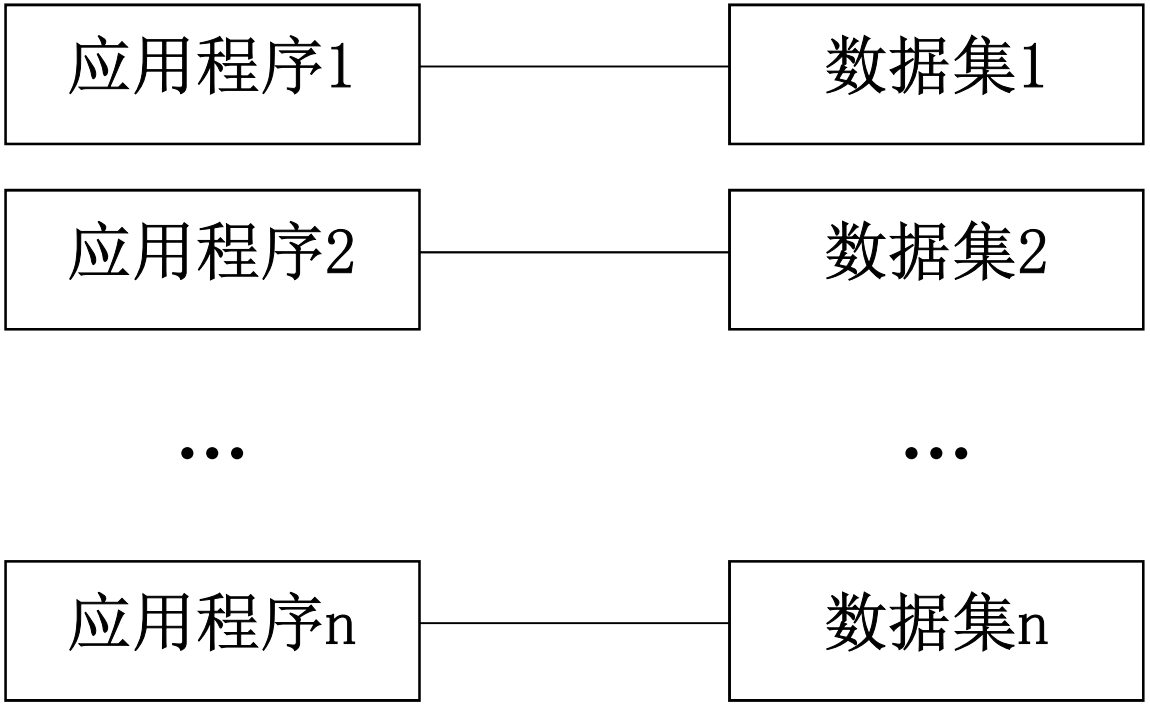
- **50**年代中期以前，计算机技术主要用于科学计算
- 外存比较简陋（只有纸带、卡片、磁带），没有直接存储设备（磁盘），没有操作系统，更没有管理数据的软件，数据处理是批处理。

数据的存储与处理分离

- 在上世纪初，美国人**Hollerith**发明的穿孔卡片用来记录美国的人口普查数据并用机械系统来处理这些卡片，不是现代意义上的计算机，而是自动化的数据处理系统。
- 后来穿孔卡片也被广泛作为数据输入计算机，穿孔卡片可以理解为数据存储介质，而所谓机械系统或计算机就是数据处理环节 。
- 数据的存储与处理两个环节是分离的

