



管理信息系统

Management Information Systems

王谦 博士/副教授

南开大学商学院管理科学与工程系

wangqian70@nankai.edu.cn



Chapter 6



数据库与信息管埋





学习目标

- * 描述在传统的文件环境中，管理数据资源的问题是如何通过数据库管理系统得以解决的
- * 描述数据库管理系统的功能以及价值
- * 掌握部分重要的数据库设计原则
- * 评估通过数据库访问来获取信息以提升商业表现的工具以及技术的价值
- * 信息政策、数据治理以及保证数据质量对公司的数据资源管理至关重要





开篇案例

* 秘鲁信贷银行为更好的数据管理所做的努力

- 问题: 冗余的过时系统, 以及重复的、非一致的数据
- 解决方案: 用单一知识库的商务智能系统取代遗留下来的无价值系统
- SAP 组合软件套装包括企业资源管理模块以及支持在企业范围内进行追踪、报告以及分析的数据仓库
- 论述在成功的数据管理中, 信息技术的作用
- 数码技术在提升企业表现时降低成本的能力

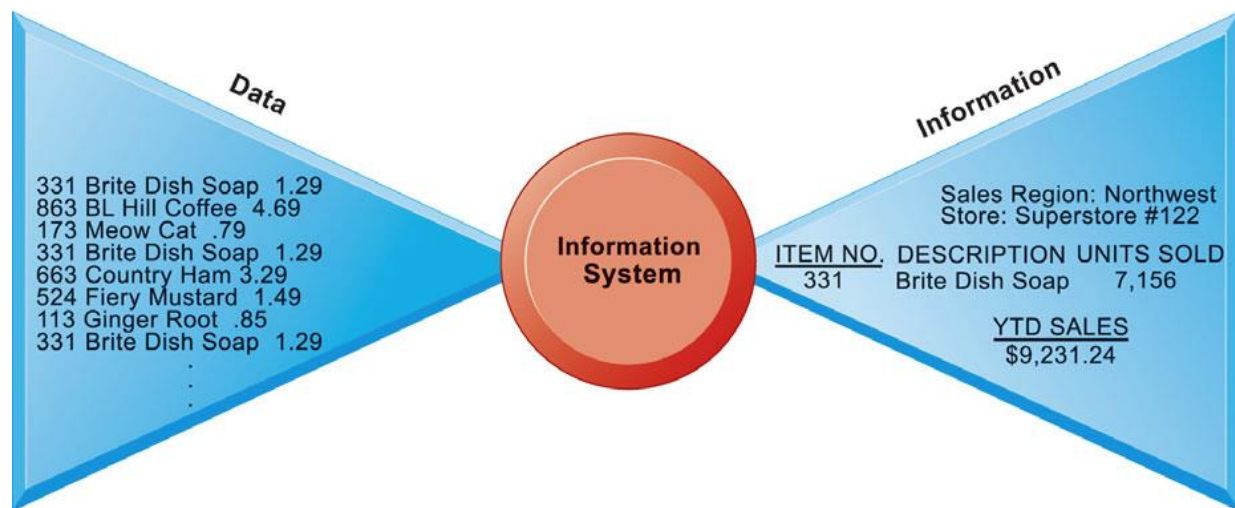




6.1 数据库概论

* 信息与数据

- 信息就是事物之间相互联系、相互作用的状态的描述，它可以帮助我们消除对事物的不确定性
- 在信息处理中，信息载体上反映信息内容且可被接收者（人或机器）识别的物理符号称为数据

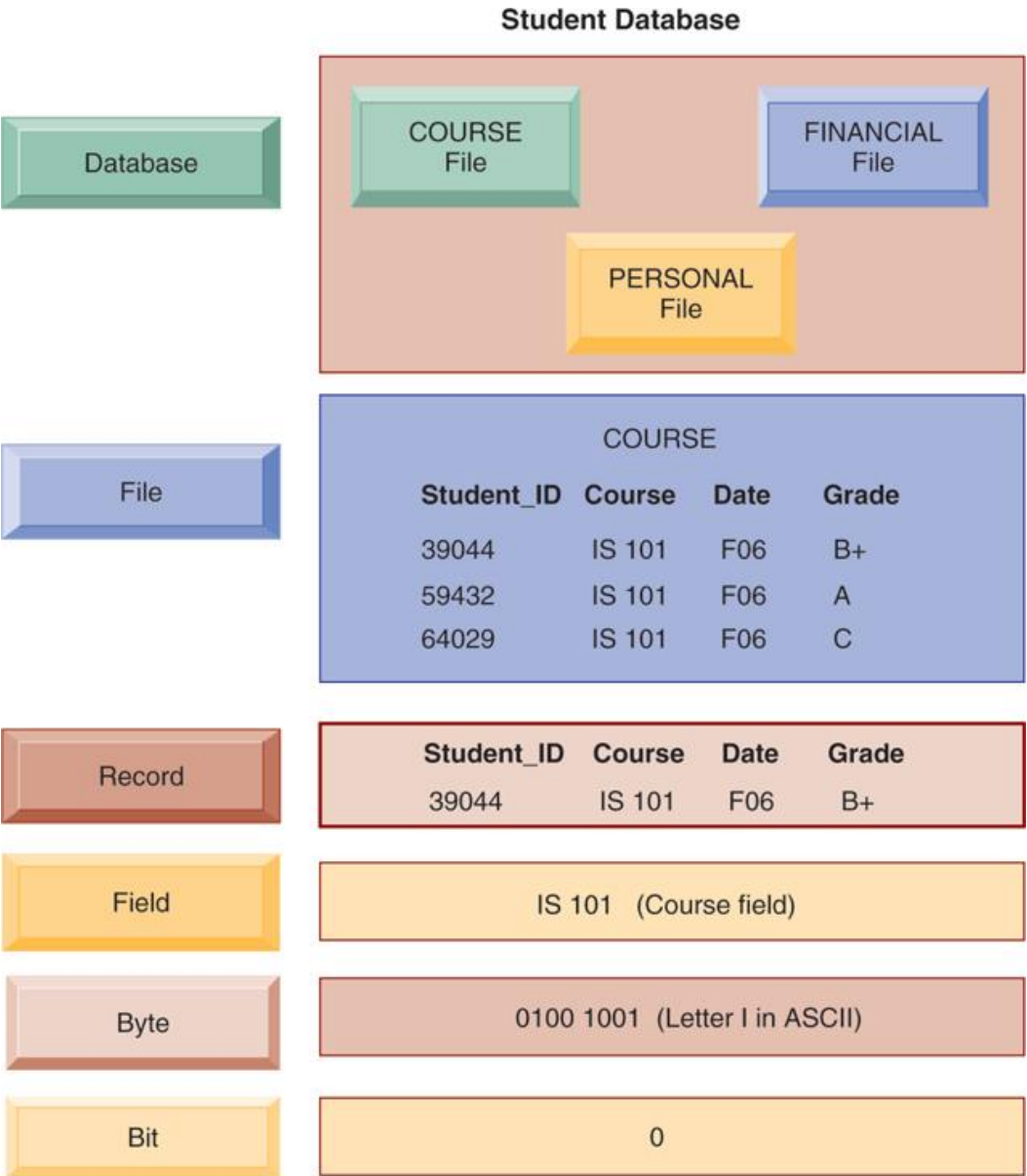




6.1 数据库概论

* 数据分层

- 计算机以层级的方式来组织数据，从比特(位) 开始第一层，以0或1表示。比特可以聚集以形成字节来体现一种特征。而字节可以聚集以形成字段，而相关的字段又可以聚集成记录。相关的记录可以汇集成文档，相关的文档汇集在一起形成数据库





6.1 数据库概论

* 传统文件环境中的数据管理问题

- 数据冗余:
 - 在多个文件中的重复数据的产生
- 数据不一致性:
 - 相同属性具有不同的值
- 程序数据的相依性:
 - 程序的变化需要程序所处理数据的变化
- 不够灵活
- 安全性不足
- 无法进行数据共享，可用性低

