数据科学导论

Introduction to Data Science

开篇实例

100年前

50年前

今天

从图书馆找一本书……

第7章
数据库系统

1 数据库的起源与发展

2 关系数据库

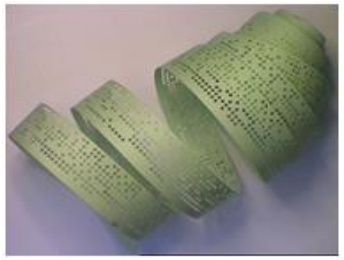
3 数据仓库与OLAP

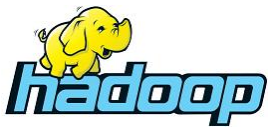
4 SQL语言

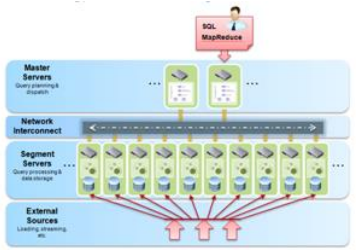
- 维基百科：数据管理，即对数据资源的管理。 而数据资源管理，致力于发展处理企业数据生命周期的适当的建构、策略、实践和程序。
- 百度百科：数据管理是利用计算机硬件和软件技术对数据进行有效的收集、存储、处理和应用的过程。其目的在于充分有效地发挥数据的作用。
- 数据管理经历了四个发展阶段：
  - 人工管理、文件系统、数据库系统、大数据管理

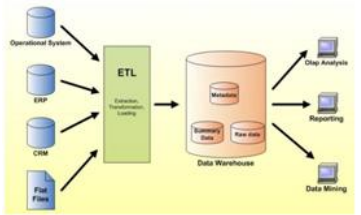
什么是数据管理

- 穿孔纸带
- 文件系统
- 数据库管理系统（DBMS）
  - 网状数据库、层次数据库、关系数据库
- 面向对象数据库
- 决策支持系统和数据仓库
- MPP数据库
- Hadoop/Spark生态系统



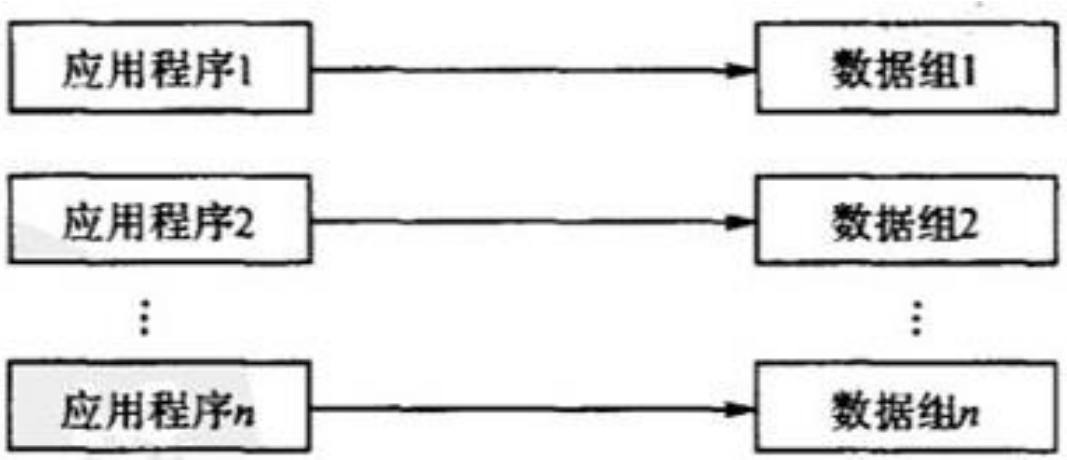






数据管理系统的演变

人工管理 → 文件系统 → 数据库系统 → 大数据管理系统

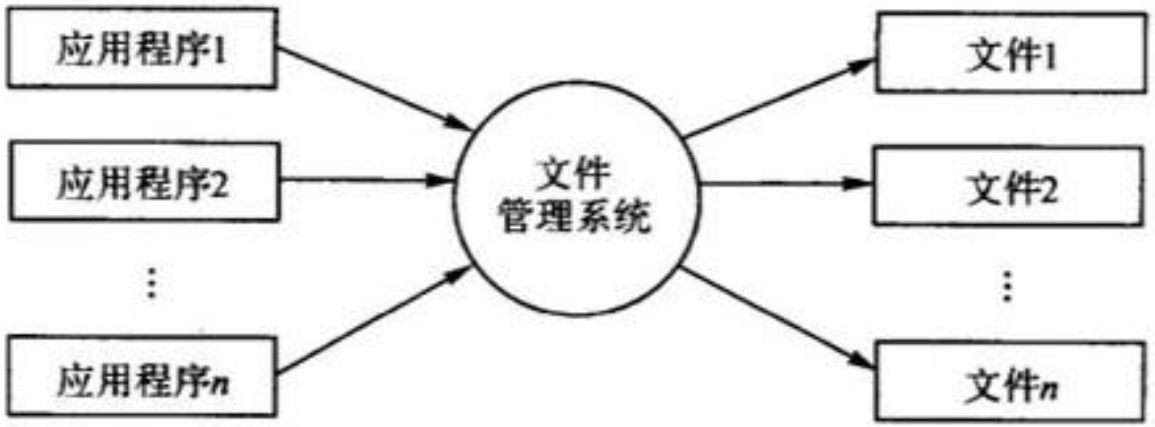


程序和数据的关系

- 数据不保存
- 数据不共享
- 数据不独立
- 应用程序管理数据

数据管理历程

人工管理 → 文件系统 → 数据库系统 → 大数据管理系统



文件管理系统

数据长期储存

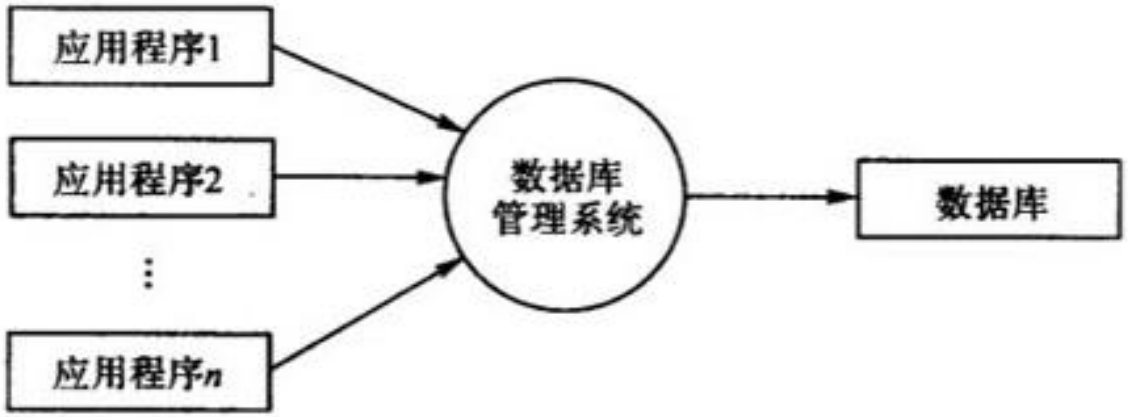
由文件系统管理数据

数据共享性差且冗余度大

文件面向应用

数据管理历程

人工管理 → 文件系统 → 数据库系统 → 大数据管理系统



数据库管理系统

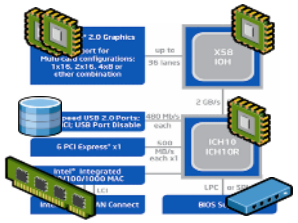
DBMS出现

数据冗余度减少

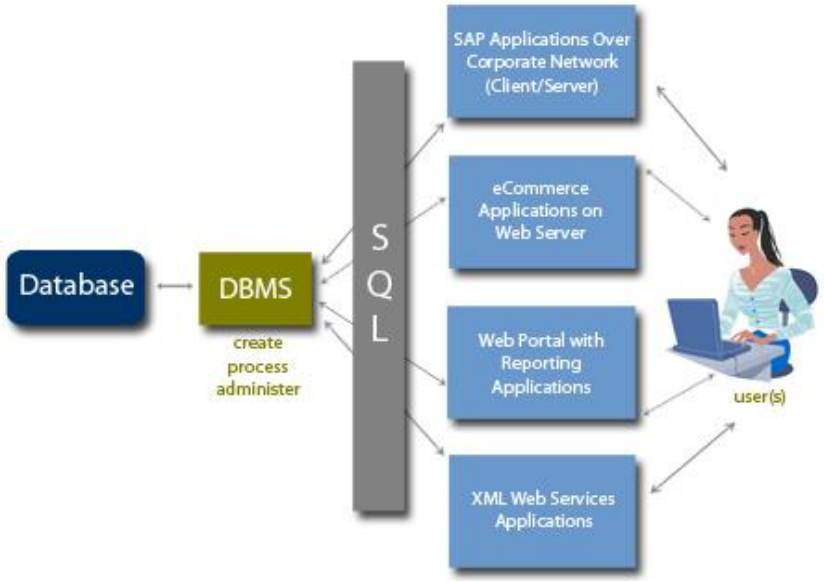
数据充分共享

数据管理历程

人工管理 → 文件系统 → 数据库系统 → 大数据管理系统



硬件基础设施
数据库服务器



数据库管理软件
DBMS

数据管理历程

人工管理 → 文件系统 → 数据库系统 → 大数据管理系统

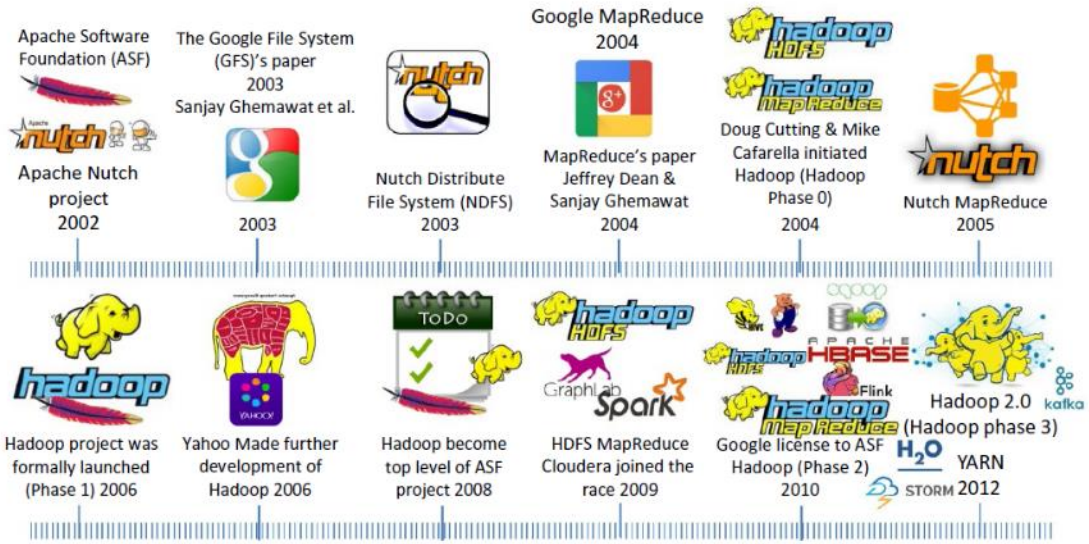
- 数据增长速度非常迅速
- 难以满足高并发读写的需求
- 难以满足对海量数据高效率存储和访问的需求
- 难以满足对数据库高可扩展性和高可用性的需求

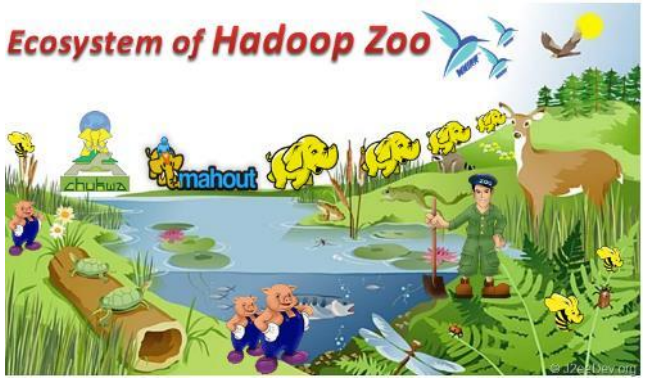
数据库和数据仓库的挑战

数据管理历程

人工管理 → 文件系统 → 数据库系统 → 大数据管理系统

- 由于大数据要处理大量、非结构化的数据，所以在各处理环节中都可以采用并行处理。
- 目前，Hadoop、MapReduce和Spark等分布式处理方式已经成为大数据处理各环节的通用处理方法。