人工管理阶段的特点

- 数据不保存（算完就撤）
- 应用程序管理数据
- 数据不共享
- 数据不独立



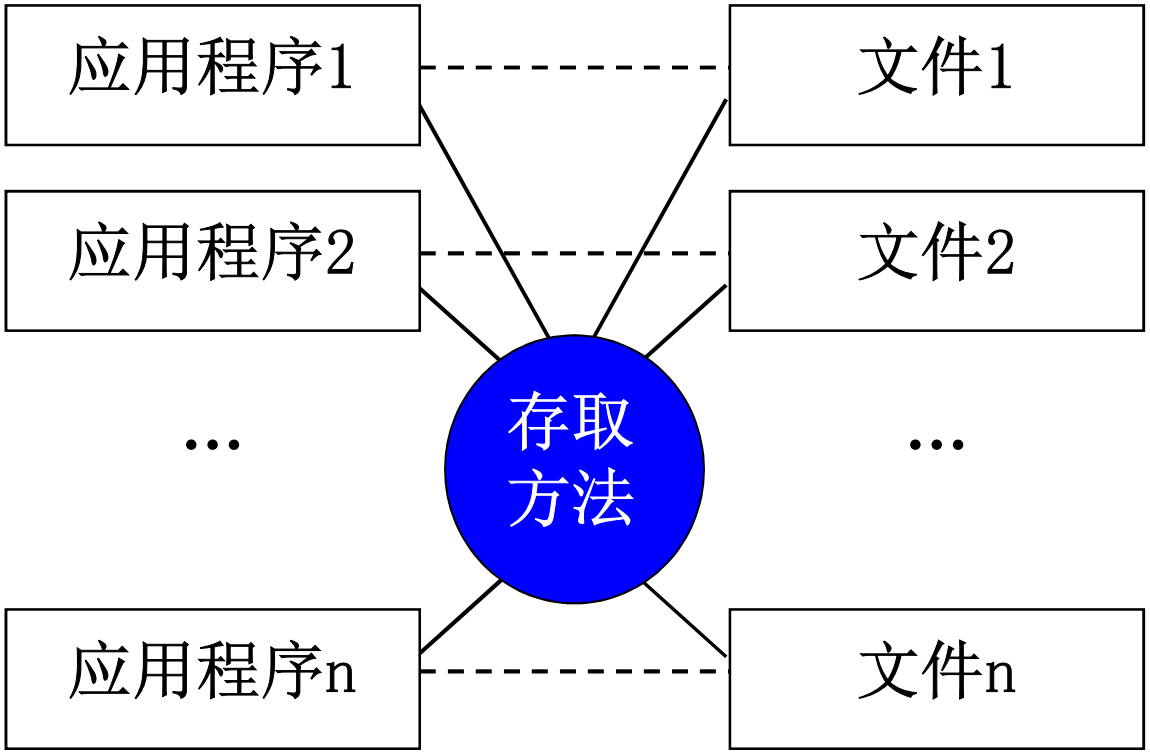
人工管理阶段应用程序与数据之间的关系

文件系统阶段

- 到了**50**年代后期到**60**年代中期硬件有了直接存取设备；
- 软件方面，操作系统中已经有了专门的数据管理软件，一般称为文件系统；
- 处理方式也不仅有批处理，而且能够联机实时处理

文件系统阶段的特点

- 数据可以长期保存
- 由文件系统管理数据
- 数据共享性差，冗余度大
- 数据独立性差



文件系统阶段应用程序与数据之间的关系

数据库系统阶段

- 初级数据库系统阶段（60年代产生，70-80年代流行），又称第一代数据库系统
- 关系数据库系统阶段（产生于70年代，80年代得到充分发展，繁荣至今），又称第二代数据库系统
- 高级数据库系统阶段（80年代中期开始，一直没有繁荣起来），又称第三代数据库系统

- 注意：不能简单理解为数据库系统取代了文件系统，文件系统也具有鲜明的特色和优势。目前的NoSQL数据库采用文件系统进行存储，如GFS、HDFS等。

初级数据库系统阶段

- 这个时期数据库系统蓬勃发展，各种数据库系统相继问世，形成了著名的“数据库时代”。
- 层次模型和网络模型的数据库系统占主流，为统一管理与共享数据提供了有力支撑。
- 缺点：
  - 这两种类型的数据库系统脱胎于文件系统，因此，受文件中的物理结构影响较大，用户在使用数据库时需要对数据的物理结构有详细的了解，这对使用数据库带来了诸多的麻烦。
  - 数据库中表示数据模式的结构方式过于烦琐，数据结构的描述复杂。

关系数据库系统阶段

- 关系型数据库系统的理论出现于**70**年代初（初级数据库系统阶段就已出现），其系统形成于**70**年代中期并在**80**年代得到了充分的发展，它具有简单的结构方式与较少的物理表示，使用与操作又极为方便
- **80**年代，关系数据库逐步取代了层次与网络型数据库成为占主导地位的数据库
- 到目前为止，关系型数据库系统仍占领数据库应用的主导地位。