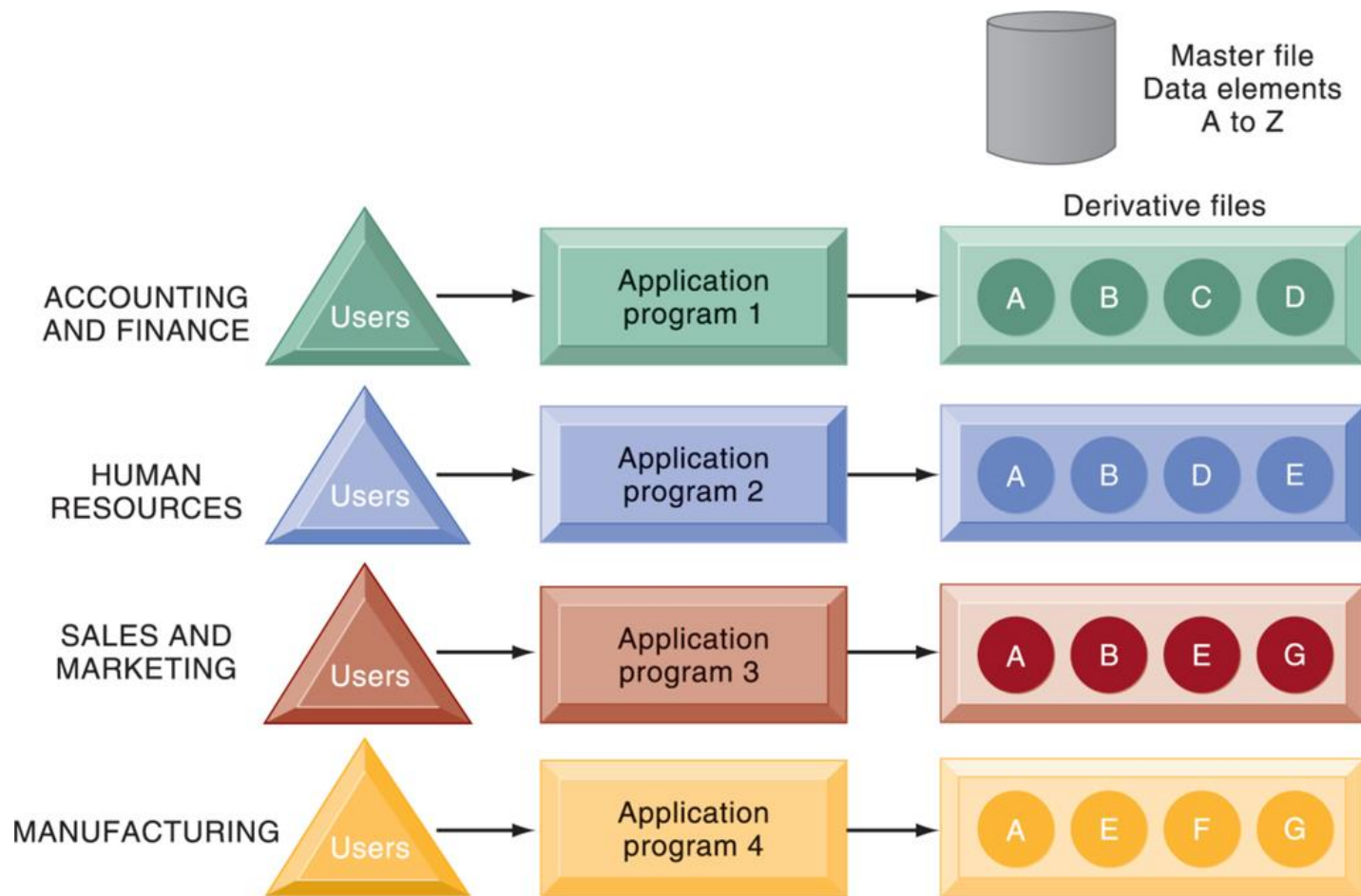




6.1 数据库概论

* 传统的文件处理方式

- 传统处理文件的方式激励企业的每个职能部门发展专门的应用。每个应用都需要单独的数据字段，这些字段可能是主文件的一个子集。这些子集导致了数据发生冗余以及不一致，这会降低数据处理的灵活性并且浪费了存储资源。

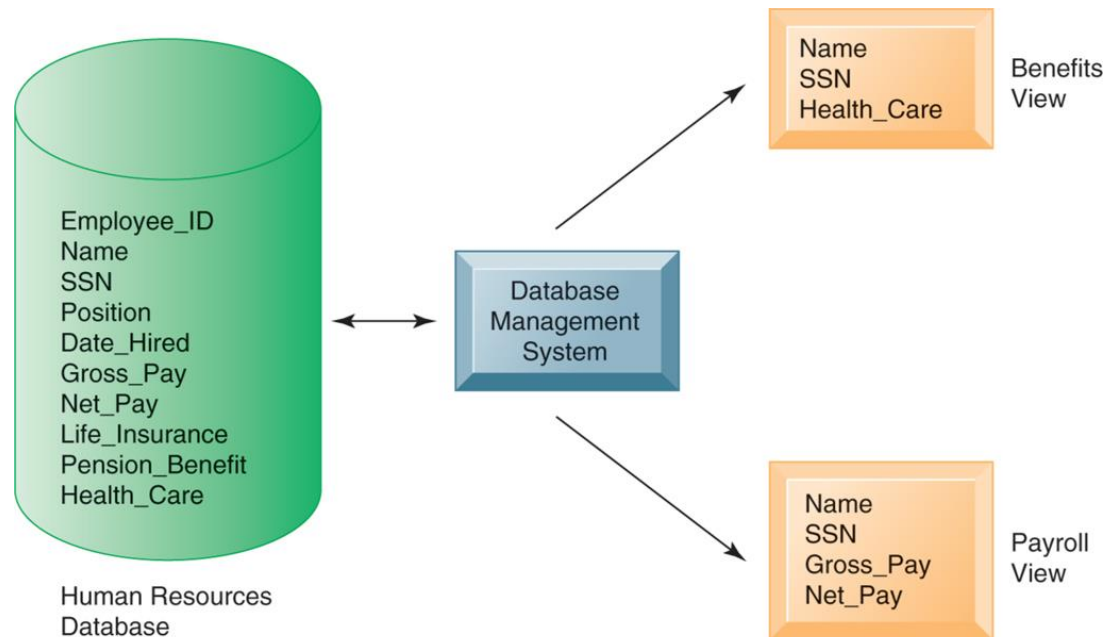




6.1 数据库概论

* 采用数据库管理数据的好处

- 应用与物理层数据文件之间的接口
- 做到了数据在逻辑结构与物理结构之间的隔离
- 解决了传统文件环境中可能发生的问题
 - 控制冗余
 - 消除不一致性
 - 区分开程序和数据
 - 使得企业能够中心化的管理数据并且保证数据安全

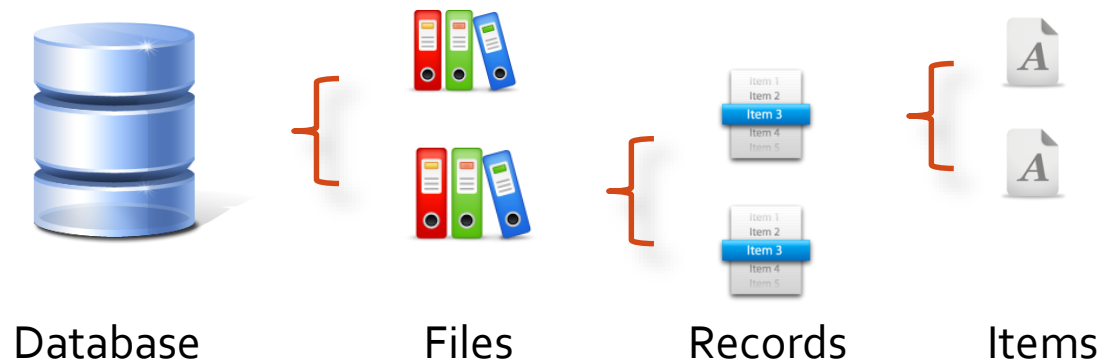




6.1 数据库概论

* 数据组织的层次

- **数据项** (Item) : 具有确定逻辑意义 (即可描述信息内容) 的数据的最小单位
- **记录** (Record) : 具有一定关系的数据项的一个有序集合
- **文件** (File) : 同类记录的有序集合
- **数据库** (Database) : 数据库是存贮起来的相关数据的集合

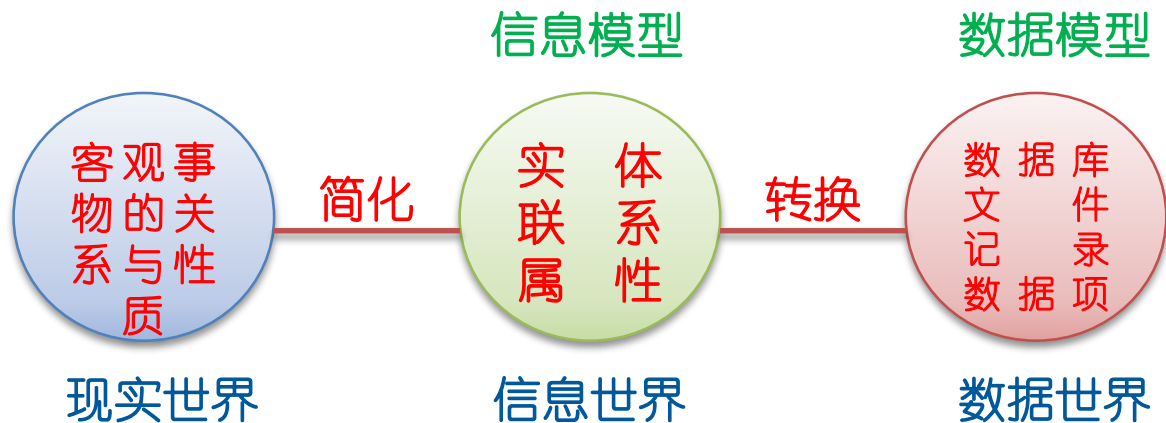




6.1 数据库概论

* 数据库概述

- 对客观世界的认识、描述是一个逐步的过程，可以有层次之分，它们分成三个层次
 - **现实世界**。它是客观存在的事物及其相互联系，客观存在的事物分为“对象”和“性质”两个方面，同时事物之间有广泛的联系
 - **信息世界**。它是客观存在的现实世界在人们头脑中的反映
 - **数据世界**。对信息世界中的有关信息经过加工、编码、格式化等具体处理，便进入了数据世界





6.1 数据库概论

* 信息模型

- 信息模型的要素

➤ 实体 (Entity)

- 定义：任何客观存在的事物，人或物，实际的东西或概念的东西
- 分为两个层次：个体和实体集（总体）

➤ 属性 (Attribute)

- 定义：实体某一方面的性质或特性
- 由属性名和属性值来描述
- 可分为原子属性和可分属性

➤ 联系 (Relationship)

- 定义：客观存在的事物之间相互关系，通常指是实体集与实体集之间的关系

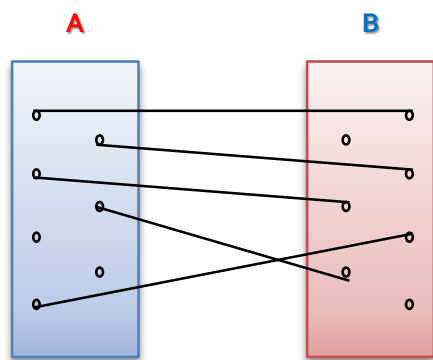


6.1 数据库概论

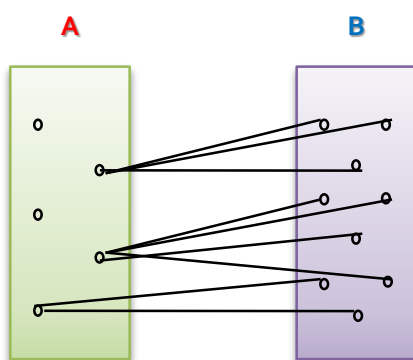
* 信息模型

- 实体联系的基本类别

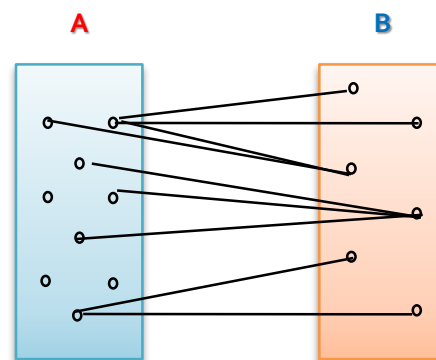
- 一对一联系，记为1: 1
- 一对多联系，记为1: N
- 多对多联系，记为N: M