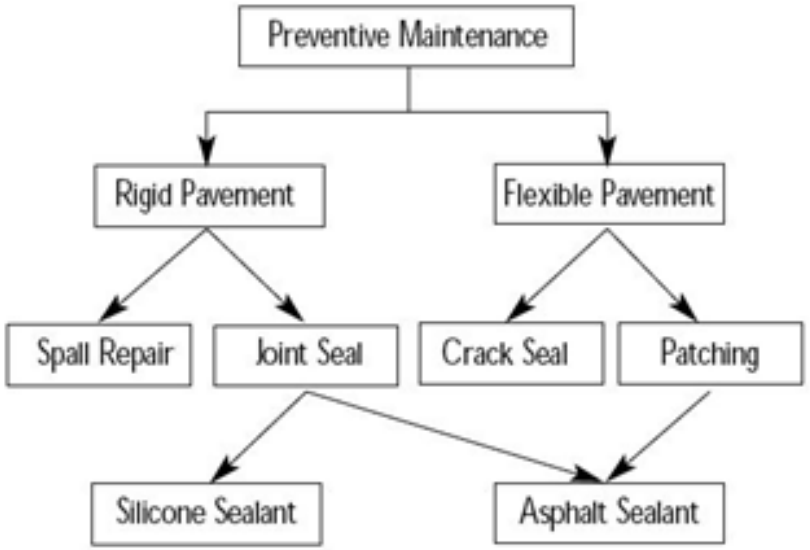


大数据技术生态

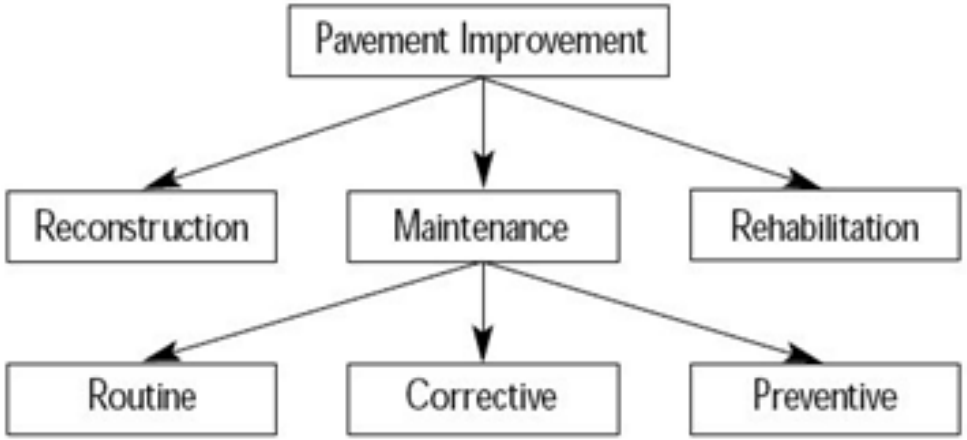
数据管理历程

第7章
数据库系统

Network Model



Hierarchical Model



Relational Model

Activity Code	Activity Name
23	Patching
24	Overlay
25	Crack Sealing

Key = 24

Activity Code	Date	Route No.
24	01/12/01	I-95
24	02/08/01	I-66

Date	Activity Code	Route No.
01/12/01	24	I-95
01/15/01	23	I-495
02/08/01	24	I-66

数据库系统的类型

1，信息准则：关系数据库中的所有信息都应在逻辑层上用表中的值显式的表示。

2，保证访问准则：依于表名，主键和列名，保证能以逻辑方式访问数据库中的每个数据项。

3，空值的系统化处理：RDBMS支持空值（不同于空的字符串或空白字符串，并且不为0）系统化的表示缺少的信息，且与数据类型无关。

4，基于关系模型的联机目录：数据库的描述在逻辑上应该和一般数据采用同样的方式，使得授权用户可以使用查询一般数据所用的关系语言来查询数据库的描述信息。

5，合理广泛的子语言准则：一个关系系统可以具有几种语言和多种终端使用方式（表格填空方式，命令方式等）。

6，视图更新准则：所有理论上可更新的视图也应该允许由系统更新。

Codd十二条法则

7，高阶的插入，更新和删除：把一个基本关系或导出关系作为一个操作对象进行数据的检索以及插入，更新和删除。

8，数据的物理独立性：无论数据库的数据在存储表示上或存取方法上做任何变化，应用程序和终端活动要都保持逻辑上的不变性。

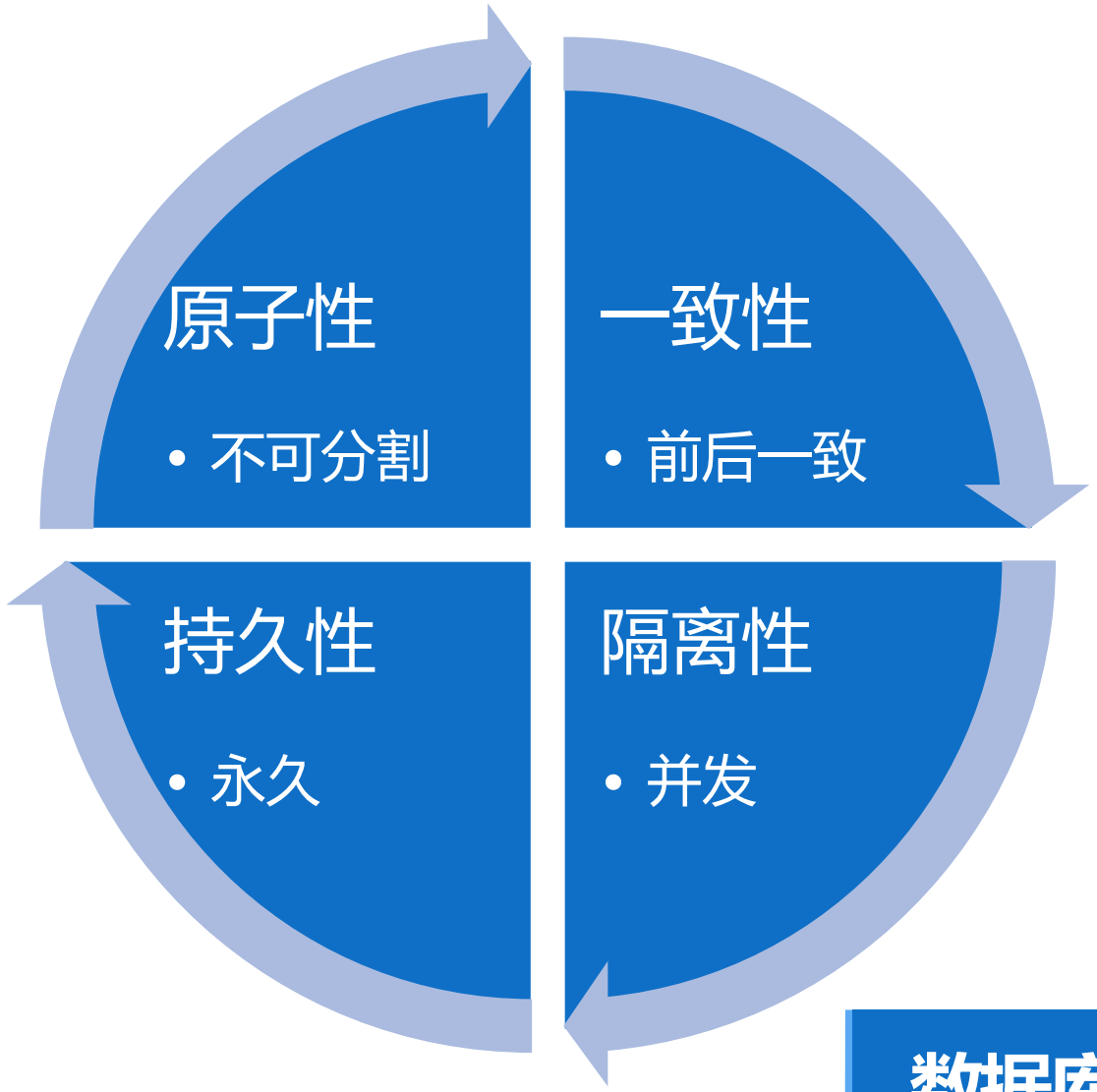
9，数据的逻辑独立性：当基本表中进行理论上信息不受损害的任何变化时，应用程序和终端和终端活动都要保持逻辑上的不变性。

10，数据完整的独立性：关系数据库的完整性约束必须是用数据子语言定义并存贮在目录中的，而不是在应用程序中加以定义的。

11，分布的独立性：一个RDBMS应该具有分布独立性。用户不必了解数据库是否是分布式的。

12，无破坏准则：若RDBMS有某种低级语言，这一低级语言不能违背或绕过完整性准则以及高级关系语言表达的约束。

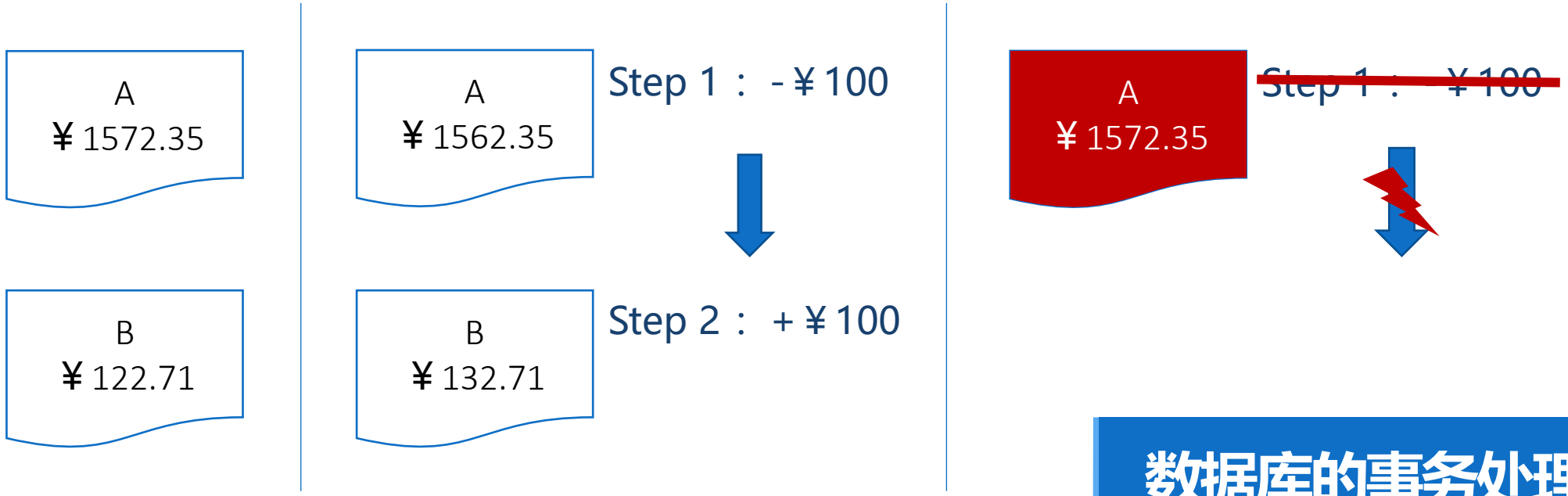
Codd十二条法则



数据库的事务处理

原子性 | 一致性 | 独立性 | 持久性

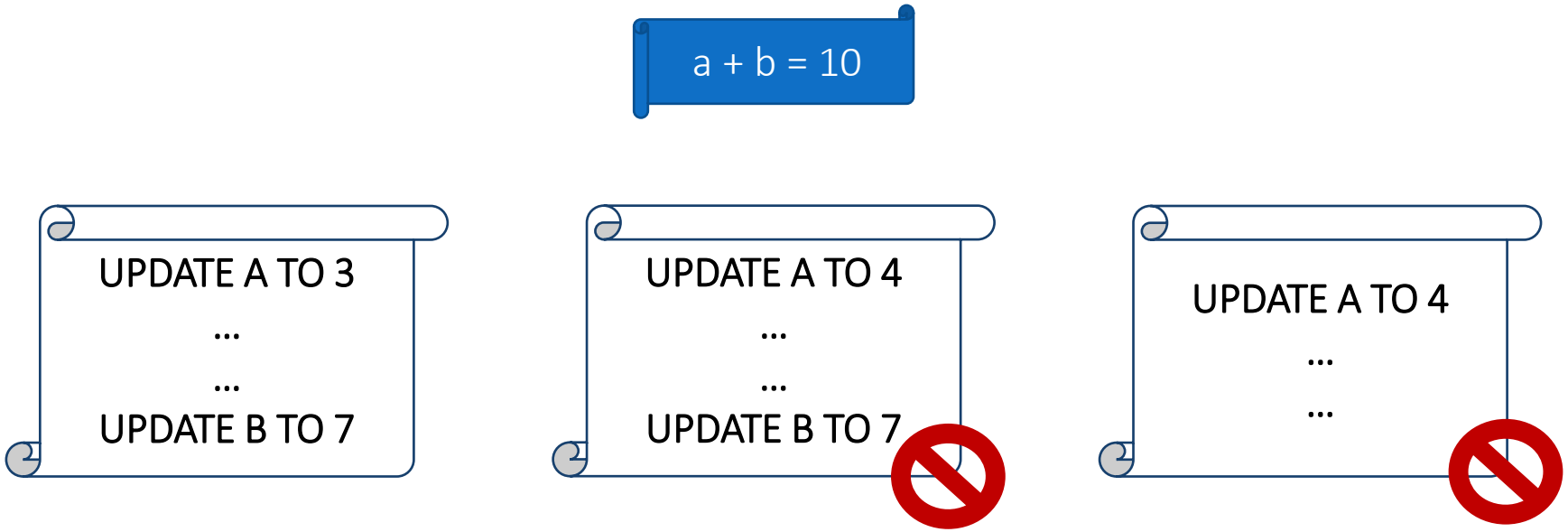
- 原子性很容易理解，就是说事务里的所有操作要么全部做完，要么都不做，事务成功的条件是事务里的所有操作都成功，只要有一个操作失败，整个事务就失败，需要回滚。



数据库的事务处理

原子性 | 一致性 | 独立性 | 持久性

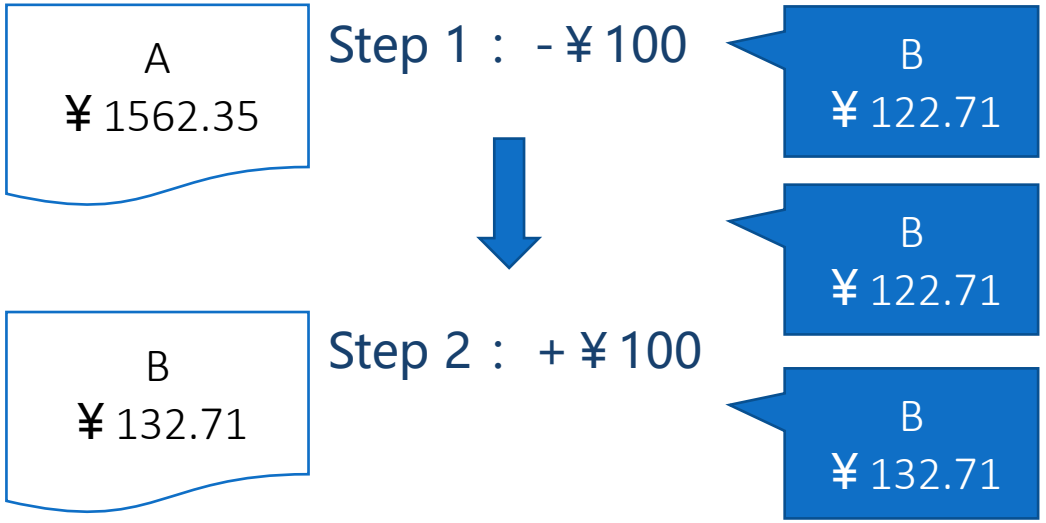
一致性就是说数据库要一直处于一致的状态，事务的运行不会改变数据库原本的一致性约束。



数据库的事务处理

原子性 | 一致性 | 独立性 | 持久性

所谓的独立性是指并发的事务之间不会互相影响，如果一个事务要访问的数据正在被另外一个事务修改，只要另外一个事务未提交，它所访问的数据就不受未提交事务的影响。



数据库的事务处理

原子性 | 一致性 | 独立性 | 持久性

- 持久性是指一旦事务提交后，它所做的修改将会永久保存在数据库上，即使出现宕机也不会丢失。

- 总结：
  - 这些原则解决了数据的一致性、系统的可靠性等关键问题，为关系数据库技术的成熟以及在不同领域的大规模应用创造了必要的条件。

数据库的事务处理

student

学号	姓名	班级	年龄	性别	住址	课号	电话
100	张三	计91	20	男	上海杨浦	上海	89150
200	李四	计92	19	男	上海徐汇	上海	88888
300	王五	计93	18	女	上海浦东	上海	77777
400	赵六	计94	19	女	上海静安	上海	99999
500	刘七	计95	21	男	上海普陀	上海	88666