NoSQL数据库

- NoSQL（Not Only SQL），泛指非关系型的数据库。
- 随着互联网Web2.0网站的兴起，传统的关系数据库在应付Web2.0网站，特别是超大规模和高并发的SNS类型的Web2.0纯动态网站已经显得力不从心，暴露了很多难以克服的问题，而非关系型的数据库则由于其本身的特点得到了非常迅速的发展。
- NoSQL数据库一般是以牺牲一致性来获得高性能的。

NoSQL数据库-MongDB

- **MongoDB**是一个介于关系数据库和非关系数据库之间的产品，是非关系数据库当中功能最丰富，最像关系数据库的。
- 支持的数据结构非常松散，是类似**json**的**bjson**格式，因此可以存储比较复杂的数据类型。
- **Mongo**最大的特点是支持的查询语言类似于面向对象的查询语言，几乎可以实现类似关系数据库单表查询的绝大部分功能，而且还支持对数据建立索引。
- 特点是高性能、易部署、易使用，存储数据非常方便

核心问题—数据模型

- 在数据库领域，数据模型是制约数据库系统的关键因素。
- **E.F. Codd** 博士（**1923-2003**）在**1970**年提出的关系模型充分考虑了企业业务数据的特点，从现实问题出发，为数据库建立了一个坚实的数学基础（关系代数）。（依用治学）
- 在整个计算机软件领域，恐怕难以找到第二个像关系模型这样概念如此简单，但却能带来如此巨大市场价值的技术。

关于模型

- 模型：现实世界特征的模拟和抽象
- 注意：模型并不是一个精确的术语

数据模型

- 数据模型（**Data Model**）：模型的一种类型，是现实世界数据特征的抽象，是按计算机系统的观点对数据进行建模的结果
- 一个重要用途：用于**DBMS**的实现
- 数据模型是数据库系统的核心和基础，**DBMS**是基于某种数据模型的
- 层次模型、网状模型和关系模型

数据模型应满足的条件

- 能比较真实地模拟现实世界
- 容易被人所理解
- 便于在计算机上实现

满足上述三个条件很困难，一般根据不同的应用采用不同的数据模型

数据模型的构成

- 数据模型构成三要素：
  - 数据结构
  - 数据操作
  - 数据的约束条件

数据结构

- 所研究对象类型的集合
- 包括两类：
  - 一类是与数据类型、内容、性质有关的对象，如网状模型中的数据项、记录，关系模型中的域、属性、关系等。数据本身的描述。
  - 一类是与数据之间联系有关的对象，如网状模型中的系型（**Set Type**）。数据之间的描述。
- 系统静态特性的描述

数据操作

- 指对数据库中各种对象（型）的实例（值）允许执行的操作的集合，包括操作及有关的操作规则。
- 数据库主要操作有检索和更新（包括插入、删除、修改）两大类操作。
- 数据模型必须定义这些操作的确切含义、操作符号、操作规则以及实现操作的语言。
- 数据操作是系统动态特性的描述。

数据的约束条件

- 约束条件是一组完整性规则的集合。完整性规则是给定的数据及其联系所具有的制约和依存规则，用以限定符合数据模型的数据状态及状态的变化以保证数据的正确、有效和相容。
- 包括
  - 实体完整性：主属性不为空
  - 参照完整性：关系间的属性取值约束
  - 用户定义完整性：属性的取值范围约束、唯一值约束和属性值间函数约束等
- 关系模型满足上述三个条件，**E.F. Codd**也是从这三个方面论述的。