一对一联系



一对多联系



多对多联系



6.1 数据库概论

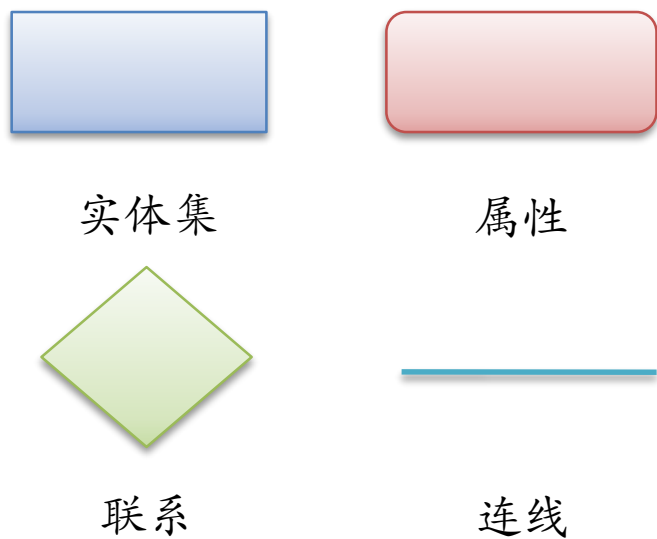
* 信息模型

- 信息模型的描述方法：实体—联系（E-R）方法

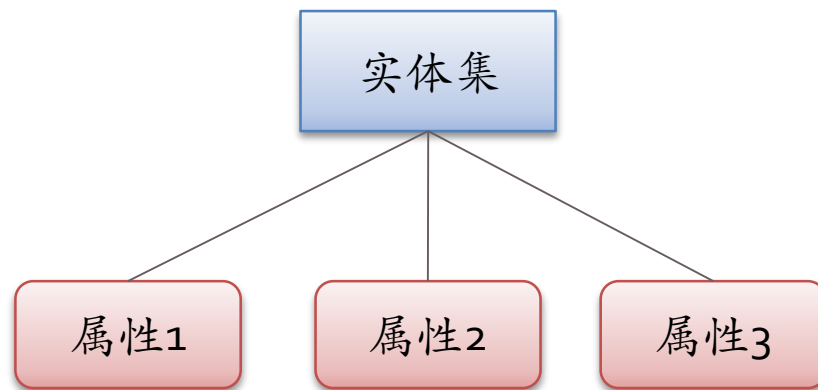
- 陈品山Peter P. S. Chen于1976年提出了建立信息模型的一种方法：实体-关系方法（Entity-Relationship Approach），简称E-R方法，应用此法建立的信息模型称之为E-R模型
- 通常给某特定的实体集命名，其名称有可能选用某个或某几个能代表此实体集的属性名，也可以另取名称