主键

course

课号	课程名	地点	教师
1	DB	5101	周老师
2	DB	5102	钱老师
3	DM	5103	金老师

主键

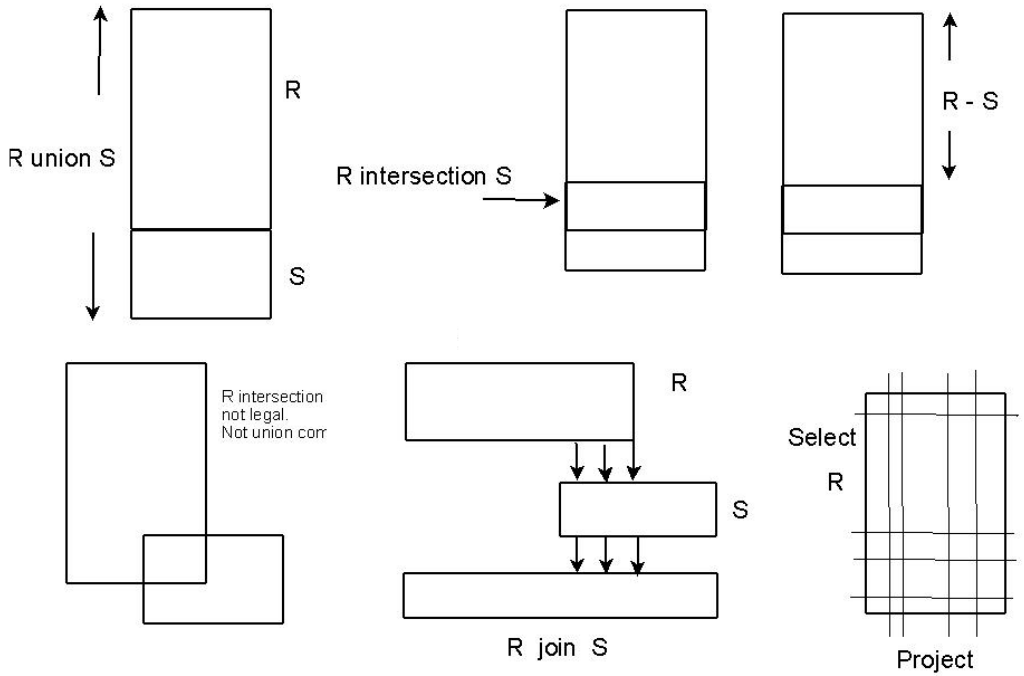
主键

ID	学号	课号	分数
1	100	1	99
2	200	1	98
3	300	2	97

grade

关系

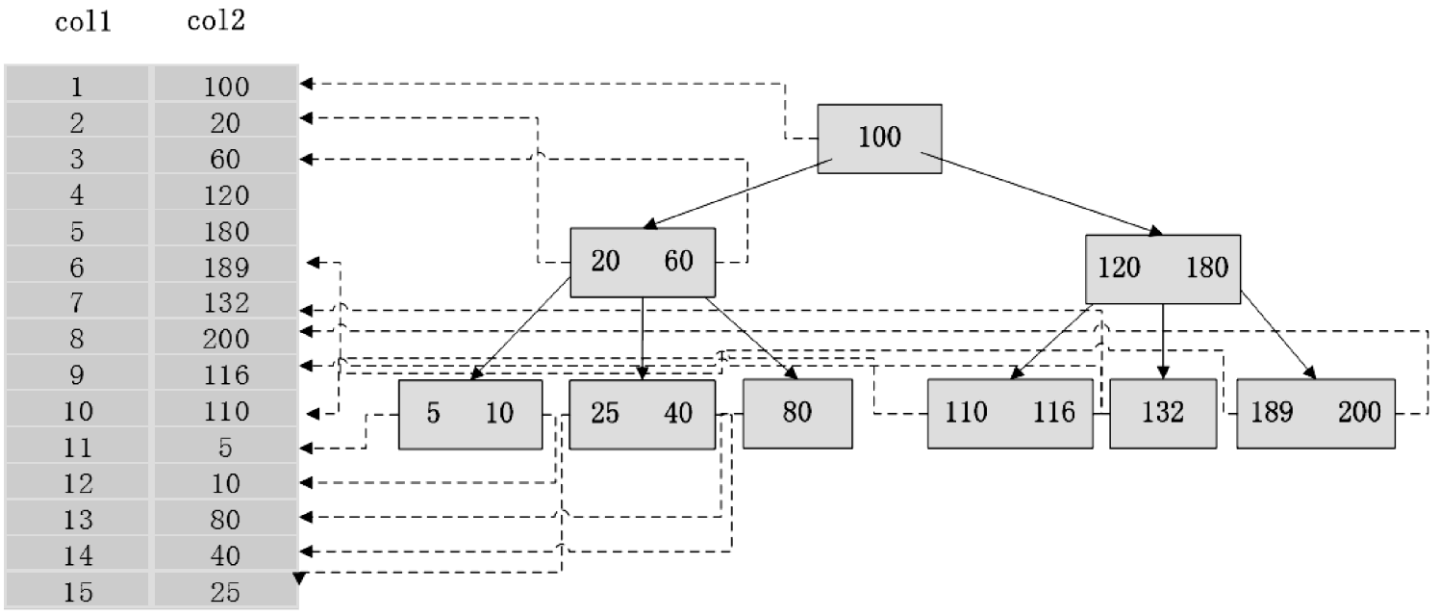
- **关系代数**是一种抽象的查询语言，用对关系的运算来表达查询，作为研究关系数据语言的数学工具。
- 集合运算
  - 并、交、差、广义笛卡尔积
- 关系运算
  - 选择（Selection）
  - 投影（Projection）
  - 连接（Join）
  - 除（Division）



关系代数

- DBMS为每个主键建立一个索引
- 为具有唯一约束性的字段自动建立索引
- 对于经常检索的字段，可为其建立额外的索引

✓ 快速查询
✓ 对排序数据的即时访问



索引

第7章
数据库系统

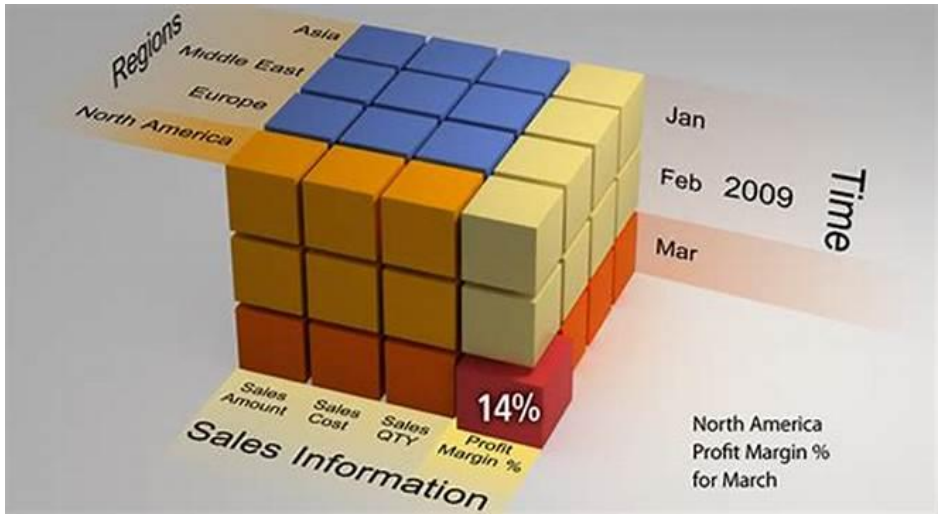
1 数据库的起源与发展

2 关系数据库

3 **数据仓库与OLAP**

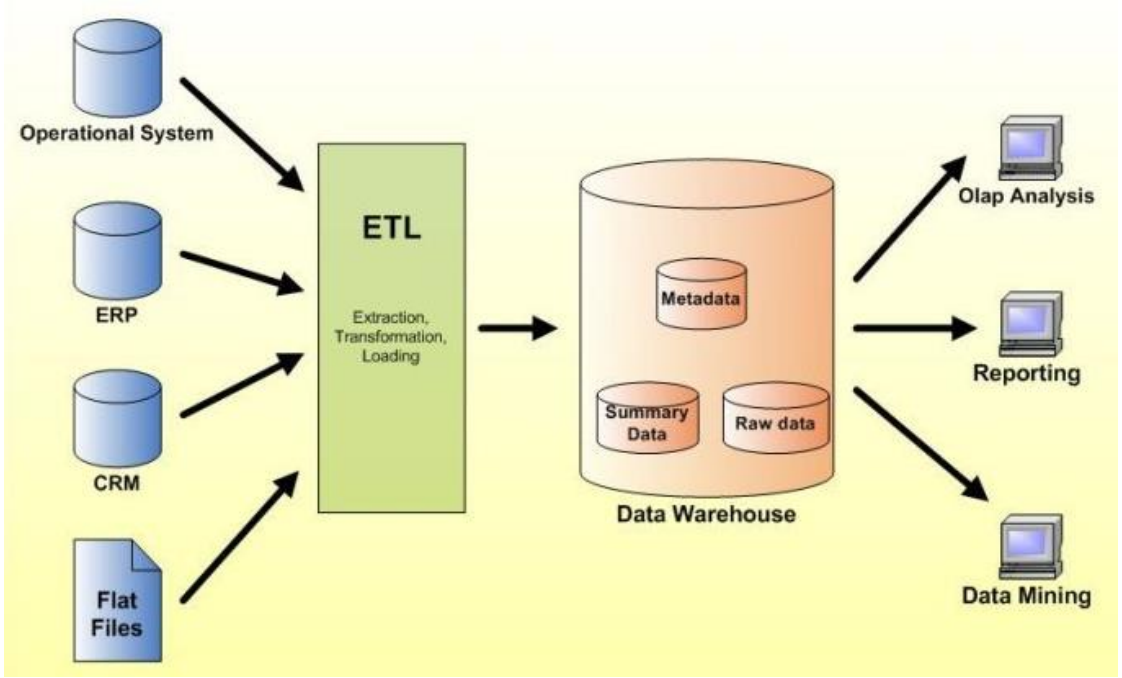
4 SQL语言

- 数据仓库一词最早是在1990年，由Bill Inmon提出的。
- 数据仓库的四个基本特征：
  - 数据仓库的数据是面向主题的（Subject Oriented）
  - 数据仓库的数据时集成的（Integrate）
  - 数据仓库的数据不可更新（Non-Volatile）
  - 数据仓库的数据时随时间不断变化（Time Variant）的



什么是数据仓库

- **数据仓库**是是一个面向主题的、集成的、相对稳定的、反映历史变化的数据集合，用于支持管理决策。
- 数据仓库的四种类型
  - 传统数据仓库
  - 实时处理数据仓库
  - 关联发现数据仓库
  - 数据集市



什么是数据仓库

数据处理的类型

- 操作型处理(OLTP)：数据的收集、整理、存储、查询和增、删、改操作。
- 分析型处理(OLAP)：数据的再加工，往往要访问大量的历史数据，进行复杂的统计分析。

数据仓库产生的原因

- 是数据库系统的主要应用
- 特点：数据存取频率高、响应时间要快、存取数据量小、数据存储正确可靠。
- 为了有效地对事务进行处理，数据库管理系统在技术和管理上采取了很多措施：
  - 提出了事务的概念
  - 采用日志、备份等恢复技术和并发控制技术
  - 采用索引技术快速定位数据

操作型数据处理

与传统数据库的区别

- 典型的分析型应用就是决策支持系统。
- 需要具备的基本功能是：建立各种数学模型，对数据统计分析，得出有用的信息作为决策的依据。
- 常规应用实例：某产品的销售经理希望通过调整该产品在各零售店的分配数量来扩大其销售量。
  - 需要查询历史数据库中各类零售店最近若干年（例如5年）内每天的销售记录。
  - 统计运算计算出近5年来各店的年度销售量。
  - 比较确定销售量增长较快的零售店。
- 决策支持系统：需要花数小时甚至更长时间的处理、需要遍历数据库中的大部分数据，进行复杂的计算，需要消耗大量的系统资源。

分析型数据处理

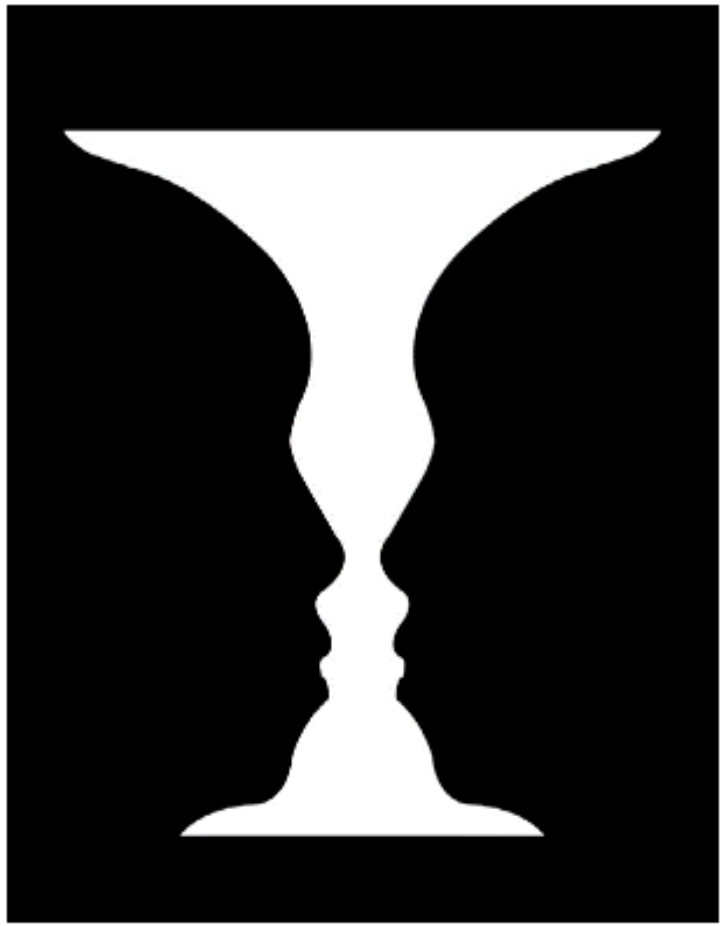
与传统数据库的区别

操作型数据	分析型数据
细节的	综合的，或提炼的
当前数据	历史数据
更新的	不可更新，只读的
生命周期符合软件开发生命周期	以数据为中心的生命周期
对性能要求高	对性能要求宽松
一个时刻操作一个单元	一个时刻操作一个集合
事务驱动	分析驱动
面向应用	面向分析
一次操作数据量小，计算简单	一次操作数据量大，计算复杂
支持日常操作	支持管理需求