现实问题的挑战

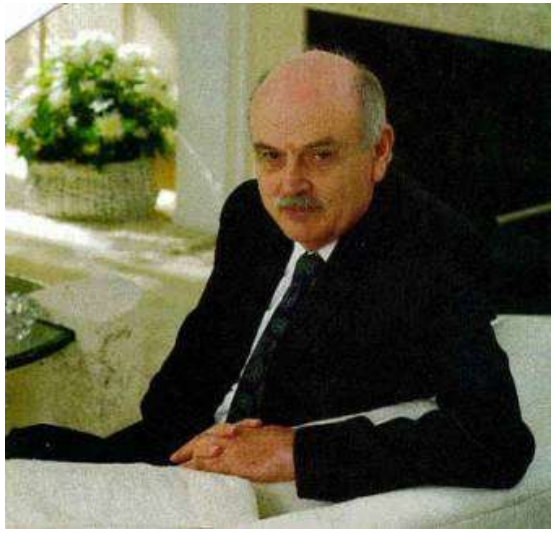
- 关系数据库理论基本成熟后，各大学、研究机构和各大公司在关系数据库管理系统（**RDBMS**）的实现和产品开发中，遇到了一系列技术问题。
- 主要体现在数据库的规模愈来愈大，数据库的结构愈来愈复杂，又有愈来愈多的用户共享数据库的情况下，如何保障数据的完整性、安全性、并发性以及故障恢复的能力，它成为数据库产品是否能够进入实用并最终为用户接受的关键因素。

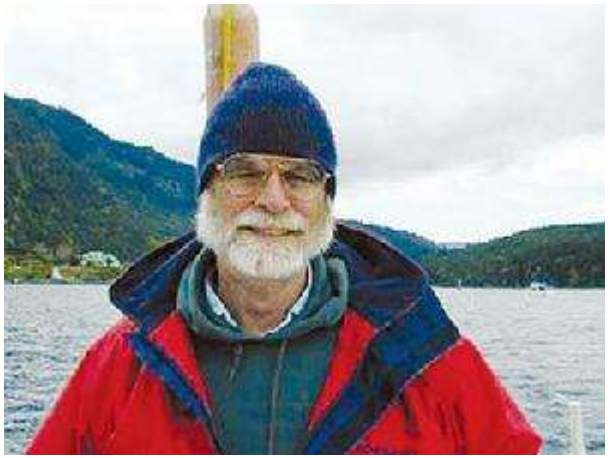
事务处理技术

- **James Gray**在解决这些重大技术问题中发挥了关键作用，使**RDBMS**技术成熟并顺利进入市场。
- 概括地说，**James Gray**解决上述问题的主要技术手段和方法是：把对数据库的操作划分为“事务”的基本单位，一个事务要么全做，要么全不做（即**all-or-nothing** 原则）。
- 具体做法：用户在对数据库发出操作请求时，需要对有关的不同数据“加锁”，防止不同用户的操作之间互相干扰；在事务运行过程中，采用“日志”记录事务的运行状态，以便发生故障时进行恢复；对数据库的任何更新都采用“两阶段封锁”策略。
- 以上围绕事务进行数据处理的方法统称为“事务处理技术”。

造就两个图灵奖

- **E.F Codd和James Gray**在关系模型和事务处理技术上的创造性思维和开拓性工作，使他们成为这一领域公认的权威，并分别于**1981**年和**1998**年成为图灵奖获得者。





关系数据库系统的最大
特点是简单，非常地简单

简单带来的最大好处是易用
易用导致流行，流行导致主流

高级数据库阶段

- 传统关系数据库系统不能完全适用于新领域的应用
- 工程设计、人工智能、多媒体、分布式等领域需要有新的数据库支撑
- **80**年代中期开始，各种适应不同领域的新型数据库系统不断涌现，如工程数据库、多媒体数据库、图形数据库、图像数据库、分布式数据库、数据仓库以及面向对象数据库等。
- 数据仓库和面向对象数据库系统由于其通用性强，适应面广而受到青睐
- 目前，**NoSQL**数据库异军突起，对于大数据以及非结构化的数据处理具有优势

什么是高级数据库阶段？

- 核心问题 — 数据模型
- 管理对象的扩展
  - 数据管理
  - 对象管理（操作或行为管理）
  - 知识管理（规则和推理机制）
- ……

高级数据库阶段的数据库仍然没有成为主流

我国数据库系统的发展

- 初级数据库系统阶段（层次、网络模型）基本没经历
- 自**80**年代中期直接进入关系数据库系统阶段。持别是以**dBASE**与**Foxbase**为代表的关系型数据库系统（小型的、个人数据库或桌面数据库）阶段和近年来以**ORACLE**、**DB2**为代表的数据库系统（大型的、分布式数据库）阶段
- 目前，对高级数据库系统的研究逐步升温，但还没有成功的大规模的产业化商品