6.1 数据库概论



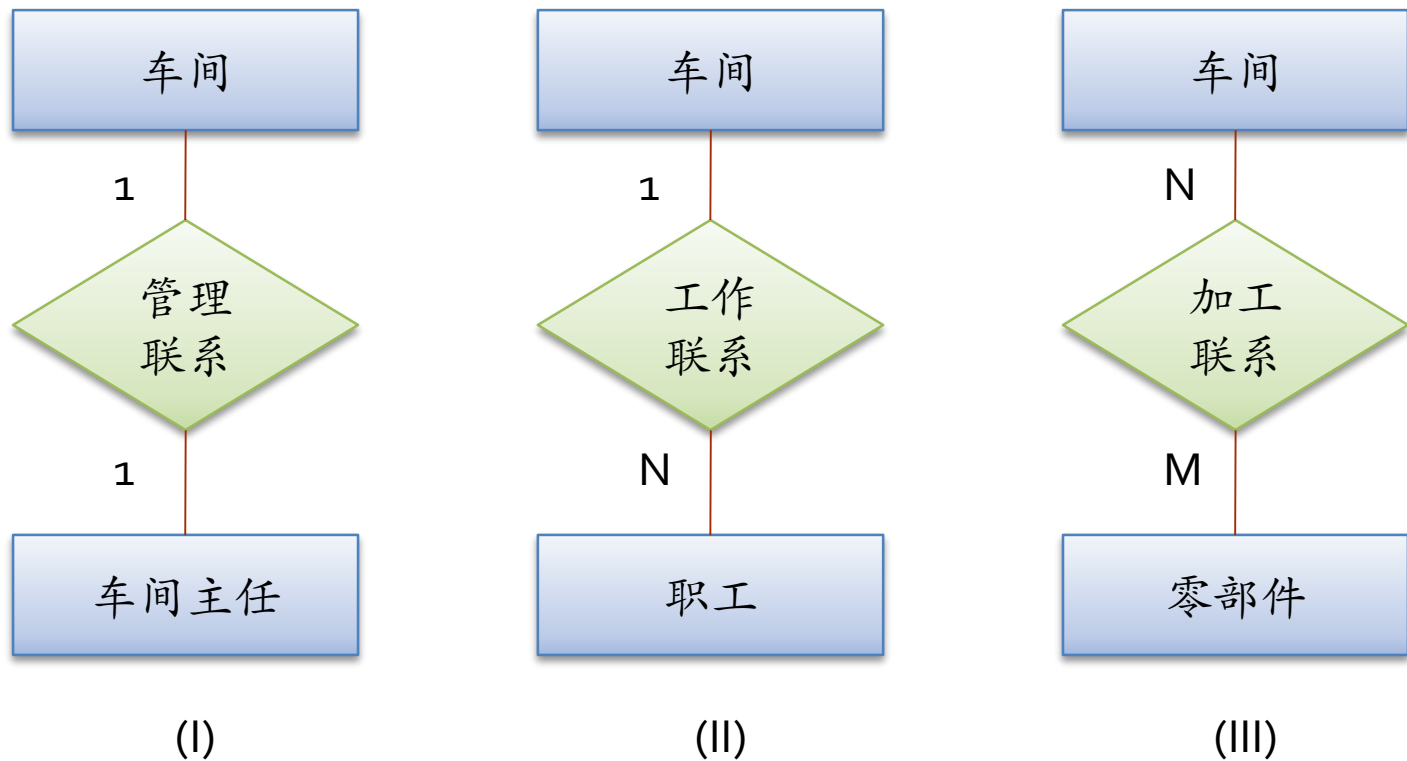
图(a) E-R符号图



图(b) 实体属性图



6.1 数据库概论

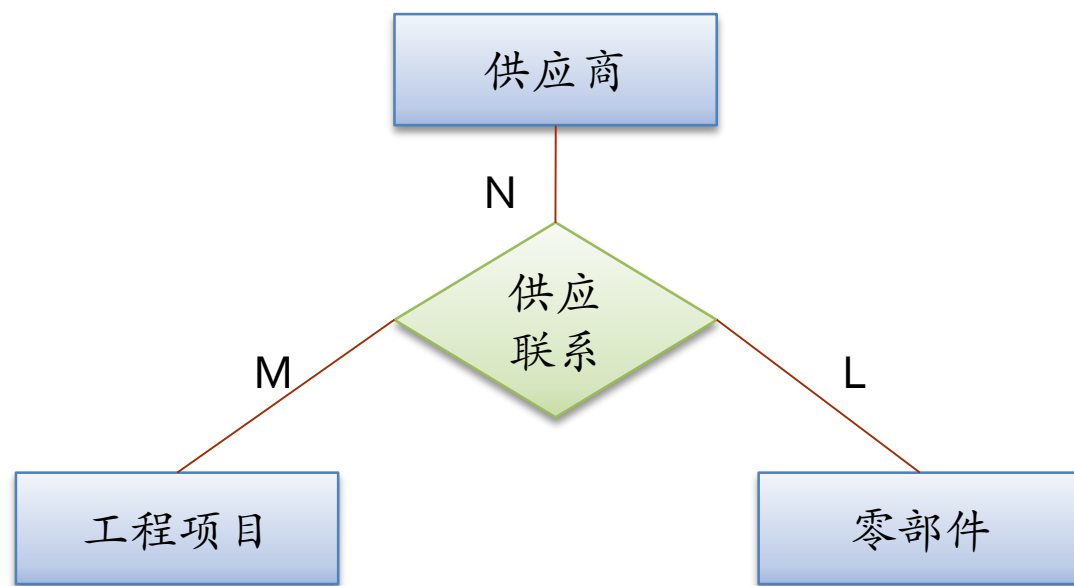


图(c) 两实体集的E-R图

6.1 数据库概论

* 信息模型

- ERD (E-R Diagram) 还可以表示多个实体之间的联系。三个实体集两两之间均为多对多联系



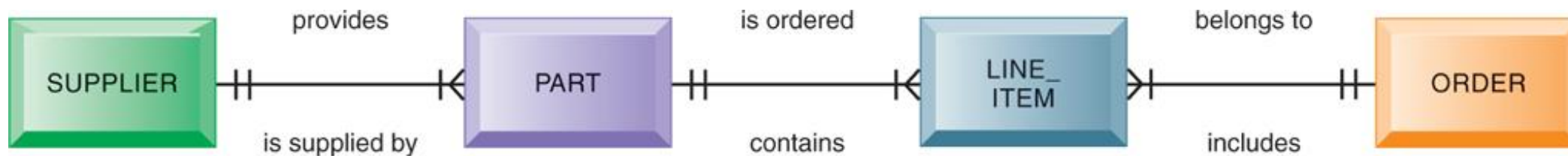
供应商联系图



6.1 数据库概论

* 信息模型

- 这幅图展示了供应商、零件、所订零件、订单这几个实体间的关系，这些实体可以建立起图10所描绘的数据模型



Copyright ©2014 Pearson Education



6.1 数据库概论

* 数据模型

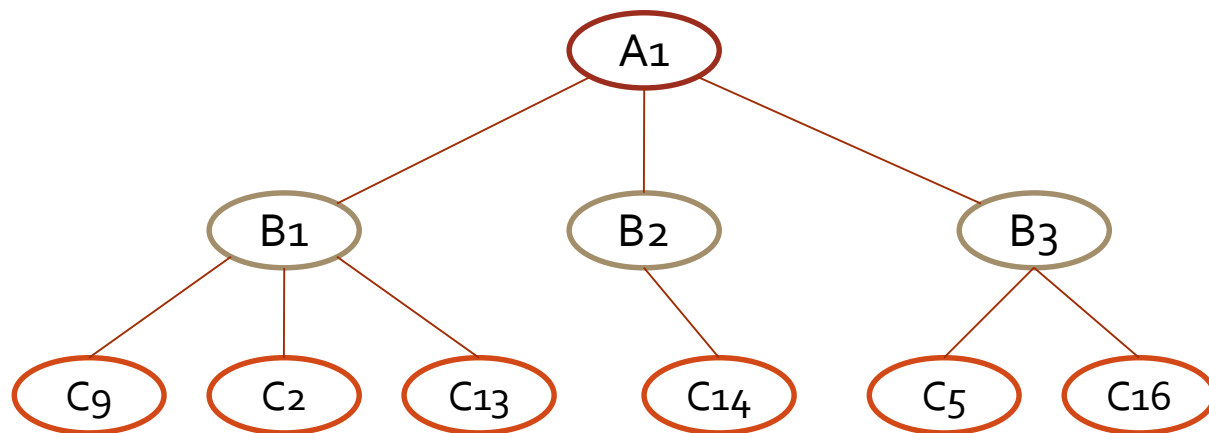
- 数据模型是**面向数据库中数据的逻辑结构的**，一个数据模型实际上是一组向用户提供的**规则**，这组规则规定了数据是如何组织的，其结构怎样以及相应的允许进行哪些操作，是数据组织中各层次内部、外部之间的联系的描述
- 数据模型以相应的信息模型为基础
- **数据模型与信息模型的关系**
 - 实体——记录；实体型——记录型；
 - 实体集——文件；个体——特定记录；
 - 属性——数据项；属性名——数据项型
 - 属性值——数据项值

6.1 数据库概论

* 数据模型

- 层次模型

- ▶ 有且仅有一个节点（实体集）无父（双亲）节点，这个节点即为树根，而其他节点有且仅有一个父节点。满足这两个条件的模型，称为层次模型



层次结构

6.1 数据库概论

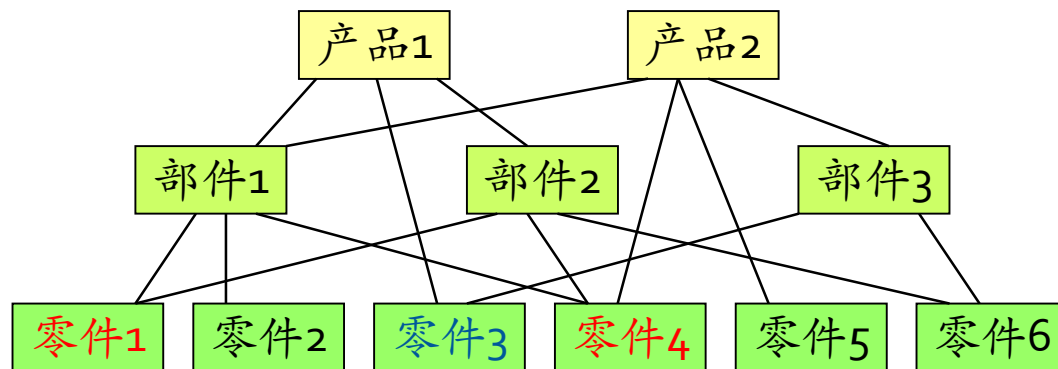
* 数据模型

- 网状模型

- 可以有一个以上节点（记录型）无父节点
- 至少有一个节点的父节点多于一个

- 层次模型与网络模型的不同之处主要有两点

- 层次模型中，从子节点到父节点的联系是唯一的，而网状模型中可以是不唯一的
- 网状模型中，还允许两个记录型间有两种以上的联系



网状结构图

6.1 数据库概论

* 数据模型

- 关系模型

- 关系模型将数据的逻辑结构归纳为满足一定条件的二维表的形式，称为一个关系（relation），关系又由关系框架和若干元组（tuple）组成，一个元组实际上就是二维表中的一行内容
- 关系模型就是指二维表格模型,因而一个关系型数据库就是由二维表及其之间的联系组成的一个数据组织

姓名	班号	选修课程名	任课教师	成绩
丁江	101911	计算机基础	刘红	90
王吉	102931	机械制图	李明	92
冯凡	103963	会计学	张扬	86
...	...	...	...	...

6.2 关系型数据库

* 关系模型与关系型数据库

- 关系型数据库是以关系模型为逻辑架构建立的
- 目前，市场上应用的产品主要是关系数据库产品
- 关系数据库产品主要有Oracle 11g、IBM DB2、Microsoft SQL Server、Sybase、Informix、MySQL等



6.2 关系型数据库

* 关系型数据库

SUPPLIER

Columns (Attributes, Fields)

Supplier_Number	Supplier_Name	Supplier_Street	Supplier_City	Supplier_State	Supplier_Zip
8259	CBM Inc.	74 5 th Avenue	Dayton	OH	45220
8261	B. R. Molds	1277 Gandolly Street	Cleveland	OH	49345
8263	Jackson Composites	8233 Micklin Street	Lexington	KY	56723
8444	Bryant Corporation	4315 Mill Drive	Rochester	NY	11344

Rows
(Records,
Tuples)

Key Field
(Primary Key)

PART

Part_Number	Part_Name	Unit_Price	Supplier_Number
137	Door latch	22.00	8259
145	Side mirror	12.00	8444
150	Door molding	6.00	8263
152	Door lock	31.00	8259
155	Compressor	54.00	8261
178	Door handle	10.00	8259

Primary Key

Foreign Key

6.2 关系型数据库

* 关系数据库的特点

- 简单、灵活、易用
- 关系数据语言具有较好的非过程化特性
- 有更高的数据独立性
- 易于实现对数据库的重组和控制，简化DBA的工作
- 易于实现各种不可预计的随机查询，从而对管理决策活动中数据库应用提供较好的支持
- 理论基础好，数据库的研究更加科学化
- 易于向推理系统和知识库系统的扩展和综合

6.2 关系型数据库

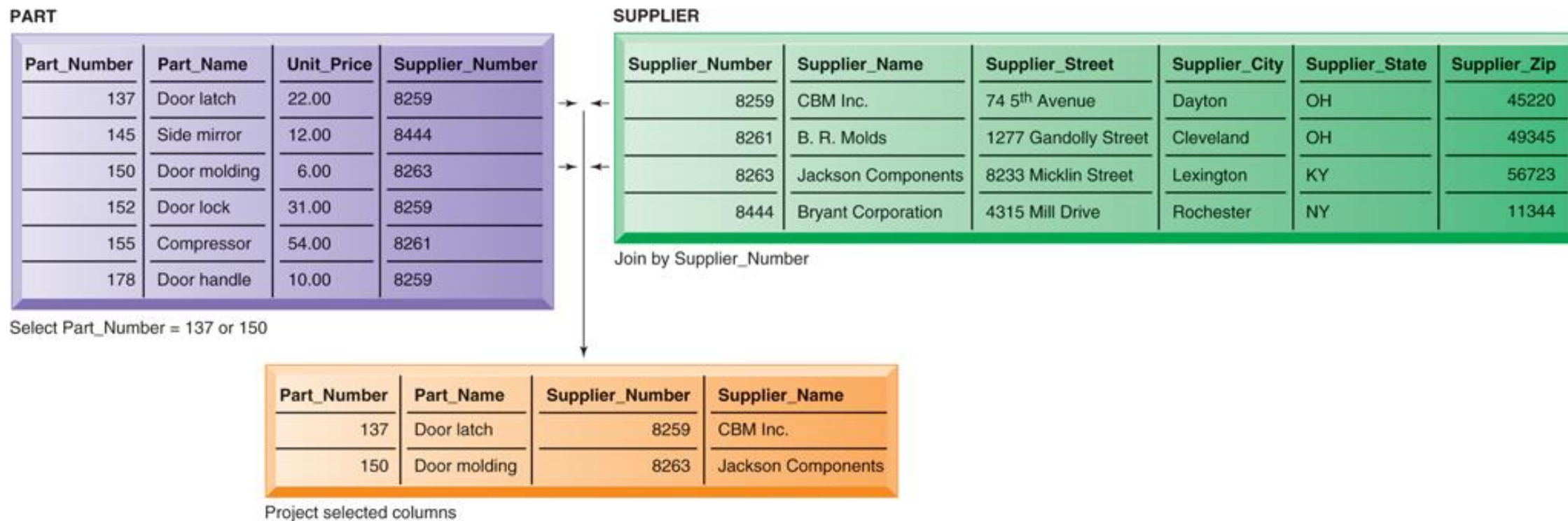
* 关系操作语言

- 关系操作集合提供了可以对数据进行的操作的描述
- 关系操作的特点是集合操作方式，即操作的对象和结果都是集合。关系操作通过关系操作语言来实现
- 对关系操作的描述，可以使用SQL（**Structured Query Language**）语言。SQL语言是介于关系代数和关系演算之间的一种查询语言
 - 数据定义语言(DDL)，例如：CREATE、DROP、ALTER等语句
 - 数据操作语言(DML)，例如：INSERT、UPDATE、DELETE等语句
 - 数据查询语言，例如：SELECT语句
 - 数据控制语言，例如：GRANT、REVOKE、COMMIT、ROLLBACK等语句
 - SQL语言包括两种主要程序设计语言类别的陈述式：数据定义语言(DDL)与数据操作语言(DML)

6.2 关系型数据库

* 关系操作语言

- 选择、连接和映射操作使得两张表被合并起来并且只列示被选择的属性



6.2 关系型数据库

* SQL查询的例子

- 上述例子说明了使用SQL语句，筛选出零件137或150的零件编号、零件名、供应商编号以及供应商名

```
SELECT PART.Part_Number, PART.Part_Name, SUPPLIER.Supplier_Number,  
SUPPLIER.Supplier_Name  
FROM PART, SUPPLIER  
WHERE PART.Supplier_Number = SUPPLIER.Supplier_Number AND  
Part_Number = 137 OR Part_Number = 150;
```

6.2 关系型数据库

* 关系的规范化

- 第一范式 (1NF)

- ✓ 关系中每个数据项必须是一个不可分的数据项
- ✓ 列是同质的。各列指定一个相异的名字，列的次序任意
- ✓ 各行相异，无重复的行，行的次序任意

- 第二范式 (2NF)