OLTP和OLAP的对比

与传统数据库的区别

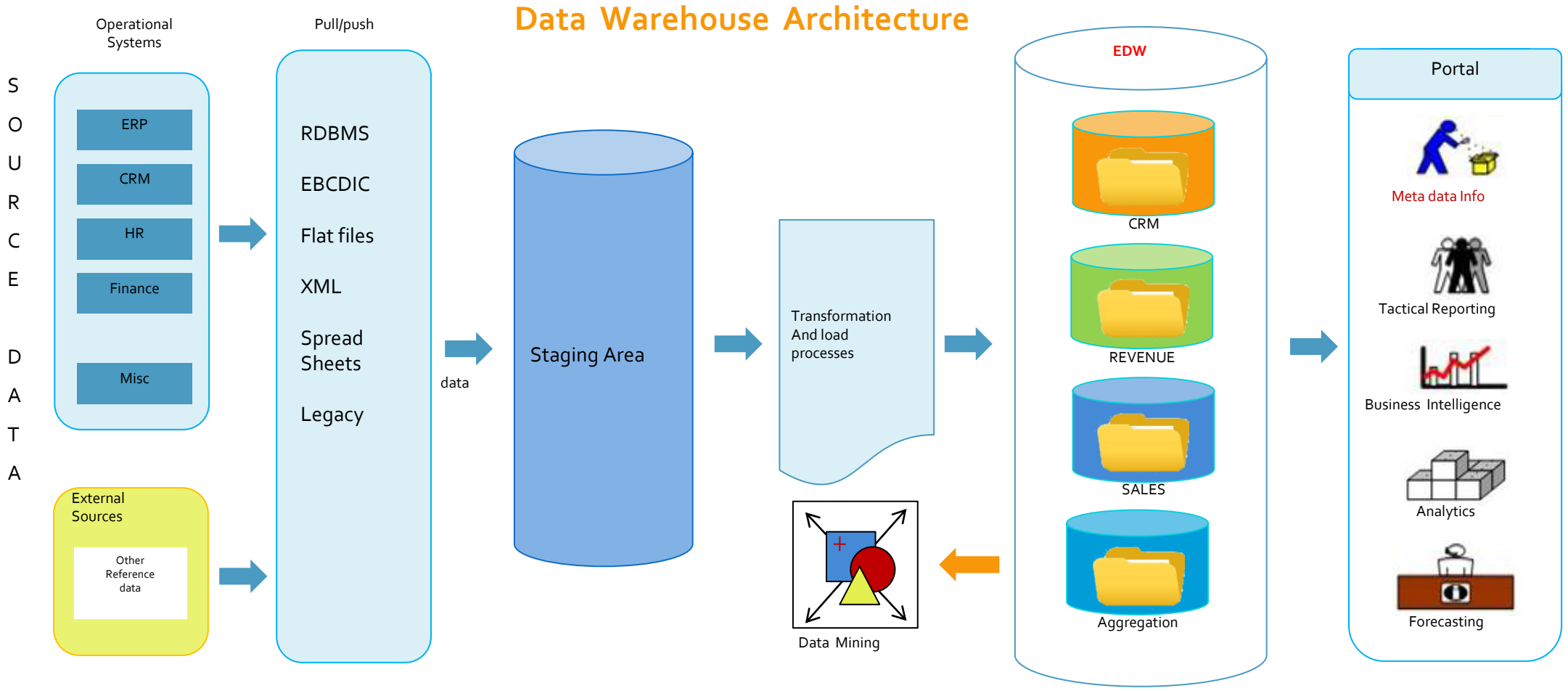
Do You Actually Know What a Data Warehouse Is?



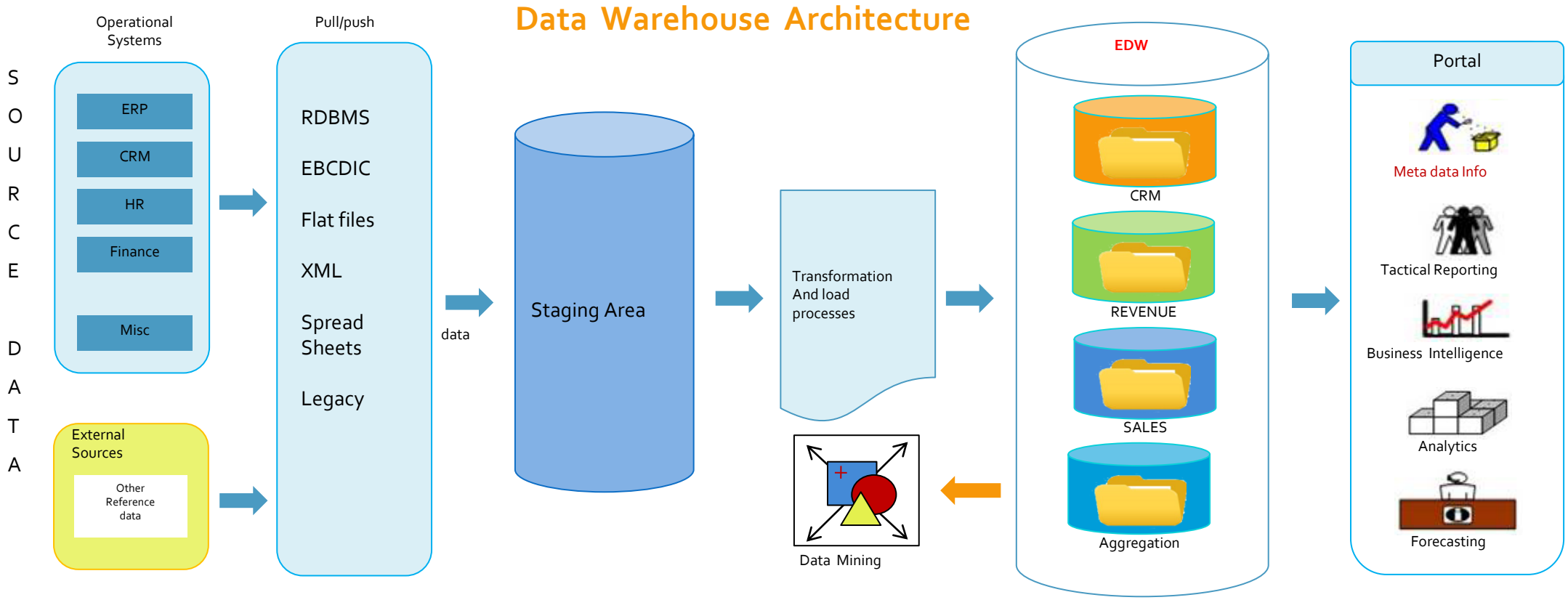
- Consolidated, integrated, subject-oriented, time-variant data management solution.
- It never was …
  - A database.
  - A DBMS.
  - An appliance.
  - Built in.
  - … done.

Things are not always what you first see, first experience, first build, first encounter.

Gartner


Data Warehouse Architecture
SOURCE DATA
Operational Systems
ERP
CRM
HR
Finance
Misc
External Sources
Other Reference data
Pull/push
RDBMS
EBCDIC
Flat files
XML
Spread Sheets
Legacy
data
Staging Area
Transformation And load processes
Data Mining
EDW
CRM
REVENUE
SALES
Aggregation
Portal
Meta data Info
Tactical Reporting
Business Intelligence
Analytics
Forecasting

传统数据仓库的架构



面临的挑战：

1. 数据量增长过快，导致运算效率下降
2. 数据抽取处理的代价过高，无法在统一的视图下处理
3. 无法处理多种类型的数据
4. 不具备进行搜索或关联分析以发现隐藏关系的能力
5. 不具备数据挖掘等高级分析的能力