绪论—目录

- 数据库系统的发展
- 数据库系统的目标
- 数据库系统的特征
- 研究领域
- 课程采用的例子

数据库系统的目标

- Drawbacks of using file systems to store data：先看看文件系统存在的不足
  - Data redundancy and inconsistency 数据冗余和不一致
    - Multiple file formats, duplication of information in different files 格式多杂，副本多杂，导致数据不一致
  - Difficulty in accessing data 不易存取
    - Need to write a new program to carry out each new task
  - Data isolation — multiple files and formats 数据孤立
  - Integrity problems 完整性不好
    - Integrity constraints (e.g. account balance > 0) become part of program code 完整性约束 如余额不能为负
    - Hard to add new constraints or change existing ones 新约束难加入

- Atomicity of problem 原子性问题导致-更新异常
  - Failures may leave database in an inconsistent state with partial updates carried out 软硬件错误会-数据不一致
  - E.g. transfer of funds from one account to another should either complete or not happen at all 如中途停电
- Concurrent access by multiple users 多用户并发
  - Concurrent accessed needed for performance 并发较快
  - Uncontrolled concurrent accesses can lead to inconsistencies 失控并发出错
    - E.g. two people reading a balance and updating it at the same time
- Security problems 安全问题

- 启示： 事务处理，精确；决策支持，模糊

数据库系统的主要目标就是
解决上述问题！

为解决上述问题
数据库系统提出了一堆概念和方法！

绪论—目录

- 数据库系统的发展
- 数据库系统的目标
- 数据库系统的特征
- 研究领域
- 课程采用的例子

数据库系统阶段的特点

- 数据结构化
- 数据共享性高，冗余度低，易扩充
- 数据独立性高
- 数据由**DBMS**统一管理和控制

数据库系统特点(1): 数据结构化

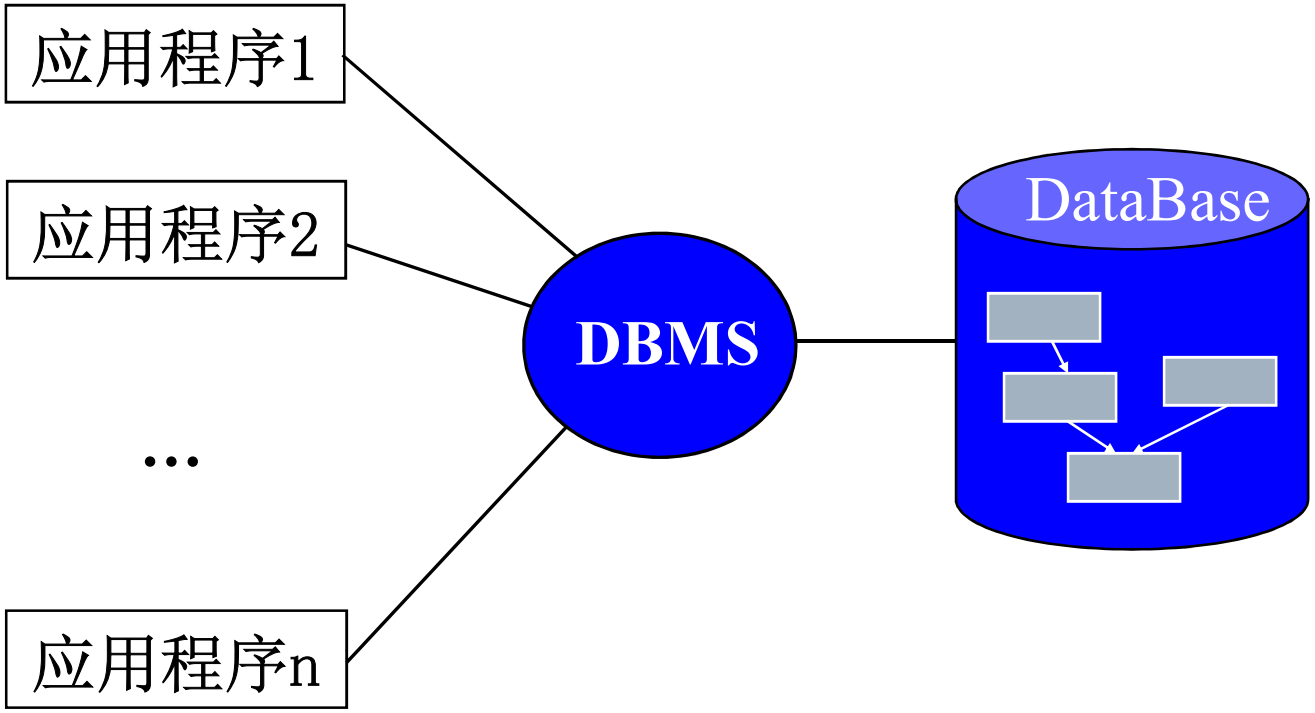
- 结构化是数据库系统的最基本特征
- 结构是信息特征的总结和抽象
- 结构不仅体现一个“数据集”中，还体现在数据集的“关联”上
- 面向整个组织，具有整体结构化特征。小到一个数据项，大到一组记录