- ✓ 从第一范式中分解出新的关系，使每个关系里都可确定一个或几个属性作为关系的主关键字 (primary key, PK)，使该关系中的其他属性都完全依赖于它而定，从而消去非主属性对主关键字的部分依赖性 (partial dependence)，所得关系叫第二范式关系
- ✓ 关键字(学号, 课程名称) → 非关键字(姓名, 年龄, 成绩, 学分) 不满足2NF，此处姓名和年龄这两个非关键属性依赖于学号，而学号是关键字的一部分，因此称为部分依赖，这是违反2NF的。

6.2 关系型数据库

* 关系的规范化

- 第三范式 (3NF)

- ✓ 在第二范式基础上，进一步规范化就是要消去非主属性对主关键字的传递依赖性，变为第三范式
- ✓ 不应存在关键字段 $\rightarrow$ 非关键字段 $x \rightarrow$ 非关键字段的关系，如(学号) $\rightarrow$ (姓名, 年龄, 所在学院, 学院地点, 学院电话)

6.2 关系型数据库

* 关系的规范化

- 一个未规范化的关系会包含重复的组，比如一个订单可能会有很多的零件以及零件的供应商，但是在“订单号码”和“订单日期”之间只有一对一关系

ORDER (Before Normalization)

Order_ Number	Order_ Date	Part_ Number	Part_ Name	Unit_ Price	Part_ Quantity	Supplier_ Number	Supplier_ Name	Supplier_ Street	Supplier_ City	Supplier_ State	Supplier_ Zip
------------------	----------------	-----------------	---------------	----------------	-------------------	---------------------	-------------------	---------------------	-------------------	--------------------	------------------

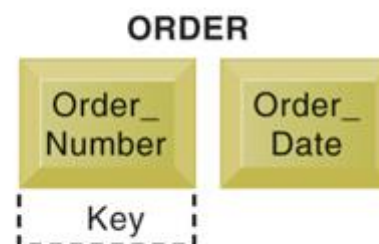
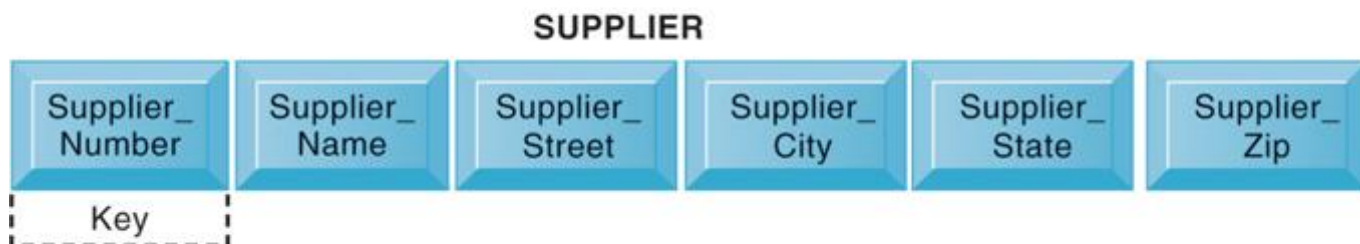
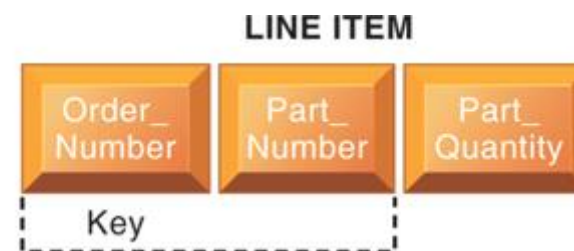
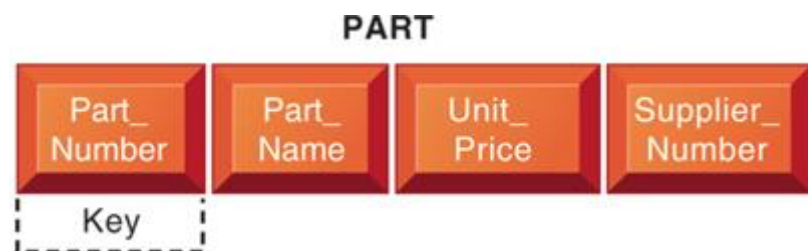
Copyright ©2014 Pearson Education

未规范化的订单关系

6.2 关系型数据库

* 关系的规范化

- 在规范化之后，原来的ORDER关系被分解成为四个小的关系，“订单”关系只剩下两个属性，“所订零件”关系有一个由“订单号码”和“零件号码”组成的联合键或连接键



Copyright ©2014 Pearson Education

规范化后的订单关系

6.2 关系型数据库

* 关系模型

- 完整性约束包括
 - 实体完整性
 - 参照完整性
 - 用户定义完整性

6.2 关系型数据库

* 全关系系统应遵循的基本准则

- ① 一个关系形的关系数据库管理系统必须能完全通过它的关系能力来管理数据库
- ② 信息准则 关系数据库管理系统的所有信息都应该在逻辑一级上用表中的值这一种方法显式的表示
- ③ 保证访问准则 依靠表名、主码和列名的组合，保证能以逻辑方式访问关系数据库中的每个数据项
- ④ 空值的系统化处理 全关系的数据库管理系统支持空值的概念，并用系统化的方法处理空值
- ⑤ 基于关系模型的动态的联机数据字典 数据库的描述在逻辑级上和普通数据采用同样的表述方式

6.2 关系型数据库

* 全关系系统应遵循的基本准则

- ⑥ 统一的数据子语言 一个关系数据库管理系统可以具有几种语言和多种终端访问方式，但必须有一种语言，它的语句可以表示为严格语法规定的字符串，并能全面的支持各种规则。
- ⑦ 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。
- ⑧ 高级的插入、修改和删除操作 系统应该对各种操作进行查询优化。
- ⑨ 数据的物理独立性 无论数据库的数据在存储表示或访问方法上作任何变化，应用程序和终端活动都保持逻辑上的不变性。

6.2 关系型数据库

* 全关系系统应遵循的基本准则

- ⑩ 数据逻辑独立性 当对基本关系进行理论上信息不受损害的任何改变时，应用程序和终端活动都保持逻辑上的不变性。
- ⑪ 数据完整的独立性 关系数据库的完整性约束条件必须是用数据库语言定义并存储在数据字典中的。
- ⑫ 分布独立性 关系数据库管理系统在引入分布数据或数据重新分布时保持逻辑不变。
- ⑬ 无破坏准则 如果一个关系数据库管理系统具有一个低级语言，那么这个低级语言不能违背或绕过完整性准则。

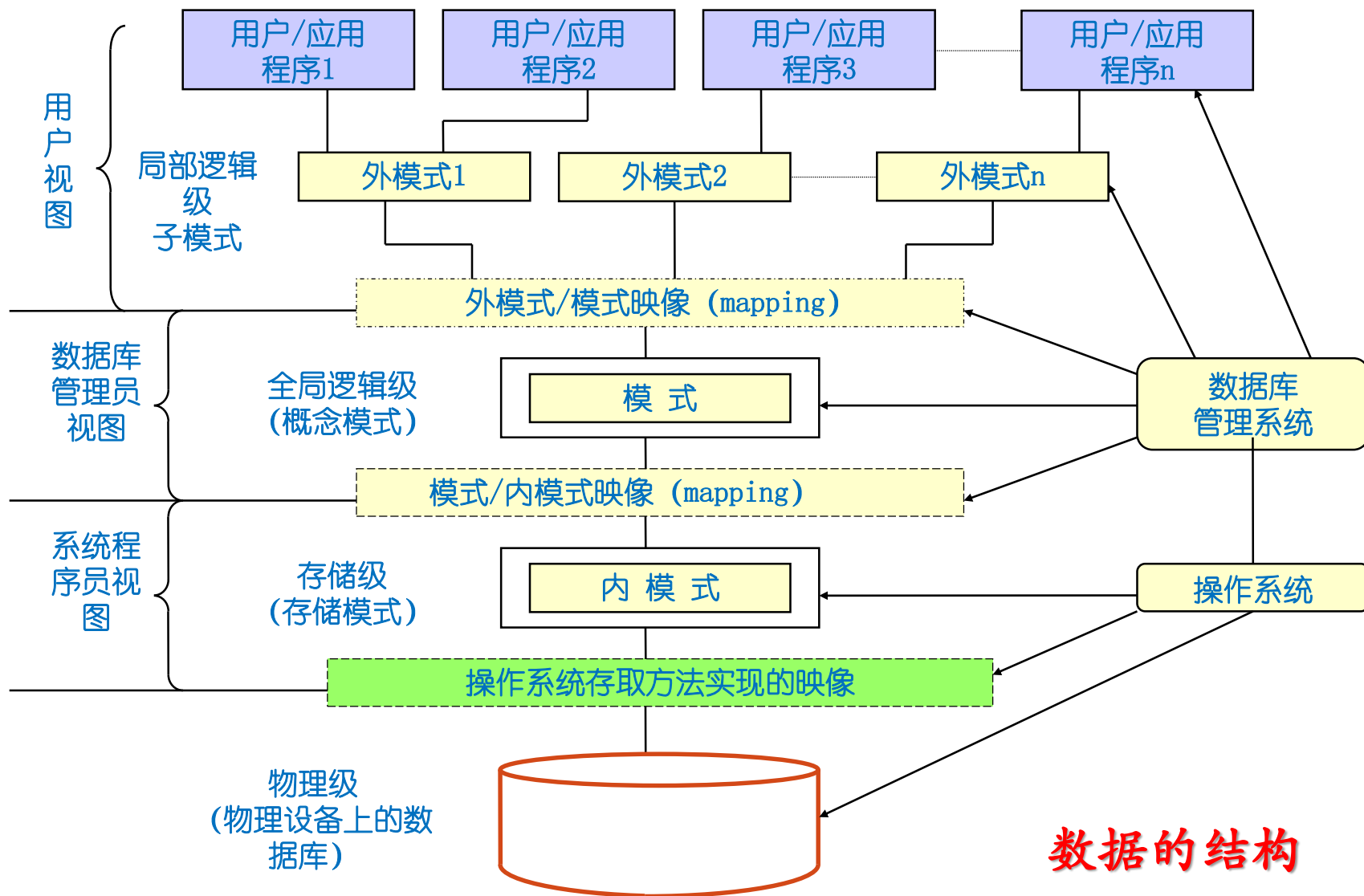
6.3 数据库管理系统

* 数据库系统结构

— 数据库系统的逻辑结构

- **外模式** (External schema) : 外模式也称子模式 (Subschema) , 它是局部逻辑级的结构, 是用户可以看到并获准使用的那部分数据的逻辑结构, 因此外模式又叫做用户视图
- **模式** (Schema) : 模式也叫做概念模式 (Conceptual Schema) , 它是全局逻辑级的, 是数据库的整体逻辑结构。
- **内模式** (Internal schema) : 内模式是数据库的存储结构, 是为了提高模式对存储设备的独立性而设立的。