传统数据仓库的架构

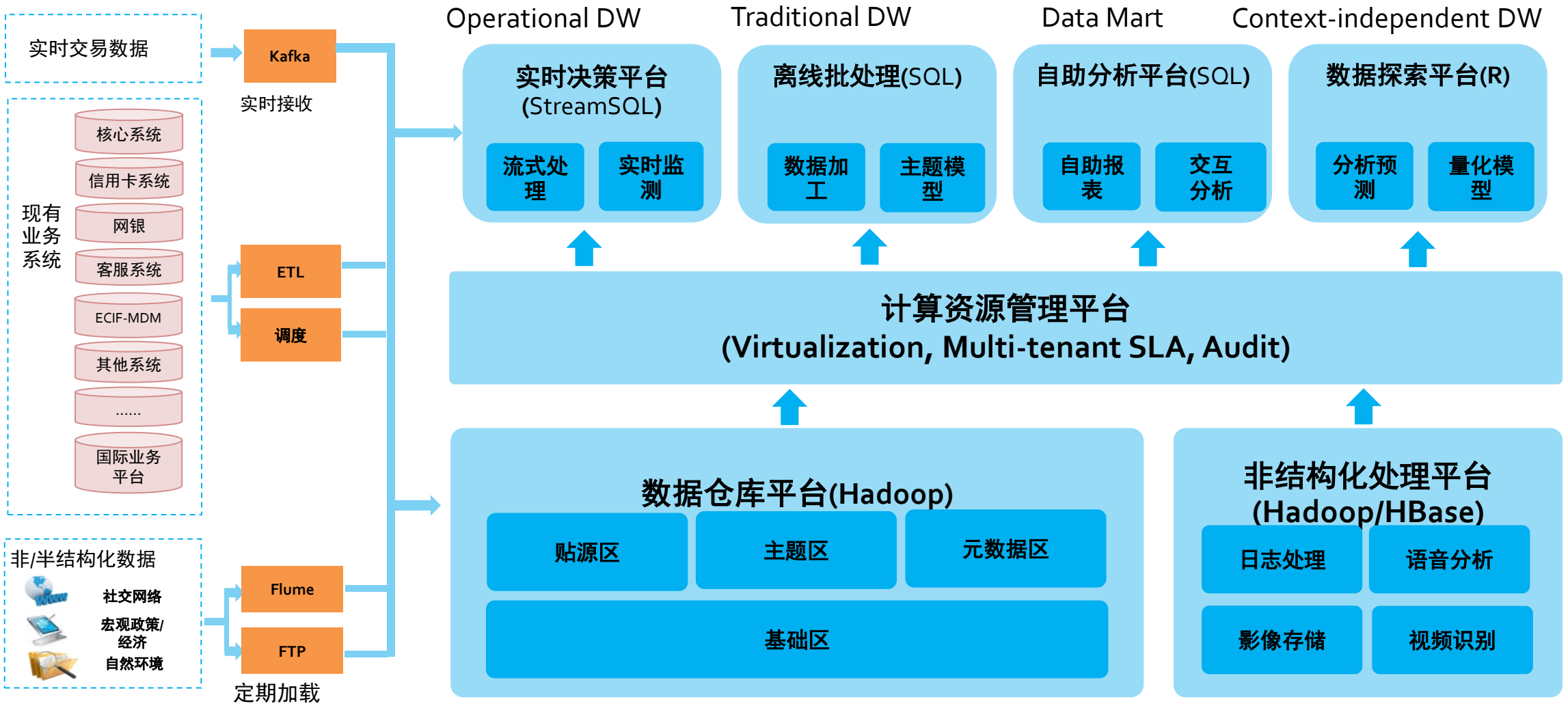
数据增长速度迅速

数据源类型众多

需要数据库虚拟化技术

需要数据挖掘和机器学习的支持

数据仓库架构的挑战

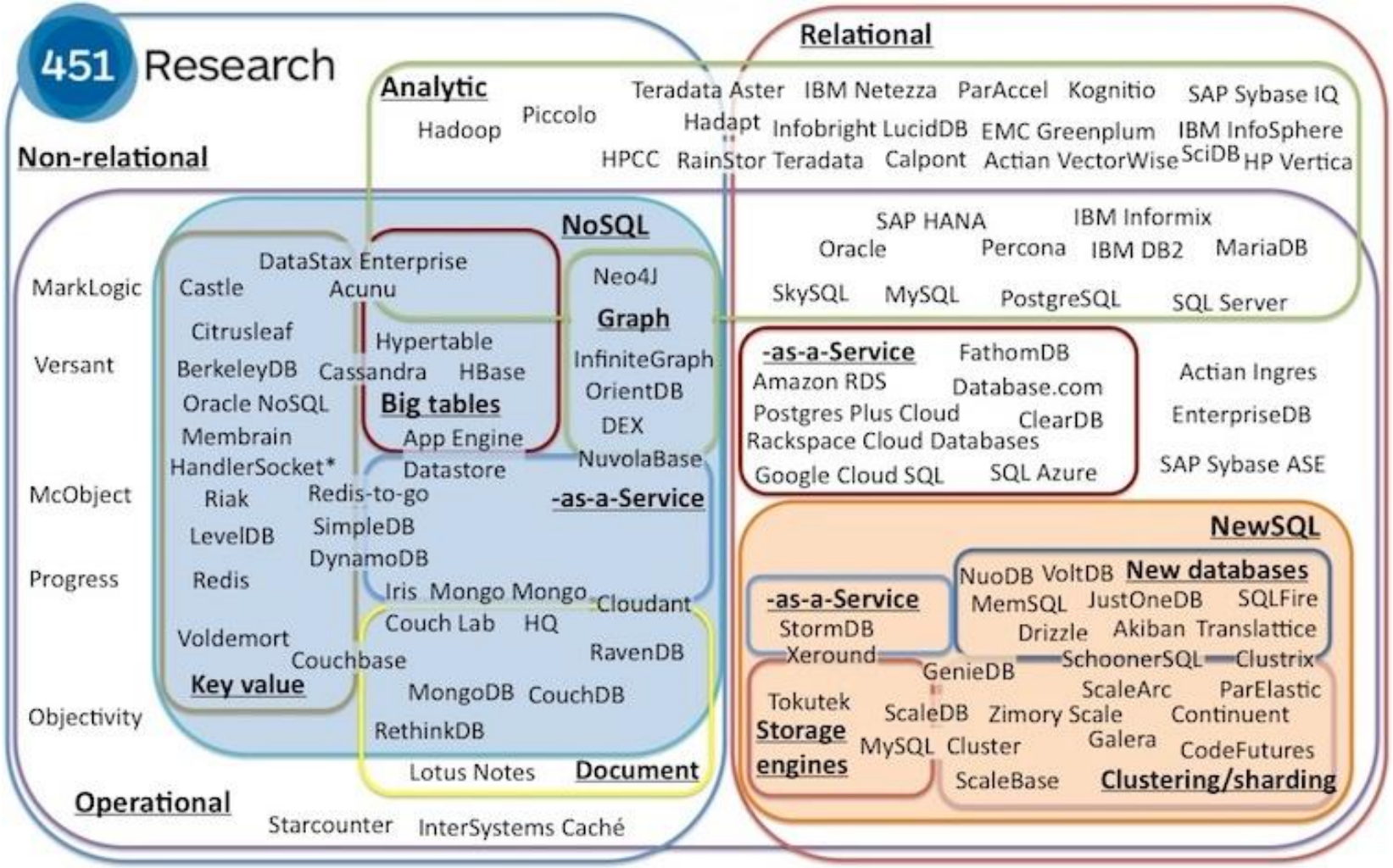


基于大数据技术的数据仓库架构

- 关系型数据库的局限性
  - 难以满足高并发读写的需求
  - 难以满足对海量数据高效率存储和访问的需求
  - 难以满足对数据库高可扩展性和高可用性的需求

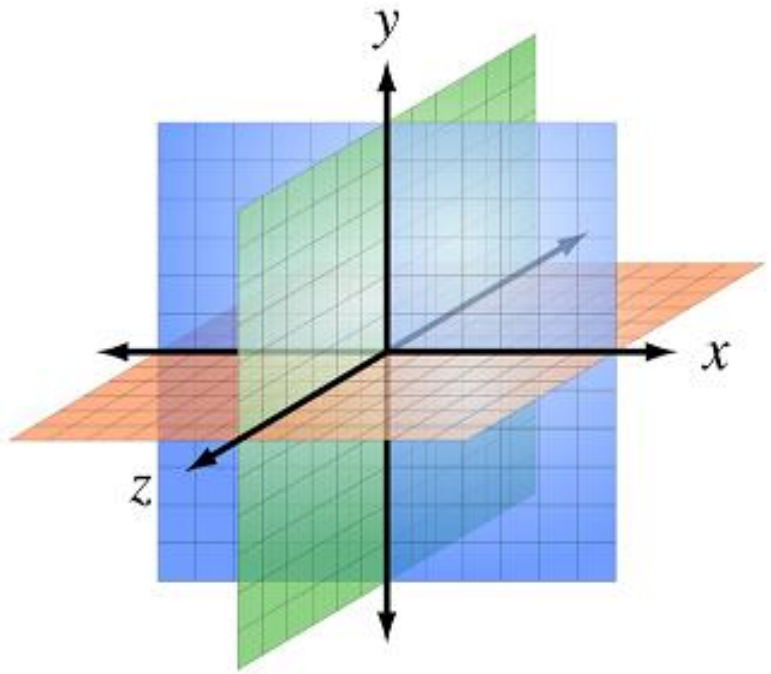
- NoSQL 数据存储
  - 不需要固定的表结构，通常也不存在连接操作。在大数据存取上具备关系型数据库无法比拟的性能优势。
  - 非关系型数据库以键值对存储，它的结构不固定，每一个元组可以有不一样的字段，每个元组可以根据需要增加一些自己的键值对，这样就不会局限于固定的结构，可以减少一些时间和空间的开销。

关系型数据库与NoSQL数据存储

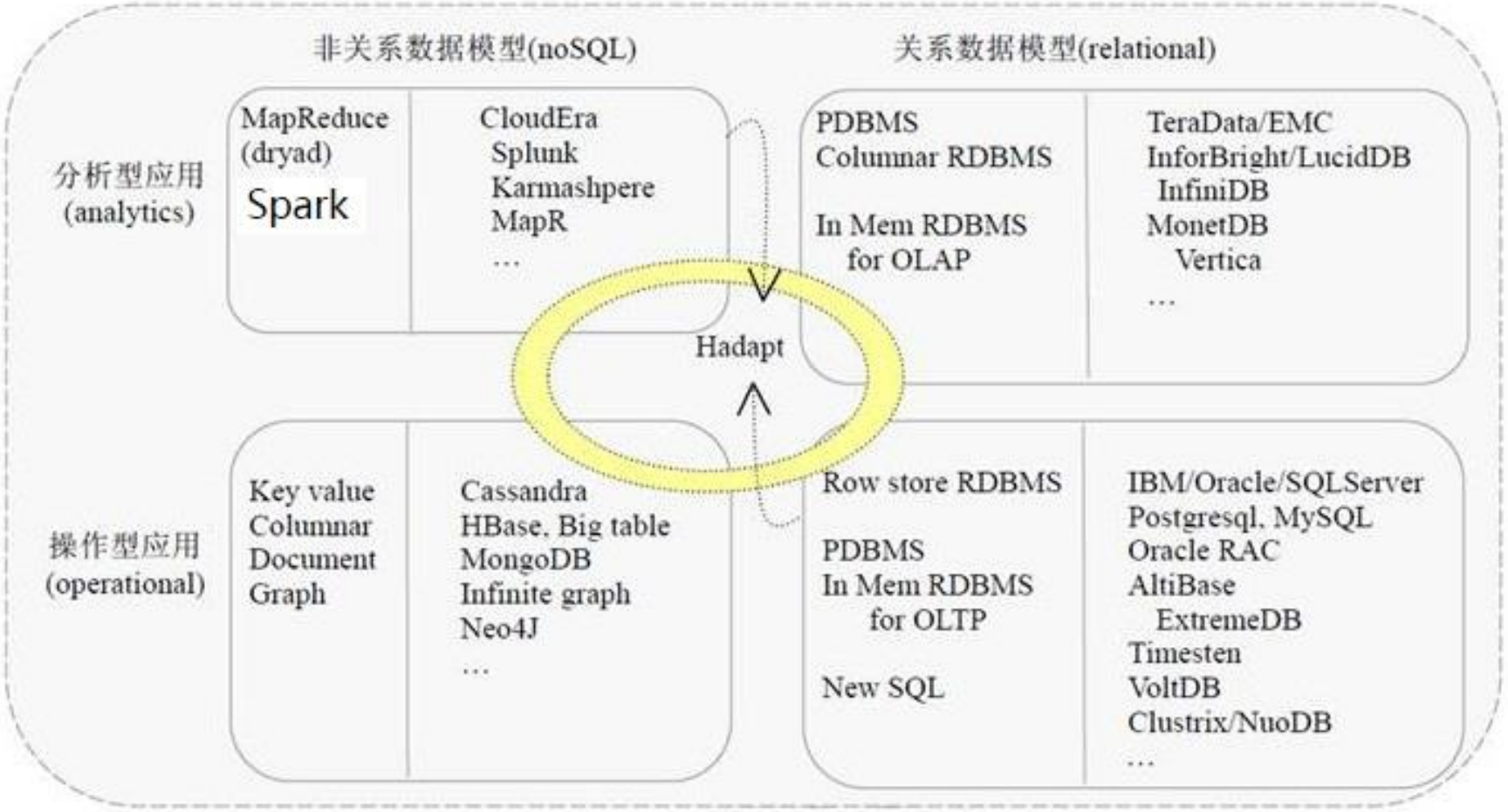


NoSQL / NewSQL …

- 根据应用的需求，可以从两个维度入手。
- 应用类型维度
  - 操作型应用
  - 分析型应用
- 数据模型维度
  - 关系模型
  - NoSQL 数据模型



大数据管理技术的四维象限



大数据管理技术的新格局

面向操作型应用的关系数据库技术

传统的基于行存储的关系数据库系统	• IBM的DB2、Oracle、微软的SQL Server等
面向实时计算的内存数据库系统	• Altibase，Timesten，Hana等
面向OLTP应用的new SQL	• VoltDB、Clustrix、NuoDB等

大数据管理技术的新格局

面向分析型应用的关系数据库技术

数据仓库	• 领头羊TeraData
列存储数据库系统	• MonetDB、InfiniDB、LucidDB，Vertica，SybaseIQ等
基于列存储技术的内存数据库	• MonetDB和Vertica

面向操作型应用的 NoSQL 技术

NoSQL 数据库系统相对于关系数据库系统具有的优势：

- 数据模型灵活、支持多样的数据类型
- 高度的扩展性。

NoSQL数据库系统

- 键值（key-value）存储数据库：Dynamo
- 列式存储数据库：BigTable、HBase
- 文档存储数据库：MongoDB

面向分析型应用的 NoSQL 技术

- MapReduce技术以其创新的设计理念、高度的扩展性和容错性，获得了学术界和工业界的青睐，围绕MapReduce的数据分析生态系统已经在几年前形成。
- 为了进一步提升计算性能和数据的实时分析能力，Hadoop与内存计算模式进行混合，目前已经成为实现高实时性的大数据查询和计算分析新的趋势。这种混合计算模式之集大成者当属UC Berkeley AMP Lab开发的Spark生态系统。

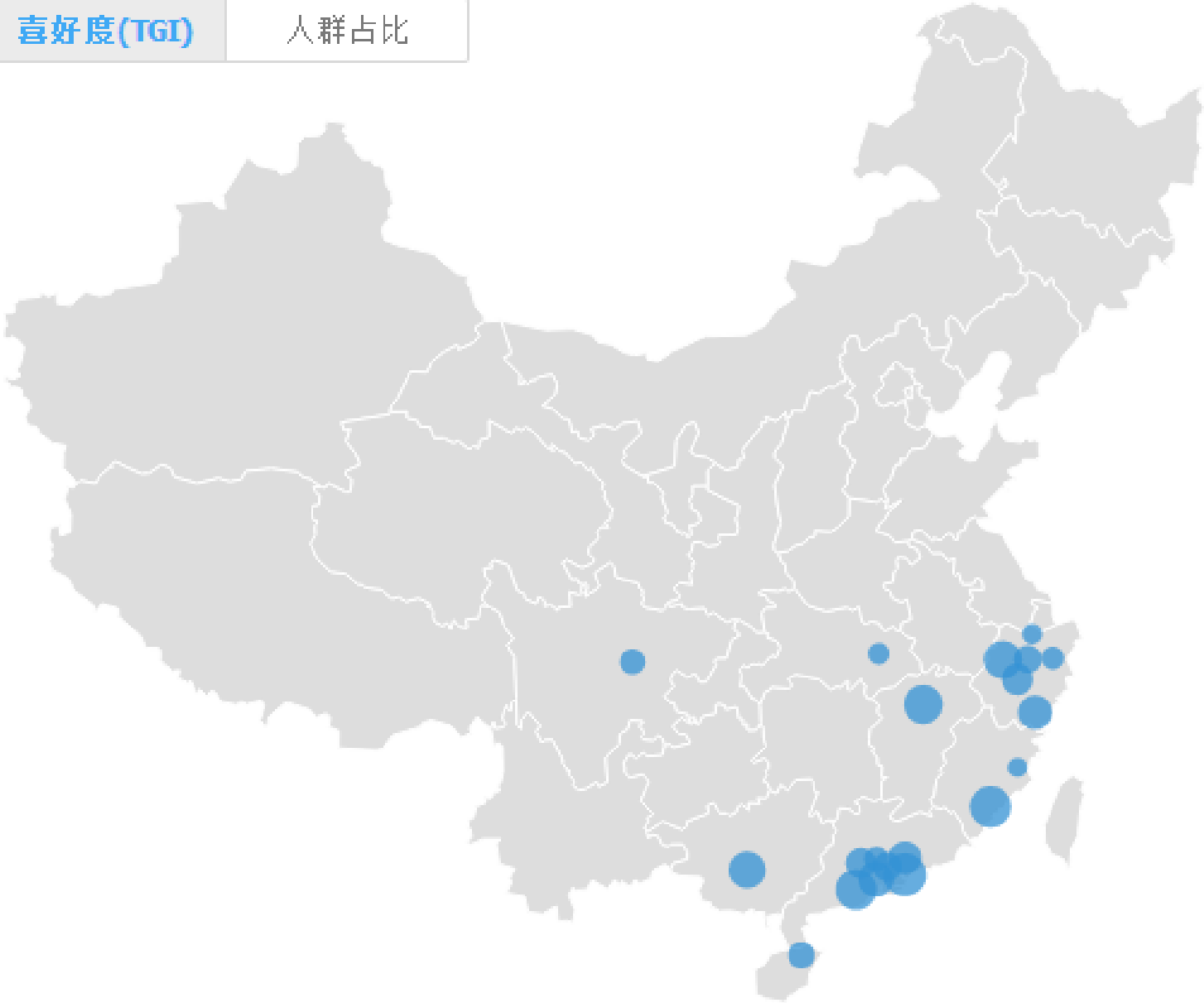

搜索指数
成交指数
搜索与成交指数
iphone5s： 搜索指数
2014.03.01 - 2014.08.31
245,935
217,850
189,766
161,681
133,596
105,511
03-14
04-07
05-01
05-25
06-18
07-12
08-05
趋势简报
"iph..."最近七天的搜索指数环比↓6.6%，与去年同期相比↑1362.0%。
"iph..."最近三十天的搜索指数环比↑5.5%，与去年同期相比↑2277.7%。
"iph..."未来一周内的总体趋势预测：保持平稳。
去阿里指数查看供货情况>
相关知识
搜索指数：
指数化的搜索量，反映搜索趋势，不等同于搜索次数。
成交指数：
由搜索带来的成交量，并进行指数化处理。反映成交趋势，不等同于成交量或成交金额。
数据来源：
淘宝网和天猫的总数据。
详细信息>


淘宝指数能告诉你...
长周期走势
淘宝上连衣裙的搜索趋势是怎样的？
任一关键词（如商品、行业、事件等）的搜索和成交走势。
人群特性
淘宝上搜索、购买iPhone4的都是什么样的人？
用淘宝指数查看不同商品的消费人群特征。
成交排行
最近7天淘宝最火的搜索词、行业和品牌是？
基于淘宝搜索和成交的排行榜，宏观数据清晰呈现。
TOP15热销类目
1 女装
2 手机
3 美容护肤
4 数码配件
5 男装
市场细分
北京女白领和20岁大学生都买过什么面膜？
淘宝指数告诉你不同标签的人买过什么商品。
20 - 25岁
女
白领