数据库系统特点(2): 数据共享性

- 整体的结构性导致数据面向整个系统而不是某个应用，可以被多用户共享（共享容易导致数据不一致）
- 减少数据冗余
- 克服了数据不一致性（事务管理技术）
- 与具体应用无关，很容易引进新的应用

数据库系统特点(3): 数据独立性

- 物理独立性（应用程序与数据存储）：应用程序不需要了解数据的存储
- 逻辑独立性（应用程序与数据逻辑结构）：数据的逻辑结构变更可以不会影响应用程序的运行
- 上述两个独立性由**DBMS**实现
- 数据独立性的目的是解耦合，通过一定的机制来避免存储、结构与应用之间的“连锁反应”。



数据库系统阶段应用程序与数据之间的关系

数据库系统特点(4): DBMS

- **DBMS具有以下数据控制功能：**
  - 数据安全性保护。权限管理
  - 数据完整性检查。正确性、有效性
  - 并发控制
  - 数据恢复，如事务回滚

绪论—目录

- 数据库系统的发展
- 数据库系统的目标
- 数据库系统的特征
- 研究领域
- 课程采用的例子

数据库的研究领域（1）

- 四个主要的研究领域
  - **DBMS**
  - 数据库设计
  - 数据库理论
  - 数据库应用（领域交叉）
- **DBMS**
  - 数据库系统的基础，也是基础软件系统。包括工具软件和中间件。目标是提高系统的可用性、可靠性、可伸缩性，提高性能和提高用户的生产率。

数据库的研究领域（2）

- 数据库设计
  - 主要研究内容是基于某**DBMS**，按照应用的需求，为某一应用设计一个合理的、易用的、高效的数据库及其应用系统。
  - 主要的研究方向是数据库设计方法学和设计工具，包括数据库设计方法、工具和理论的研究，数据模型和数据建模的研究，数据库设计规范和标准的研究等。

数据库的研究领域（3）

- 数据库理论
  - 数据库理论的研究主要集中在关系的规范化理论、关系数据理论等。
  - 近年来，随着人工智能与数据库理论的结合以及并行计算技术等的发展，知识推理、知识发现(KDD)、数据仓库、并行算法等成为新的理论研究方向。

数据库的研究领域（4）

- 数据库与其他技术（领域）的交叉融合
  - 分布式数据库系统：与分布式处理技术融合
  - 并行数据库系统：与并行处理技术融合
  - 知识库系统：与AI技术融合
  - 多媒体数据库系统：与多媒体技术融合
  - 数据仓库：与DSS融合
  - 工程数据库：与CAD/CAM/CIM融合
  - 统计数据库：与统计技术融合
  - 空间数据库：与GIS融合
  - ……