6.3 数据库管理系统



6.3 数据库管理系统

* 数据库系统结构

— 数据库系统 (DBS) 的构成

- 硬件
- 软件：主要包括数据库管理系统 (DBMS) 软件以及支持DBMS运行的操作系统和有关开发工具
- 数据库系统中相关人员：数据库管理员 (DBA)、系统分析员、应用程序员和用户是管理、开发和使用数据库的主要人员
- 数据库：它是保存在存贮介质上的大量的相关数据的集合，这也是数据库系统的重要组成部分

6.3 数据库管理系统

* 数据库系统结构

— 数据库管理系统 (DBMS) 的主要功能

- 定义数据库的功能
- 管理数据库的功能
- 维护数据库的功能
- 数据通信的功能

— 数据库管理系统 (DBMS) 的组成

- 数据描述语言 (Data Description Language, DDL)
- 数据操纵语言 (Data Manipulation Language, DML)
- 数据库管理例行程序

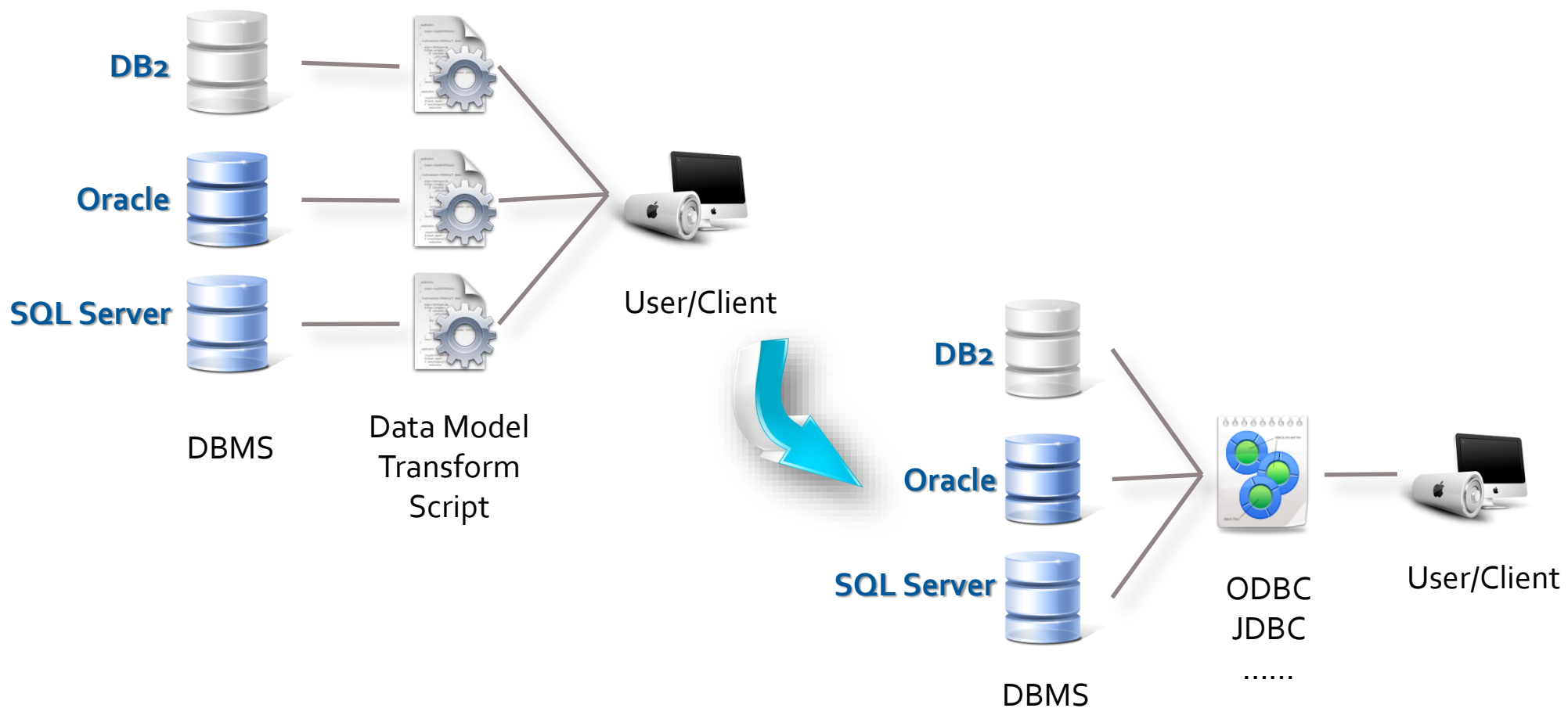
6.3 数据库管理系统

* 异构数据源的访问与数据交换

- 对于那些相对独立的软件供应商而言，经常要为每一个DBMS编写一个版本的应用程序，此时就需要开放的数据库连接。为此，在应用系统中广泛使用中间件技术，以隐藏各种复杂性，屏蔽各种系统之间的差异。
- 开放数据库连接（Open Database Connectivity, ODBC）和Java数据库互连（Java Database Connectivity, JDBC）是两种优秀的中间件技术，它们的出现为数据库应用系统的开发指明了道路

6.3 数据库管理系统

* 异构数据源的访问与数据交换



6.3 数据库管理系统

* 面向对象数据库

- 面向对象数据库（OODB）是一种数据库系统，充分支持面向对象概念与机制。它的用户主要是应用软件和系统软件的开发人员。这类系统的优点是可以与面向对象程序设计语言一体化，开发人员不必再学习新的数据库语言

* 对象—关系数据库

- 从传统的关系数据库加以扩展，增加面向对象特性，把面向对象技术与关系数据库相结合，可建立对象—关系数据库管理系统（即ORDBMS）

* 多媒体数据库

- 多媒体数据库系统是指能够存储、查询和管理多媒体数据的数据库系统，实现对格式化和非格式化的多媒体数据的存储、管理和查询

6.3 数据库管理系统

* 分布式数据库

- 分布式数据库系统通常使用较小的计算机系统，每台计算机可单独放在一个地方，每台计算机中都可能存有DBMS的一份完整拷贝副本，或者部分拷贝副本，并具有自己局部的数据库，位于不同地点的许多计算机通过网络互相连接，共同组成一个完整的、全局的逻辑上集中、物理上分布的大型数据库
- 分布式数据库系统抽象为四层的结构模式，分别为全局外层、全局概念层、局部概念层和局部内层，在各层间还有相应的层间映射。这种4层模式适用于同构型分布式数据库系统，也适用于异构型分布式数据库系统

6.3 数据库管理系统

* 分布式数据库优势

- 优势

- ① 降低了数据传送代价。大多数的对数据库的访问操作都是针对局部数据库的，而不是对其他位置的数据库访问；
- ② 系统的可靠性提高了很多。当网络出现故障时，仍然允许对局部数据库的操作，而且一个位置的故障不影响其他位置的处理工作，只有当访问出现故障位置的数据时，在某种程度上才受影响；
- ③ 便于系统的扩充，增加一个新的局部数据库，或在某个位置扩充一台适当的小型计算机，都很容易实现。

- 劣势

- 有些功能要付出更高的代价。例如，为了调配在几个位置上的活动，事务管理的性能比在中心数据库时花费更高，而且甚至抵消许多其他的优点

6.3 数据库管理系统

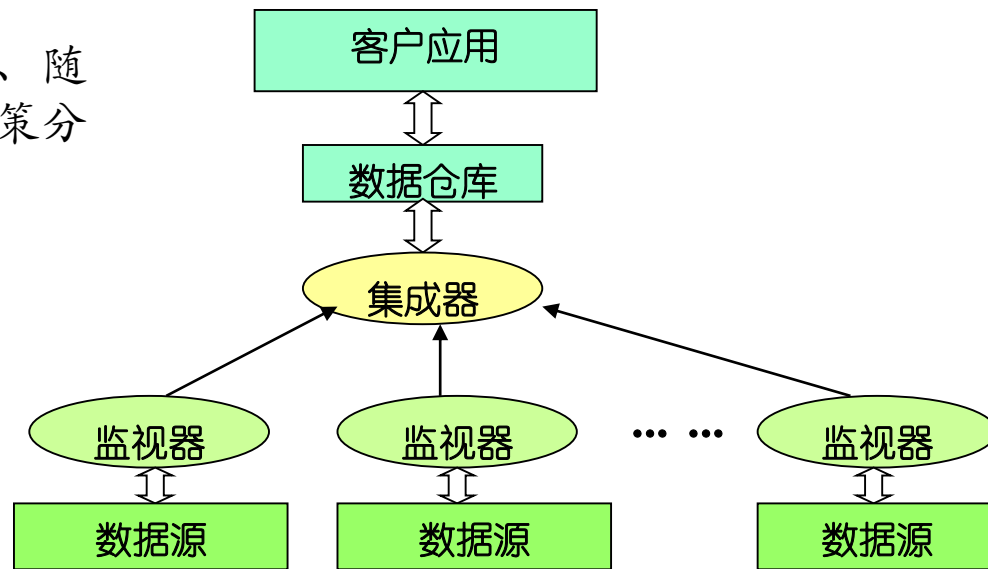
* 分布式数据库特点

- 多数处理就地完成；
- 各地的计算机由数据通信网络相联系；
- 克服了中心数据库的弱点：降低了数据传输代价；
- 提高了系统的可靠性，局部系统发生故障，其他部分还可继续工作；
- 各个数据库的位置是透明的，方便系统的扩充；
- 为了协调整个系统的事务活动，事务管理的性能花费高；

6.4 数据仓库与数据挖掘

* 数据仓库

- W. H. Inmon 于1990年提出
- 定义
 - 数据仓库就是一个面向主题的、集成的、不可更新的、随时间不断变化的数据集合，用以支持企业或组织的决策分析处理
- 特点
 - 数据仓库是面向主题的
 - 数据仓库是集成的
 - 数据仓库是稳定的
 - 数据仓库是随时间变化的
- 结构
 - 斯坦福大学“WHPS”课题组提出的一个基本的数据仓库模型