数据仓库的应用

地域细分 从2014-06-01到2014-08-31，91天来搜索 iphone5s 的消费者

喜好度(TGI) | 人群占比



喜好度（TGI）排行

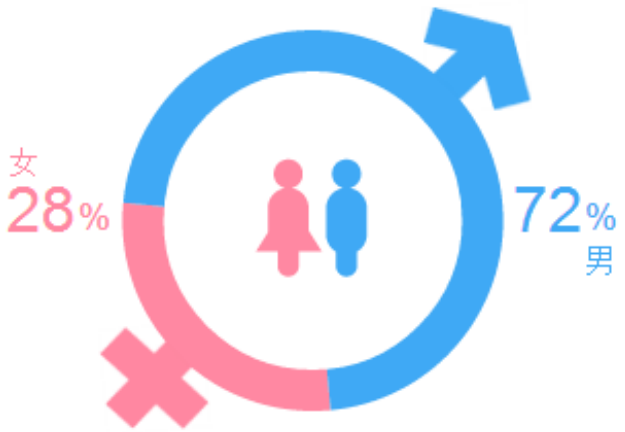
iphone5s

省份	1 江西	11 北京
	2 广西	12 江苏
	3 广东	13 山东
	4 浙江	14 陕西
	5 福建	15 天津
	6 四川	16 安徽
	7 湖北	17 甘肃
	8 湖南	18 河南
	9 上海	19 黑龙江
	10 重庆	20 辽宁
城市	1 深圳市	11 佛山市
	2 厦门市	12 东莞市
	3 江门市	13 绍兴市
	4 南昌市	14 海口市
	5 杭州市	15 成都市
	6 南宁市	16 广州市
	7 中山市	17 宁波市
	8 温州市	18 武汉市
	9 惠州市	19 嘉兴市
	10 金华市	20 福州市

iphon... 城市数据起止于：2014-06-01/2014-08-31

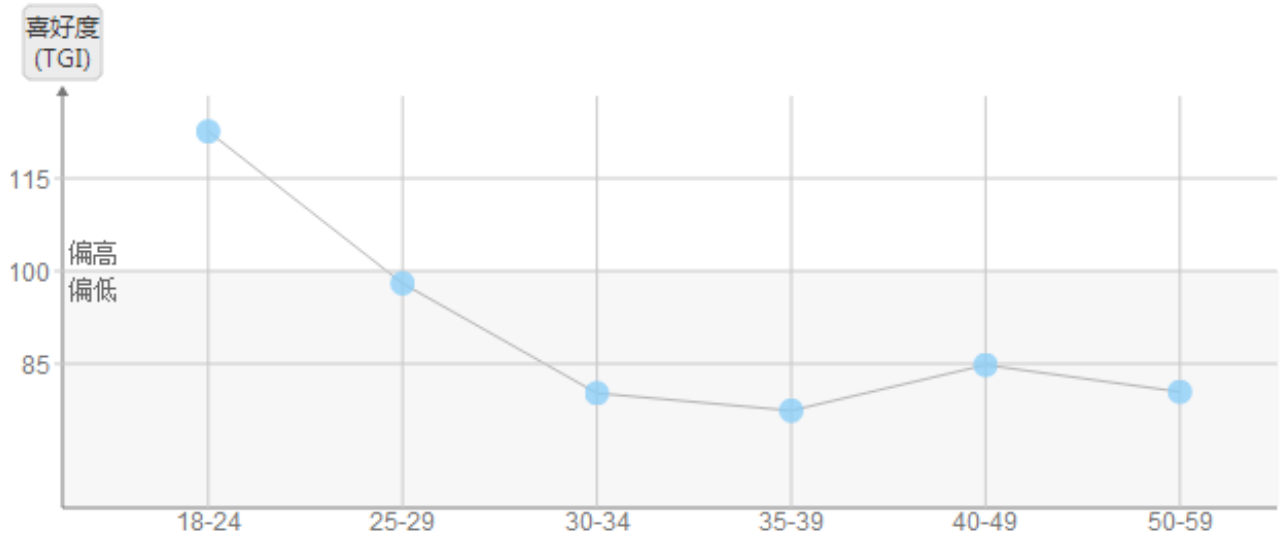
人群定位 从2014-06-01到2014-08-31，91天来搜索 iphone5s 的消费者

性别比例

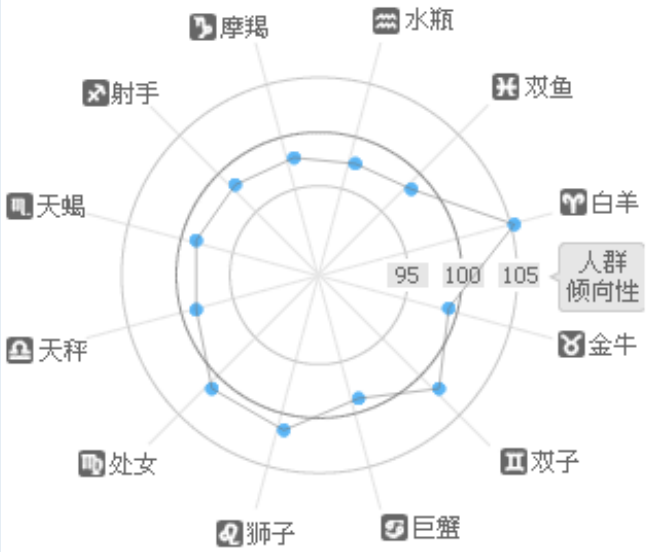


分享到

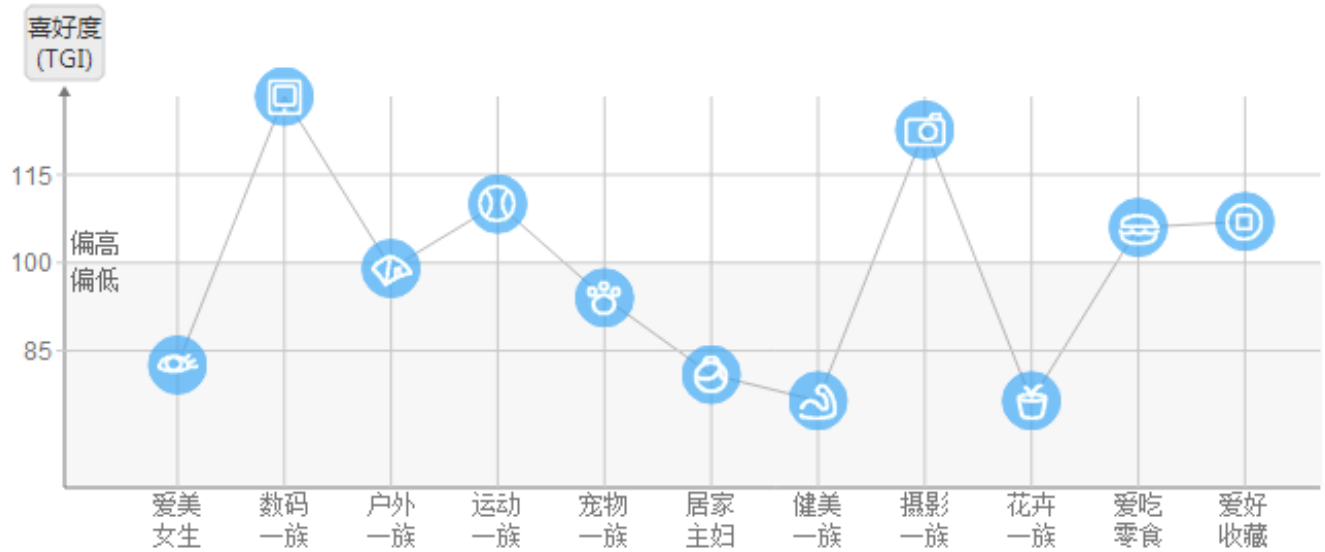
年龄 显示人群占比



星座 显示人群占比



爱好 显示人群占比



第7章
数据库系统

1 数据库的起源与发展

2 关系数据库

3 数据仓库与OLAP

4 **SQL语言**

- 1974年，IBM的Boyce和Chamberlin将Codd关系数据库的12条准则的数学定义以简单的关键字语法表现出来，里程碑式地提出了SQL（Structured Query Language）语言。
- 1976年IBM的Codd发表了一篇里程碑的论文“R系统：数据库关系理论”，介绍了关系数据库理论和查询语言SQL。
- 随后，Oracle的创始人Larry Ellison仔细地阅读了这篇文章，几个月后，Ellison他们就开发了Oracle 1.0。

结构性查询语言（SQL）

适用于数据分析的自然语言	高效的语言	优化的处理	持续创新
• 基于关系代数、面向集合	• 声明式语言：直接表述想要的结果，而非获得方式	• 将结果与方式脱钩有助于持续优化SQL引擎	• 内部处理、语言结构和数据访问一直在增强

SQL的强大功能

“……对于我们业务逻辑的任一部分来说，处理非 ACID 数据存储都非常复杂，如果不用 SQL 查询，我们的业务根本没法开展。”

Google，VLDB 2013

Google

“[Facebook] 开始是用 Hadoop。现在我们在引入关系型数据库系统来增强 Hadoop…… [我们] 意识到使用错误的技术来解决某些问题可能比较困难。”

Ken Rudin，Facebook，TDWI 2013

facebook

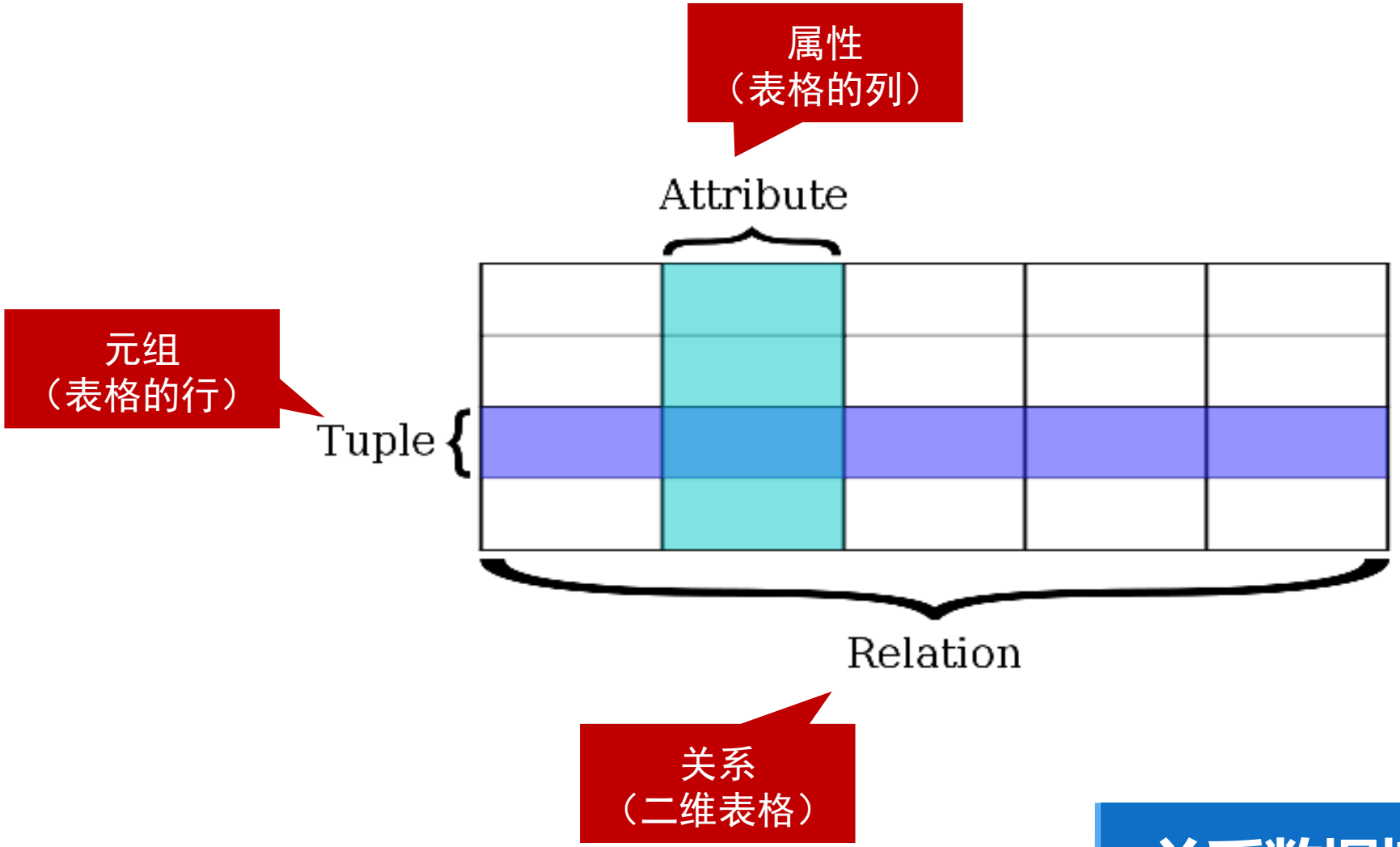
SQL的强大功能


ORACLE
DATABASE
Microsoft
SQL Server
IBM
DB2
MySQL
PostgreSQL
HIVE
Stinger
presto
cloudera
IMPALA
SHARK
TAJO

SQL广受欢迎

- A *data model* is a collection of concepts for describing data.
- A schema is a description of a particular collection of data, using a given data model.
- The relational model of data is the most widely used model today.
  - Main concept: *relation*, basically a table with rows and columns.
  - Every relation has a *schema*, which describes the columns, or fields.