关注热点1：云数据库

- 云计算：云数据库，多种数据模型混合使用
  - 需求：海量数据存储及处理，需要高性能、高可靠性、安全性
  - 解决方法：
    - 云端：超算中心（大型并行计算机、集群），存储能力海量，数据处理性能极高。云端虚拟化为云客户端独占模式（虚拟化：似乎是，以假乱真）。
    - 云客户端：“瘦骨嶙峋”，价格便宜（甚至免费），云终端。低成本
    - 网络：高速通信、**WLAN**（**wifi**，移动、自由）、随时随地。
    - 商业模式：卖服务
  - 瓶颈：安全性、传输效率（带宽）、通信费用、覆盖范围等。

关注热点2：大数据

- 大数据(big data)
  - 指的是所涉及的数据规模巨大到无法通过目前主流软件工具，在合理时间内达到查询、管理、处理、并整理成为帮助企业经营决策更积极目的的信息（数据、规则、知识）
- 大数据的4V特点：
  - **Volume（PB级，2^{20}GB，≈10^6GB）**
  - **Variety（结构化、非结构化，MapReduce、Key-value）**
  - **Value（最为重要，DW、DM）**
  - **Velocity（高速处理）**
- 大数据的价值挖掘是大数据概念存在的价值
- **NoSQL数据库：HBase、Mongodb、Casssandra**

关注热点3：物联网数据库

- 物联网：实时数据库
  - 需求：物物相连，随时反映物的状态以及物物之间的关系，跨空间，严格要求时间
  - 解决方法：数据模型引入时间概念，如某以商品何时何地被销售出去，库存在某一时刻的状态等。数据量可能会大幅度增加，海量存储，云数据库。
  - 瓶颈：网络通信的可靠性、数据处理的实时性、数据采集的实时性等。

绪论—目录

- 数据库系统的发展
- 数据库系统的目标
- 数据库系统的特征
- 研究领域
- 课程采用的例子

课程采用的例子说明

- 课程采用《数据库系统概念》提供的一个关于某银行账户管理信息系统的数据库，一个关系数据库。
- 该例子短小精悍，非常适合数据库的课程教学。

本课程采用的通例-表

customer_name	*customer_street*	*customer_city*
Adams	Spring	Pittsfield
Brooks	Senator	Brooklyn
Curry	North	Rye
Glenn	Sand Hill	Woodside
Green	Walnut	Stamford
Hayes	Main	Harrison
Johnson	Alma	Palo Alto
Jones	Main	Harrison
Lindsay	Park	Pittsfield
Smith	North	Rye
Turner	Putnam	Stamford
Williams	Nassau	Princeton

Customer-客户

account_number	*branch_name*	*balance*
A-101	Downtown	500
A-102	Perryridge	400
A-201	Brighton	900
A-215	Mianus	700
A-217	Brighton	750
A-222	Redwood	700
A-305	Round Hill	350

Account-账户

customer_name	*account_number*
Hayes	A-102
Johnson	A-101
Johnson	A-201
Jones	A-217
Lindsay	A-222
Smith	A-215
Turner	A-305

Depositor-存款人

loan_number	*branch_name*	*amount*
L-11	Round Hill	900
L-14	Downtown	1500
L-15	Perryridge	1500
L-16	Perryridge	1300
L-17	Downtown	1000
L-23	Redwood	2000
L-93	Mianus	500