Stanford WHPS Data
Warehouse Model

6.4 数据仓库与数据挖掘

* 数据仓库涵义

- 数据仓库是一种资讯系统的资料储存理论，此理论强调利用某些特殊资料储存方式，让所包含的资料，特别有利于分析处理，以产生有价值的资讯并依此作决策。
- 利用数据仓库方式所存放的资料，具有一旦存入，便不随时间而更动的特性，同时存入的资料必定包含时间属性，通常一个数据仓库皆会含有大量的历史性资料，并利用特定分析方式，自其中发掘出特定资讯
- 数据仓库的建制不仅只是资讯工具技术面的运用，在规划和执行面更需对产业知识、行销管理、市场定位、策略规划等相关条件有深入的了解，才能真正发挥数据仓库以及后续分析工具的价值，提升组织竞争力

6.4 数据仓库与数据挖掘

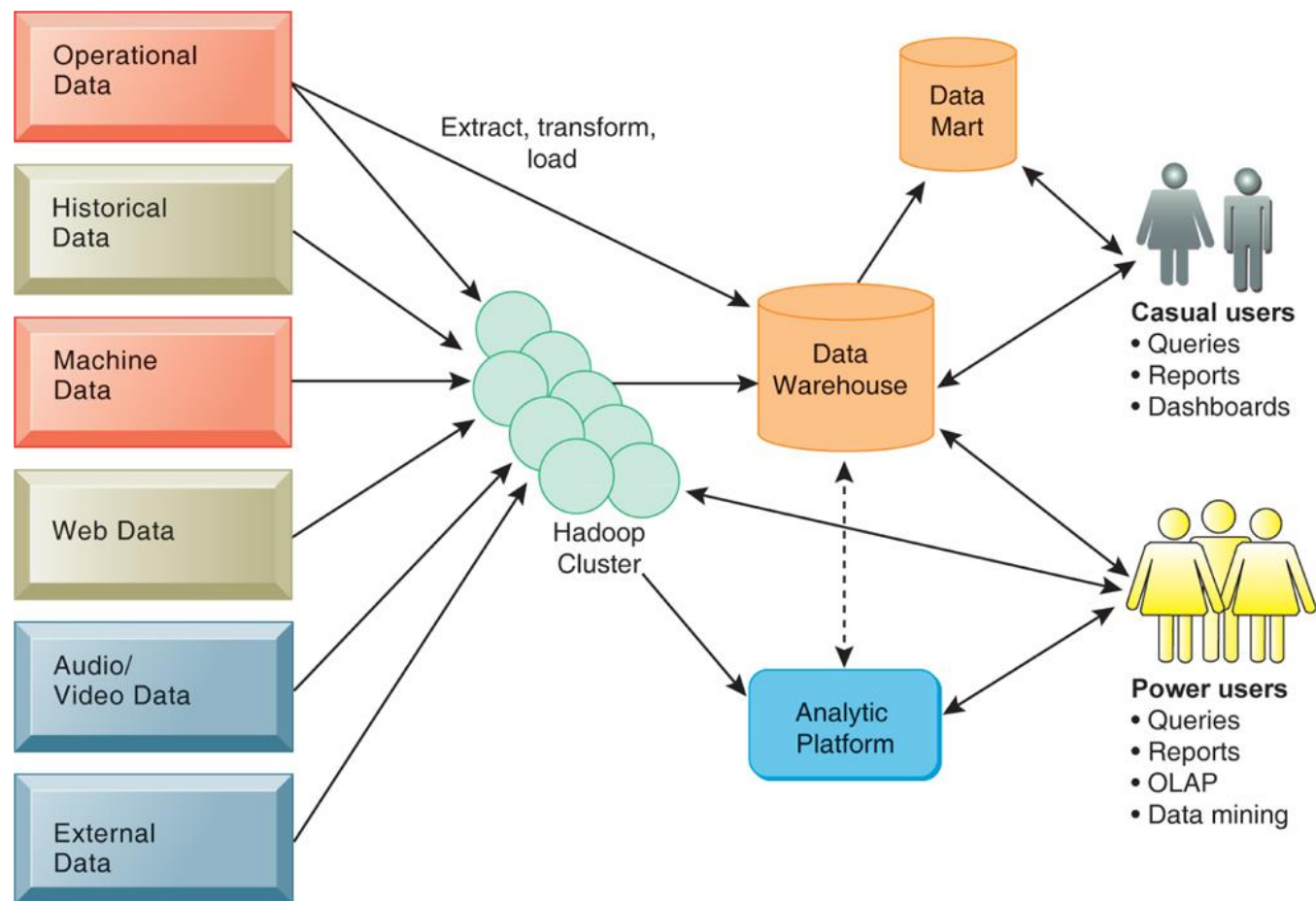
* 数据仓库涵义

- 有别于一般联机交易处理系统，数据模型设计是一个数据仓库设计的地基，目前两大主流理论分别为采用**正规方式(normalized approach)**或**多维方式(dimensional approach)**进行数据模型设计，数据模型可以分为逻辑与实体数据模型
- **逻辑数据模型**陈述业务相关数据的关系，基本上是一种**与数据库无关的结构设计**，通常均会**采用正规方式设计**，主要精神是从企业业务领域的角度及高度订出Subject Area Model，再逐步向下深入到Entities、Attributes，在设计时不会考虑未来采用的数据库管理系统，也不需考虑分析性能问题
- **实体数据模型**则**与数据库管理系统有关**，是建置在该系统上的数据架构，故设计时需考虑数据型态(Data Type)、空间及性能相关的议题。实体数据模型设计，则**较多有采用正规方式或多维方式的讨论**，但从实务上来说，不执著于理论，能与业务需要有最好的搭配，才是企业在建置数据仓库时的正确考量

6.4 数据仓库与数据挖掘

* 数据仓库组件

- 现代商务智能基础设施具有管理和分析来自多个数据源的大规模不同类型数据的能力与工具。其中包括了为一般企业用户设计的易用的查询和报表生成工具，以及为高级用户设计的复杂的分析工具



6.4 数据仓库与数据挖掘

* 数据仓库特性

- **主题导向**(Subject-Oriented)。有别于一般OLTP系统，数据仓库的资料模型设计，着重将资料按其意义归类至相同的主题区(subject area)，因此称为主题导向。举例如Party、Arrangement、Event、Product等
- **集成性**(Integrated)。资料来自企业各OLTP系统，在数据仓库中是集成过且一致的
- **时间差异性**(Time-Variant)。资料的变动，在数据仓库中是能够被纪录以及追踪变化的，有助于能反映出能随着时间变化的资料轨迹
- **不变动性**(Nonvolatile)。资料一旦确认写入后是不会被取代或删除的，即使资料是错误的亦同。(i.e.错误的后续修正，便可因上述时间差异性的特性而被追踪)

6.4 数据仓库与数据挖掘

* 数据仓库

— 关键技术

- 数据的抽取是数据进入仓库的入口。数据抽取在技术上主要涉及互连、复制、增量、转换、调度和监控等方面
- 数据的存储和管理是数据仓库的关键。对数据进行清洁、增强、变换、和加载的处理并且根据在元数据中所描述的检验规则和专家知识，消除数据的不一致性和不合理性
- 数据的表现。数据仓库表现工具包括报表生成工具、OLAP、数据挖掘（Data Mining）、决策支持工具
- 数据仓库就是一个面向主题的、集成的、不可更新的、随时间不断变化的数据集合，它用以支持企业或组织的决策分析处理

6.4 数据仓库与数据挖掘

* 数据仓库

- 当今的数据处理大致可以分成两大类
 - 联机事务处理**OLTP** (On-Line Transaction Processing), OLTP是传统的关系型数据库的主要应用, 主要是基本的、日常的事务处理, 例如银行交易
 - 联机分析处理**OLAP** (On-Line Analytical Processing), OLAP是数据仓库系统的主要应用, 支持复杂的分析操作, 侧重决策支持, 并且提供直观易懂的查询结果。

6.4 数据仓库与数据挖掘

	OLTP	OLAP
用户	操作人员，低层管理人员	决策人员，高级管理人员
功能	日常操作处理	分析决策
DB 设计	面向应用	面向主题
数据	当前的, 最新的细节的, 二维的分立的	历史的, 聚集的, 多维的集成的, 统一的
存取	读/写数十条记录	读上百万条记录
工作单位	简单的事务	复杂的查询
用户数	上千个	上百个
DB 大小	100MB-GB	100GB-TB

6.4 数据仓库与数据挖掘

下雨吃蛋糕
气温上升三明治销量增长

* 数据挖掘技术概述

— 数据挖掘技术的发展

- 20世纪80年代，人们在新的神经网络理论的指导下，用机器学习的方法处理大型商业数据库，从而产生了一个新的术语—数据库中的知识发现（Knowledge Discovery In Database, KDD）

— 数据挖掘技术的定义

- 从技术的角度看，数据挖掘就是应用一系列技术从大型数据库或数据仓库的数据中提取人们感兴趣的信息和知识，这些知识或信息是隐含的、事先未知而潜在有用的，提取的知识多表示为概念、规则、规律和模式等形式

6.4 数据仓库与数据挖掘

* 数据挖掘技术概述

- 数据挖掘的任务
 - 数据挖掘的任务分为两类：描述和预测
 - 通过挖掘，我们一般可以得到以下的几种模式
 - 概念描述
 - 关联规则
 - 分类和预测
 - 聚类
 - 时序模式和序列模式

6.4 数据仓库与数据挖掘

* 数据挖掘、OLAP和数据仓库

- 数据仓库可以作为数据挖掘和OLAP等分析工具的资料来源，由于存放于数据仓库中的资料，必需经过筛选与转换，因此可以避免分析工具使用错误的资料，而得到不正确的分析结果。
- 数据挖掘和OLAP同为分析工具，其差别在于OLAP提供用户一便利的多维度观点和方法，以有效率的对数据进行复杂的查询动作，其预设查询条件由用户预先设定，而数据挖掘，则能由资讯系统主动发掘资料来源中，未曾被查觉的隐藏资讯，和透过用户的认知以产生知识。
- 数据挖掘技术是经由自动或半自动的方法探勘及分析大量的资料，以创建有效的模型及规则，而企业透过数据挖掘更了解他们的客户，进而改进他们的行销、业务及客服的运作。数据挖掘是数据仓库的一种重要运用。基本上，它是用来将你的资料中隐藏的资讯挖掘出来，所以 Data Mining 其实是所谓的 Knowledge Discovery 的一部份，Data Mining 使用了许多统计分析与 Modeling 的方法，到资料中寻找有用的特征（Patterns）以及关连性（Relationships）。Knowledge Discovery 的过程对 Data Mining 的应用成功与否有重要的影响，只有它才能确保 Data Mining 能获得有意义的结果。