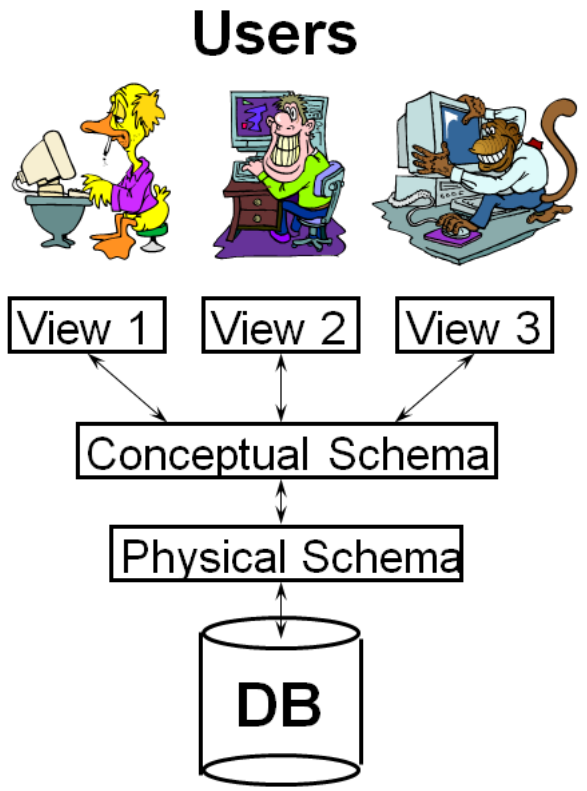
描述数据的方法：数据模型



关系数据模型

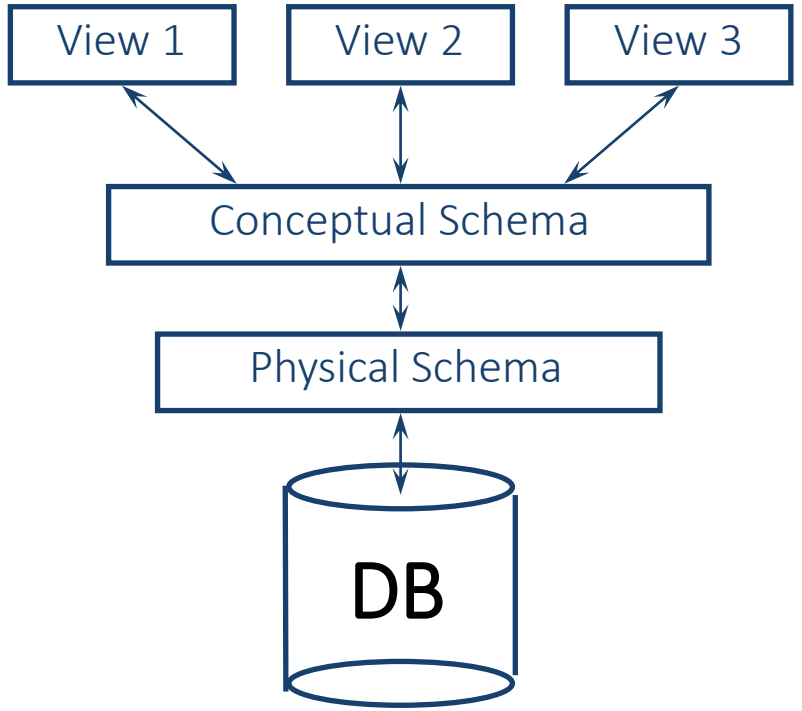
- Views describe how users see the data.
- Conceptual schema defines logical structure
- Physical schema describes the files and indexes used.

  (sometimes called the ANSI/SPARC model)



抽象级别

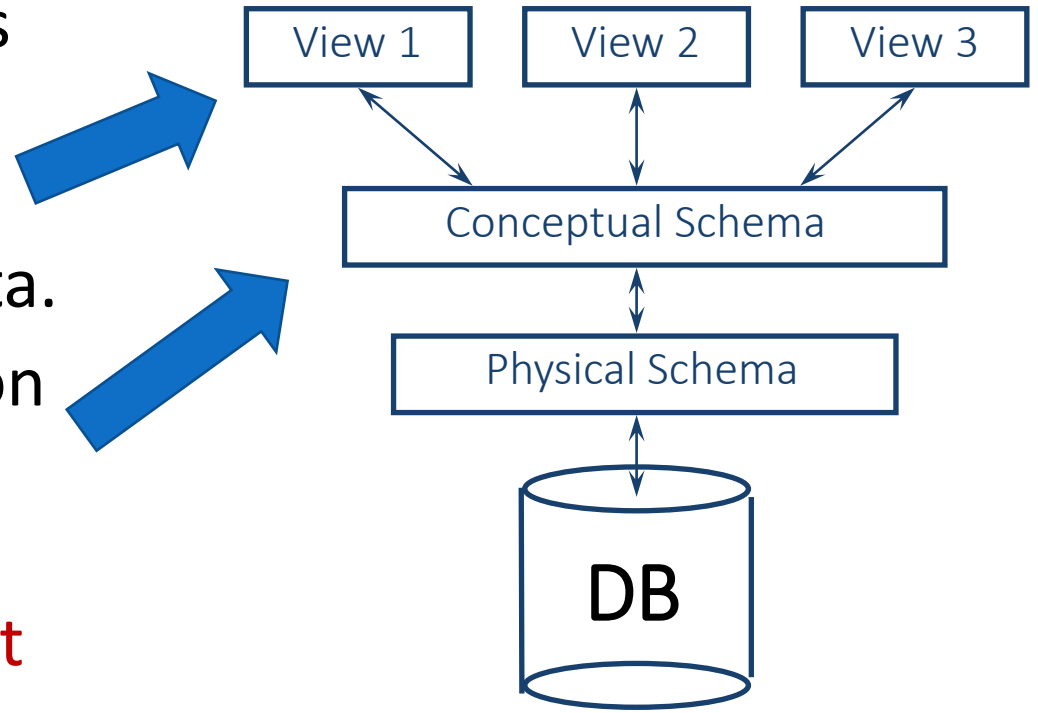
- Conceptual schema:
  - *Students*(sid: *string*, name: *string*, login: *string*, age: *integer*, gpa:*real*)
  - *Courses*(cid: *string*, cname:*string*, credits: *integer*)
  - *Enrolled*(sid:*string*, cid:*string*, grade:*string*)
- External Schema (View):
  - *Course_info*(cid:*string*,enrollment:*integer*)
- Physical schema:
  - Relations stored as unordered files.
  - Index on first column of Students.



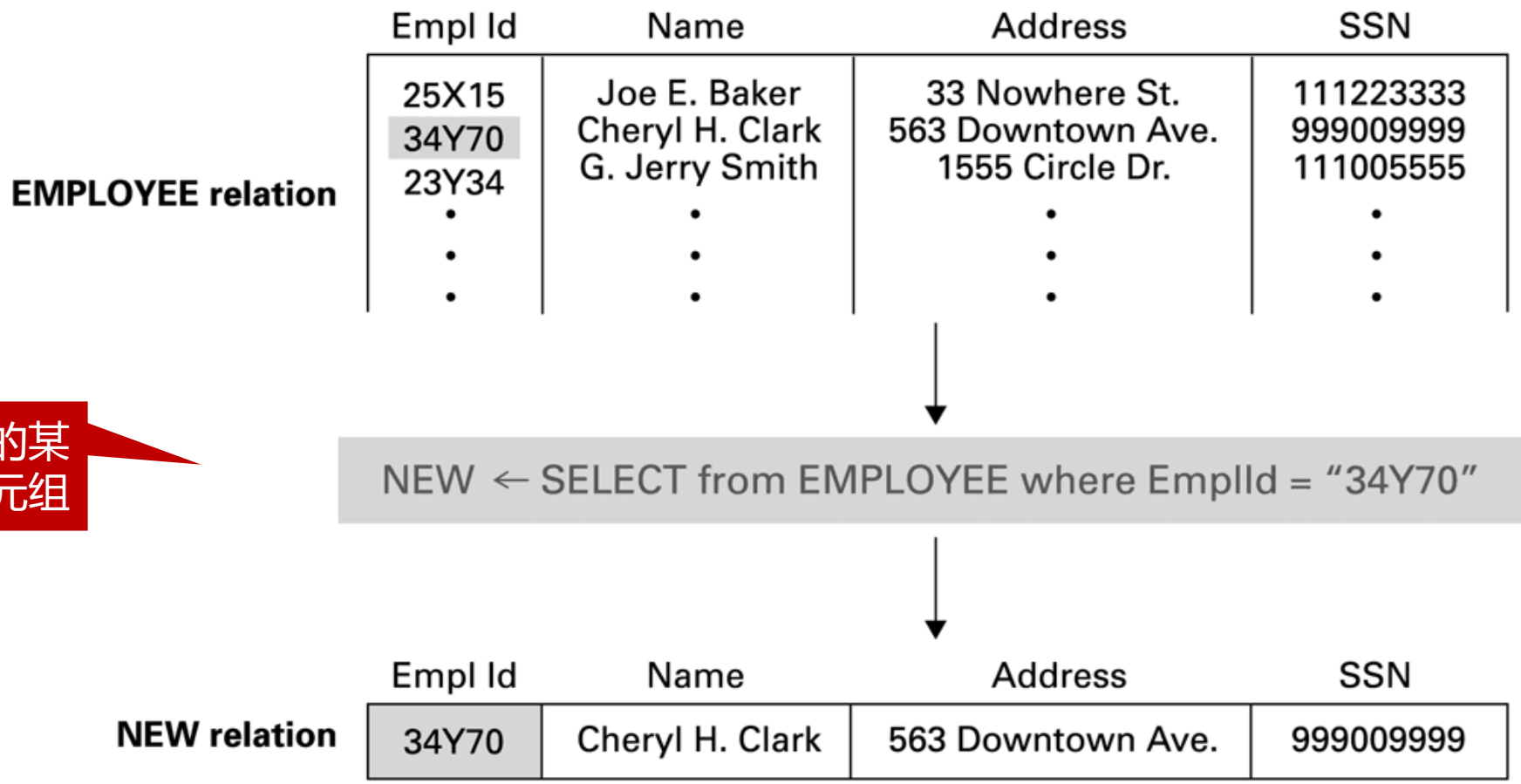
Example: University Database

抽象级别

- Applications insulated from how data is structured and stored.
- Logical data independence: Protection from changes in *logical* structure of data.
- Physical data independence: Protection from changes in *physical* structure of data.
- Q: Why are these particularly important for DBMS?