Loan-贷款

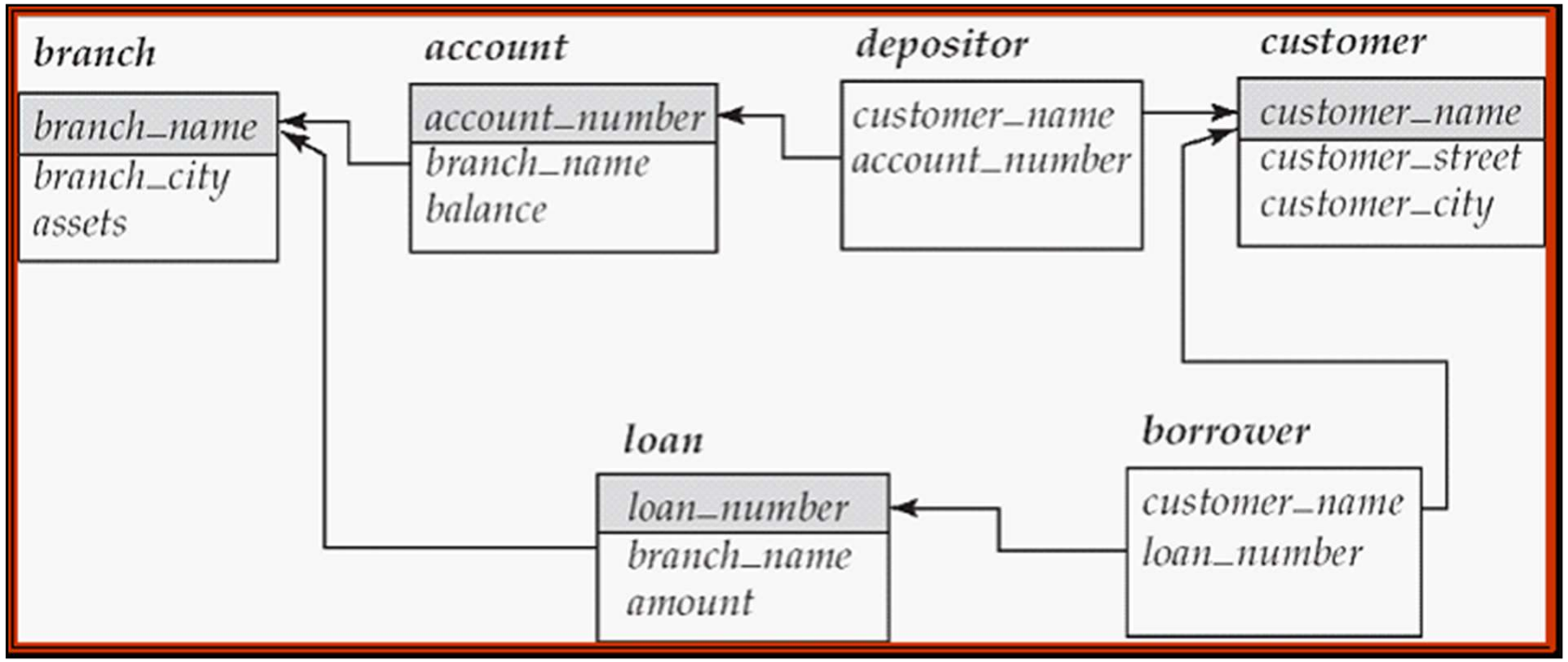
branch_name	*branch_city*	*assets*
Brighton	Brooklyn	7100000
Downtown	Brooklyn	9000000
Mianus	Horseneck	400000
North Town	Rye	3700000
Perryridge	Horseneck	1700000
Pownal	Bennington	300000
Redwood	Palo Alto	2100000
Round Hill	Horseneck	8000000

Branch-支行

customer_name	*loan_number*
Adams	L-16
Curry	L-93
Hayes	L-15
Jackson	L-14
Jones	L-17
Smith	L-11
Smith	L-23
Williams	L-17

Borrower贷款人

本课程采用的通例-数据库模式



总结

- 数据库技术发展大致经历了三个阶段，关系数据库是主流
- 什么是第三代数据库系统仍无定论
- 核心问题是数据模型，数据模型是关键
- 数据库技术与其他技术或领域交叉融合会产生新的应用，但目前成功应用仍没有脱离关系数据库的基础理论
- 数据库系统是一个大家族，很多新的数据库不断涌现，很容易令人迷惑，数据模型是最后一根“稻草”，抓住它就抓住了问题的实质
- 新一代数据库将造就另一个图灵奖？

背景知识

- 文献：
  - **E.F.Codd, “A Relational Model of Data for Large Shared Data Banks”, Communication of the ACM, Volume 13, Number 6(1970), pages 377-387**
- 工具：商业化的数据库系统
  - **IBM DB2**
  - **Oracle**
  - **Microsoft SQL Server**
  - **Informix （IBM）**
  - **Sybase**
- 资料：**ACM的数据管理兴趣组，www.acm.org/sigmod，国内的《软件学报》、《计算机学报》等杂志资源已开放。**