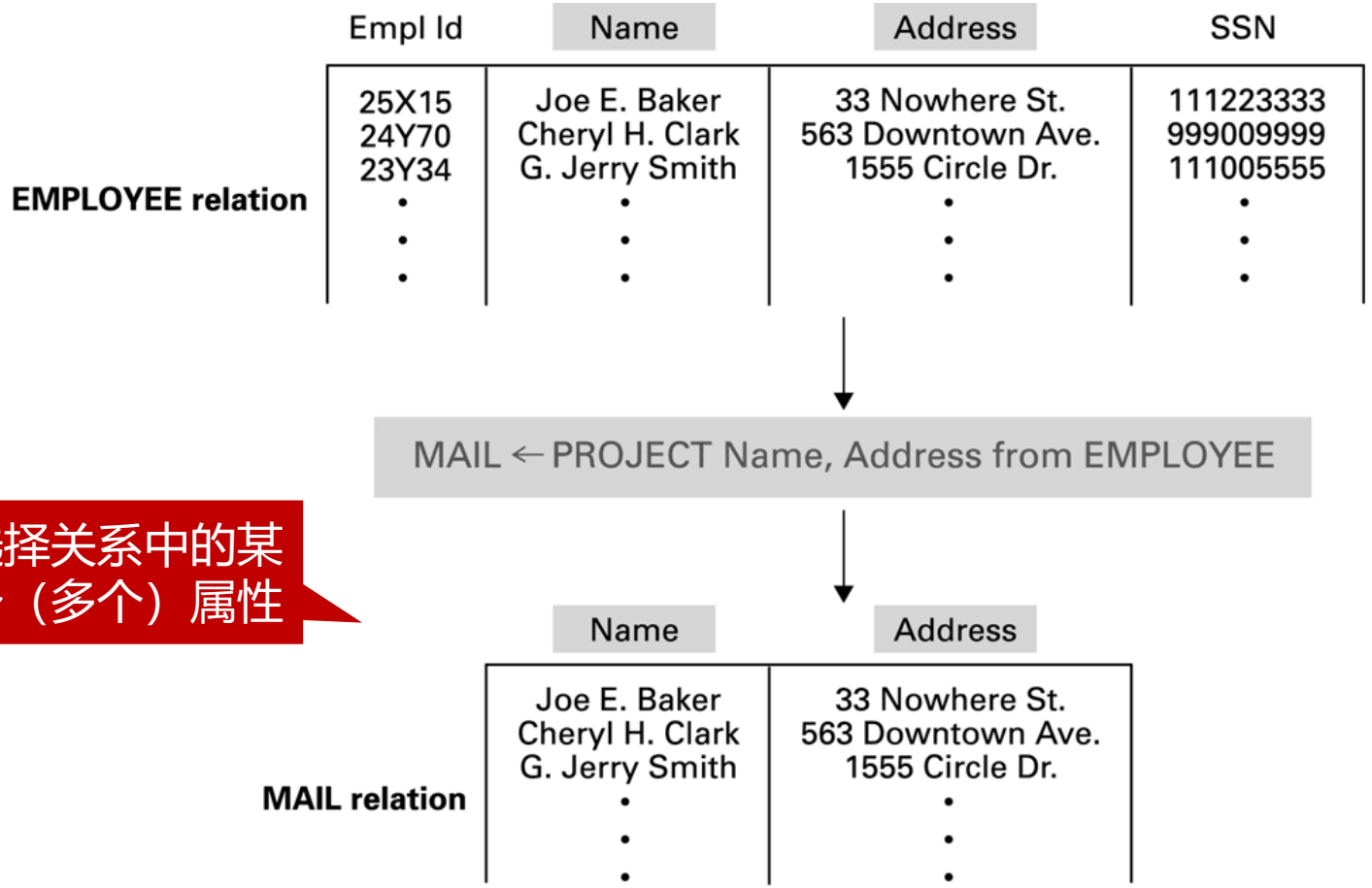


数据独立性

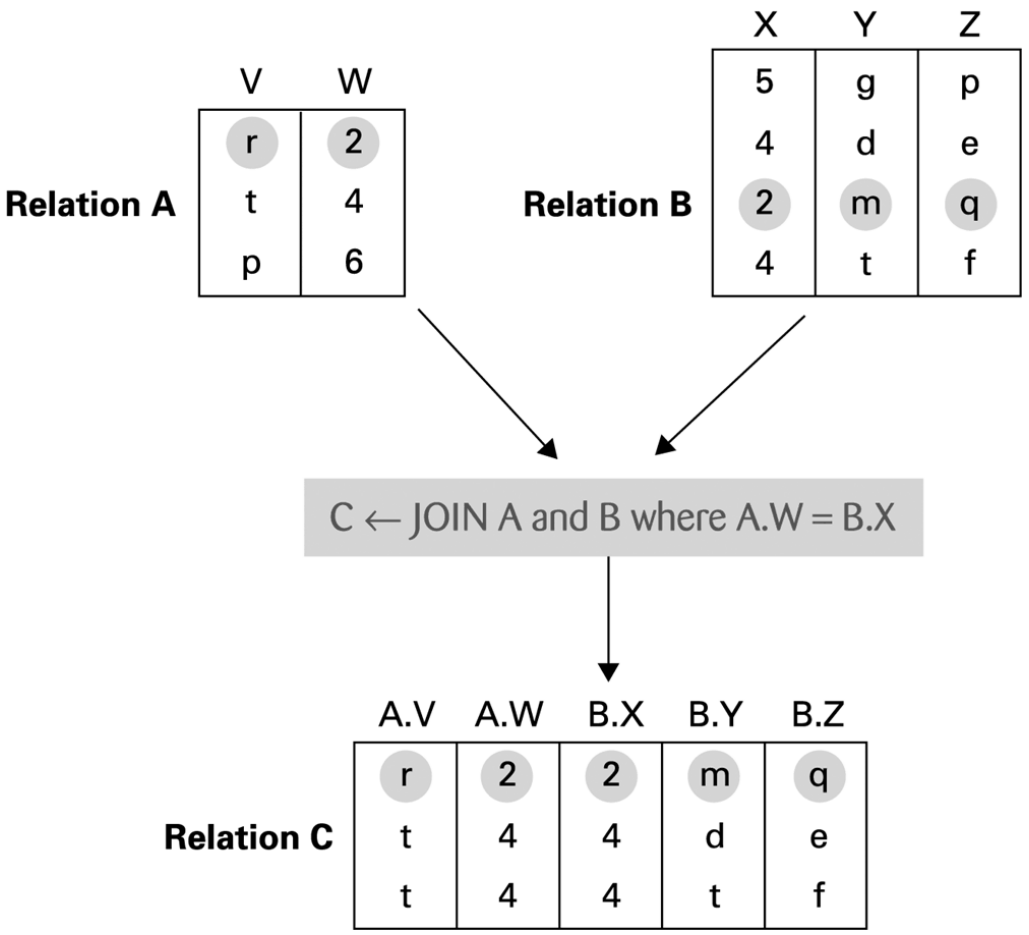


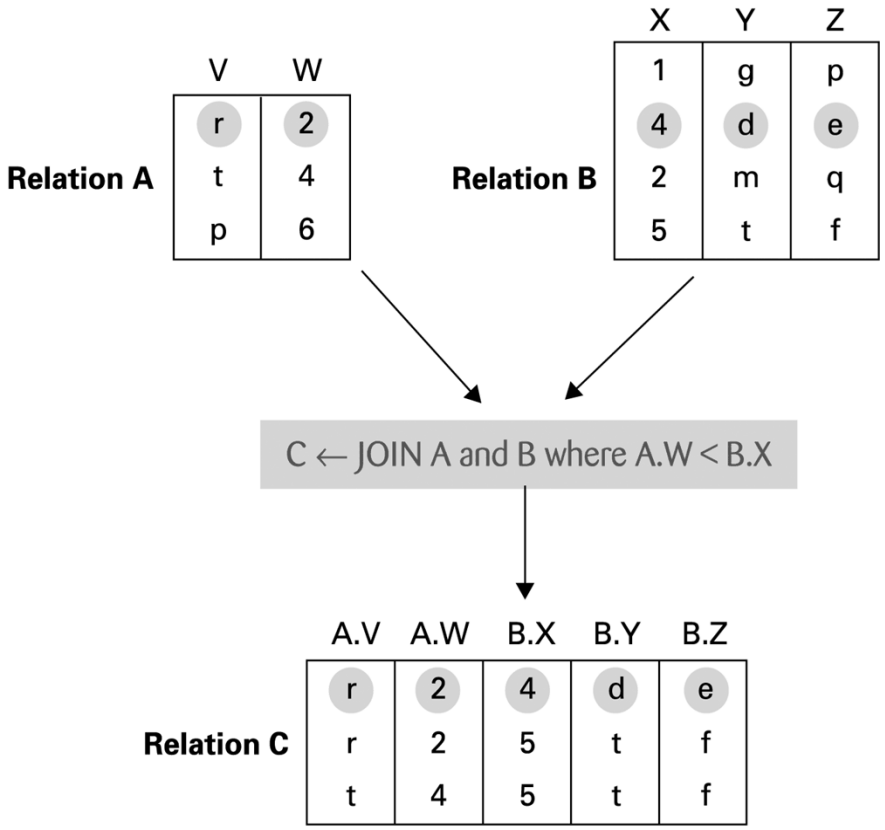
选择关系中的某个（多个）元组

关系代数：选择

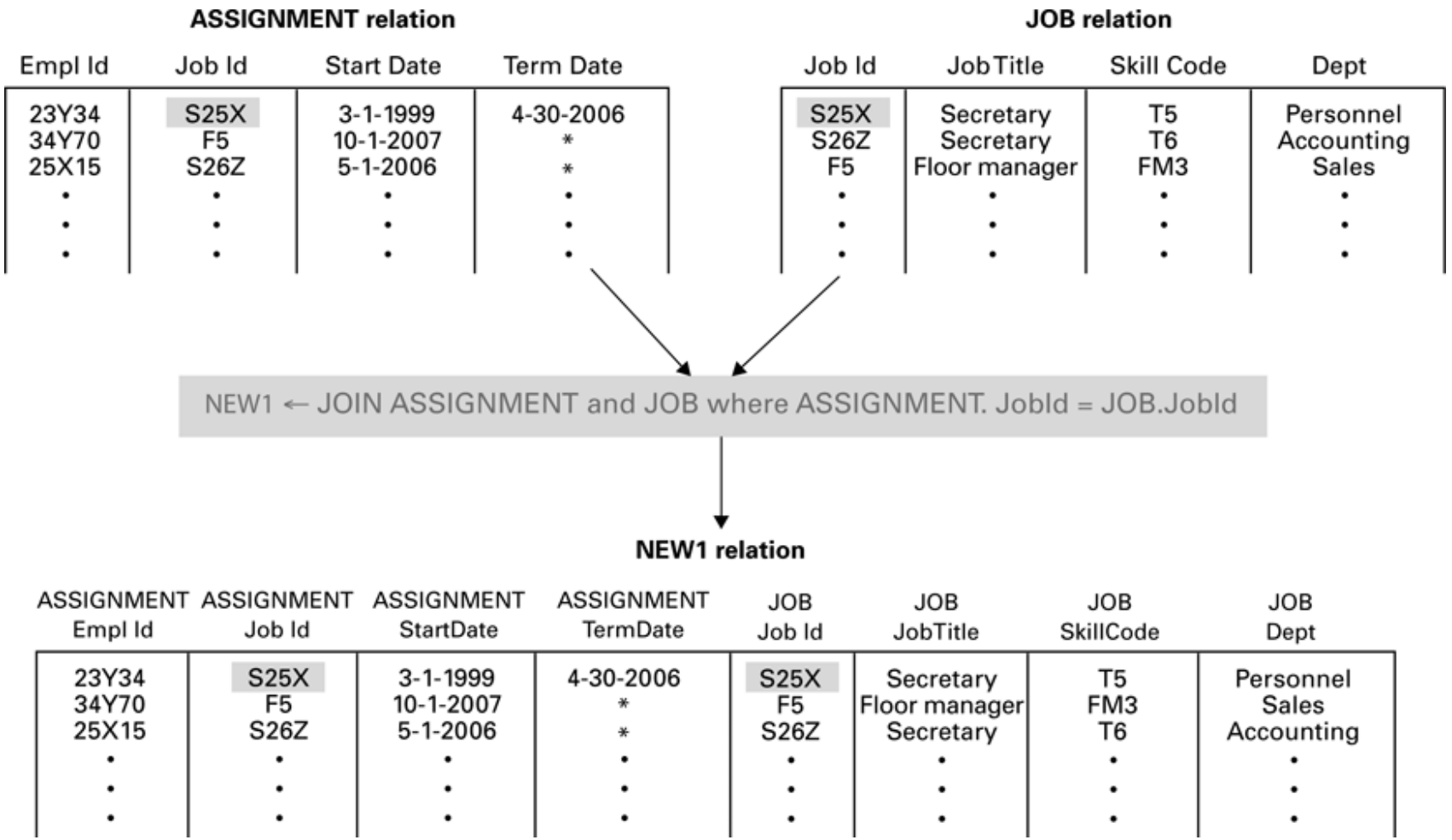

Empl Id
Name
Address
SSN
EMPLOYEE relation
25X15
24Y70
23Y34
Joe E. Baker
Cheryl H. Clark
G. Jerry Smith
33 Nowhere St.
563 Downtown Ave.
1555 Circle Dr.
111223333
999009999
111005555
MAIL ← PROJECT Name, Address from EMPLOYEE
选择关系中的某个（多个）属性
Name
Address
MAIL relation
Joe E. Baker
Cheryl H. Clark
G. Jerry Smith
33 Nowhere St.
563 Downtown Ave.
1555 Circle Dr.

关系代数：投影


Relation A
V W
r 2
t 4
p 6
Relation B
X Y Z
5 g p
4 d e
2 m q
4 t f
C ← JOIN A and B where A.W = B.X
Relation C
A.V A.W B.X B.Y B.Z
r 2 2 m q
t 4 4 d e
t 4 4 t f


Relation A
V W
r 2
t 4
p 6
Relation B
X Y Z
1 g p
4 d e
2 m q
5 t f
C ← JOIN A and B where A.W < B.X
Relation C
A.V A.W B.X B.Y B.Z
r 2 4 d e
r 2 5 t f
t 4 5 t f

关系代数：连接


ASSIGNMENT relation
Empl Id
Job Id
Start Date
Term Date
23Y34
34Y70
25X15
S25X
F5
S26Z
3-1-1999
10-1-2007
5-1-2006
4-30-2006
*
*
JOB relation
Job Id
JobTitle
Skill Code
Dept
S25X
S26Z
F5
Secretary
Secretary
Floor manager
T5
T6
FM3
Personnel
Accounting
Sales
NEW1 ← JOIN ASSIGNMENT and JOB where ASSIGNMENT. JobId = JOB.JobId
NEW1 relation
ASSIGNMENT Empl Id
ASSIGNMENT Job Id
ASSIGNMENT StartDate
ASSIGNMENT TermDate
JOB Job Id
JOB JobTitle
JOB SkillCode
JOB Dept
23Y34
34Y70
25X15
S25X
F5
S26Z
3-1-1999
10-1-2007
5-1-2006
4-30-2006
*
*
S25X
F5
S26Z
Secretary
Floor manager
Secretary
T5
FM3
T6
Personnel
Sales
Accounting

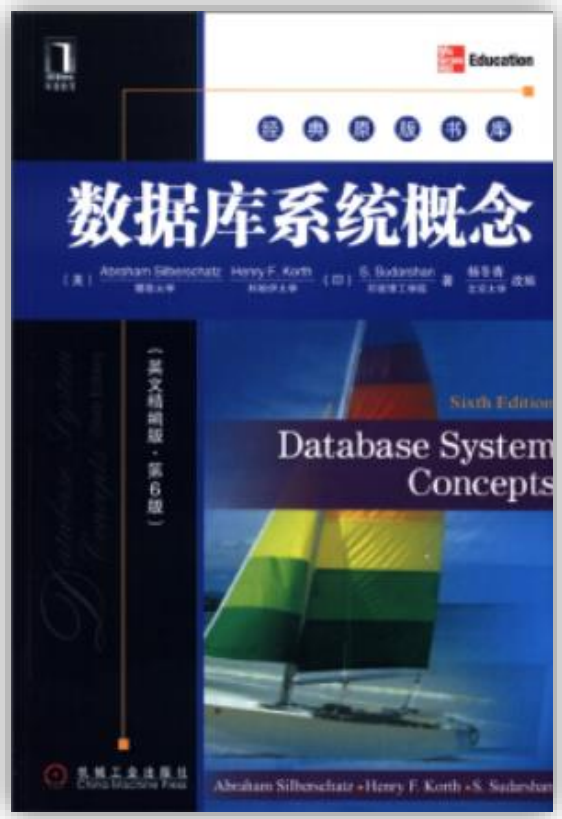
关系代数：连接

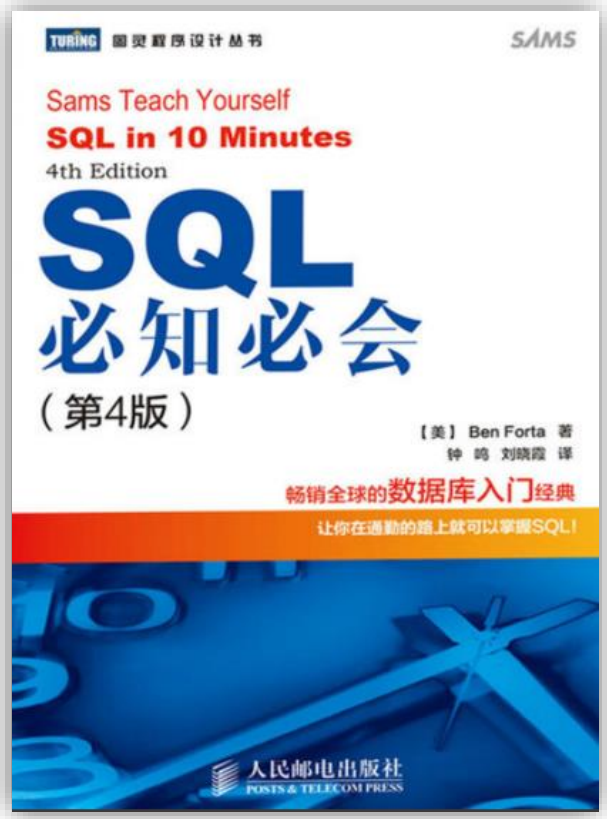
- MySQL是一个关系型数据库管理系统，由瑞典MySQL AB 公司开发，目前属于 Oracle 旗下产品。
- MySQL 是最流行的关系型数据库管理系统之一，在 WEB 应用方面，MySQL是最好的 RDBMS 应用软件。
- MySQL所使用的 SQL 语言是用于访问数据库的最常用标准化语言。
- MySQL 由于其体积小、速度快、总体拥有成本低，尤其是开放源码这一特点，一般中小型企业的开发都选择MySQL 作为数据库。

MySQL简介

- SQL这种传统数据库时代的语言在大数据时代还会有用吗？为什么？

思考题


Education
经典原版书库
数据库系统概念
(英文精简版·第6版)
Sixth Edition
Database System Concepts
Abraham Silberschatz · Henry F. Korth · S. Sudarshan


TURING
图灵程序设计丛书
SAMS
Sams Teach Yourself
SQL in 10 Minutes
4th Edition
SQL
必知必会
(第4版)
【美】Ben Forta 著
钟鸣 刘晓霞 译
畅销全球的数据库入门经典
让你在通勤的路上就可以掌握SQL!
人民邮电出版社
POSTS & TELECOM PRESS

书籍推荐

第7章
数据库系统

1 数据库的起源与发展

2 关系数据库

3 数据仓库与OLAP

4 SQL语言