6.5 大数据与云计算

* 结构化数据

- 即行数据，存储在数据库里，可以用二维表结构来逻辑表达实现的数据
- 列数据不可再细分，列内数据具有相同的类型
- 关系型数据库中的数据均为结构化数据
- 结构化数据是非结构化数据的特例

* 非结构化数据

- 指其字段长度可变，并且每个字段的记录又可以由可重复或不可重复的子字段构成的数据库，用它不仅可以处理结构化数据（如数字、符号等信息）而且更适合处理非结构化数据（全文文本、图象、声音、影视、超媒体等信息）

6.5 大数据与云计算

* 数据清洗

- 因为数据仓库中的数据是面向某一主题的数据的集合，这些数据从多个业务系统中抽取而来而且包含历史数据，这样就避免不了有的数据是错误数据、有的数据相互之间有冲突，这些错误的或有冲突的数据显然是我们不想要的，称为“脏数据”。因此需要按照一定的规则把“脏数据”“洗掉”，这就是数据清洗。
- 数据清洗的任务是过滤那些不符合要求的数据，将过滤的结果交给业务主管部门，确认是否过滤掉还是由业务单位修正之后再进行抽取。不符合要求的数据主要有不完整的数据、错误的数据、重复的数据三大类

6.5 大数据与云计算

* 非结构化数据库

- 随着网络技术的发展，特别是Internet和Intranet技术的飞快发展，使得非结构化数据的数量日趋增大。此时，主要用于管理结构化数据的关系数据库的局限性暴露地越来越明显。数据库技术相应地进入了“后关系数据库时代”，发展进入基于网络应用的非结构化数据库时代。
- 所谓非结构化数据库，是指数据库的变长纪录由若干不可重复和可重复的字段组成，而每个字段又可由若干不可重复和可重复的子字段组成。简单地说，非结构化数据库就是字段可变的数据库
- 可以预言，随着网络技术和网络应用技术的飞快发展，完全基于Internet应用的非结构化数据库，或称云数据库，将成为继层次数据库、网状数据库和关系数据库之后的又一重点、热点技术

6.5 大数据与云计算

* 大数据 Big Data 或 Metadata

- 大数据由巨型数据集组成，这些数据集大小常超出人类在可接受时间下进行截取、管理、处理、并整理成为人类所能解读信息的能力。大数据的大小经常改变，截至2012年，单一数据集的大小从数兆字节（TB）至数十兆亿字节（PB）不等
- 在总数据量相同的情况下，与个别分析独立的小型数据集（data set）相比，将各个小型数据集合并后进行分析可得出许多额外的信息和数据关系性，可用来察觉商业趋势、判定研究质量、避免疾病扩散、打击犯罪或测定实时交通路况等

* 数据集大小增长的部分原因来自于信息持续从各种来源被广泛收集，这些来源包括搭载传感设备的移动设备、高空传感科技（遥感）、软件记录、相机、麦克风、无线射频辨识（RFID）和无线传感网络

6.5 大数据与云计算

* 大数据特征

- Volume, 量, 即数据大小
- Velocity, 速度, 即数据输入输出的速度
- Variety, 多变, 即多样性
- Veracity, 真实性
- Value, 价值密度低

* 大数据处理技术

- 大数据必须借由计算机对数据进行统计、比对、解析方能得出客观结果
- 数据挖掘 (Data Mining) 则是在探讨用以解析大数据的方法

6.5 大数据与云计算

* Hadoop

- 使得大数据能够大范围在低成本的PC上进行分布式并行处理
- 关键服务
 - Hadoop 分布式文件系统(HDFS): 数据存储
 - MapReduce: 将工作分为不同的簇
 - Hbase: 非SQL数据库
- Facebook、Yahoo、NextBio都使用Hadoop

6.5 大数据与云计算

* 内存计算

- 运用于大数据分析
- 进行数据存储，以避免从存取的数据中生成需要数据的延迟
- 可以以小时/天为单位减少处理数据的时间
- 最佳利用硬件

6.5 大数据与云计算

* 分析工具 - 聚焦于关系、模式、趋势

- 为大量数据的分析提供稳固的分析方法，帮助用户更好的做出商业决策
 - 多维数据分析(OLAP)
 - 数据挖掘
 - 文本挖掘
 - 网络挖掘

6.5 大数据与云计算

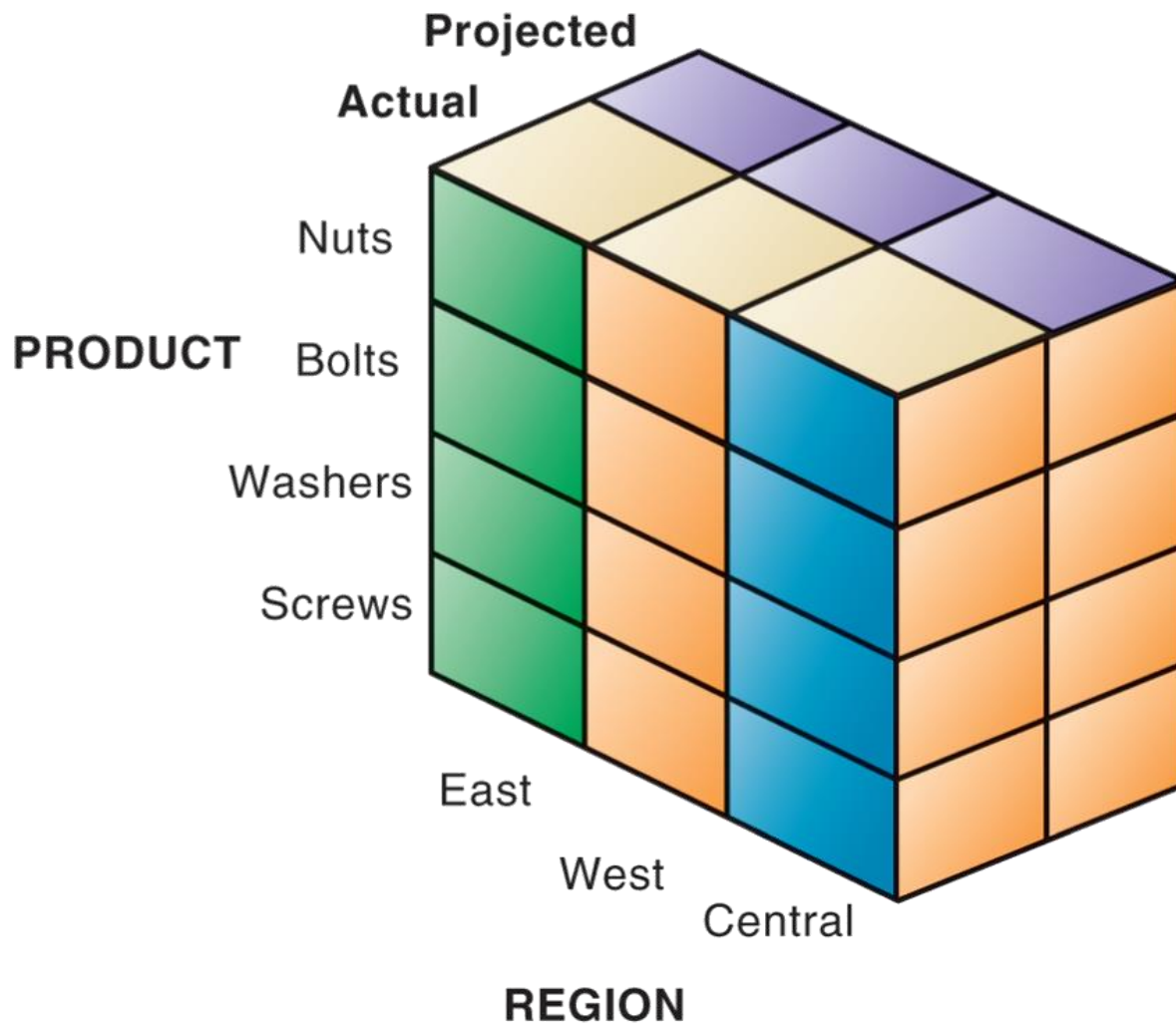
* OLAP

- 支持多维度的数据分析
 - 将数据视作多维度的
 - 信息的每一方面作为一个维度（产品、价格、成本、地区、时期）
 - 例子：六月东部地区相比于其他地区销售了多少台洗衣机？
- OLAP能够满足迫切的特定需求

6.5 大数据与云计算

* 基于OLAP的多维度数据分析

- 这个视图显示产品和区域。如果你旋转立方体90度，将展现的是产品对比实际销量和目标销量。如果你再转90度，你将会看到不同区域对比实际销量和目标销量。通过转动，也可以得到其他的视图。



6.5 大数据与云计算

* 数据挖掘

- 发现隐藏的模式以及数据之间的关系
 - 例如: 客户购买方式
- 推断出可能遵循的行为规律
- 通过数据挖掘可以进行的分析
 - 关联分析
 - 序列分析
 - 分类
 - 聚类
 - 预测

6.5 大数据与云计算

* 文本挖掘

- 从大型的非结构化数据集中提取关键元素
 - 存储的邮件
 - 呼叫中心的单据
 - 法律案例
 - 专利说明
 - 服务报告等等
- 情感分析软件
 - 在电子邮件、博客、社会化网络中挖掘观点

6.5 大数据与云计算

* 网络挖掘

- 在网络中挖掘和分析有价值的模式以及信息
 - 探索消费者行为
 - 评估网站的实效等等
- 网页内容挖掘
 - 挖掘网页
- 网页结构挖掘
 - 分析网页链接的来龙去脉
- 网页使用挖掘
 - 挖掘网站服务器中用户交互记录情感分析软件

6.5 大数据与云计算

* 大数据需要借助云计算

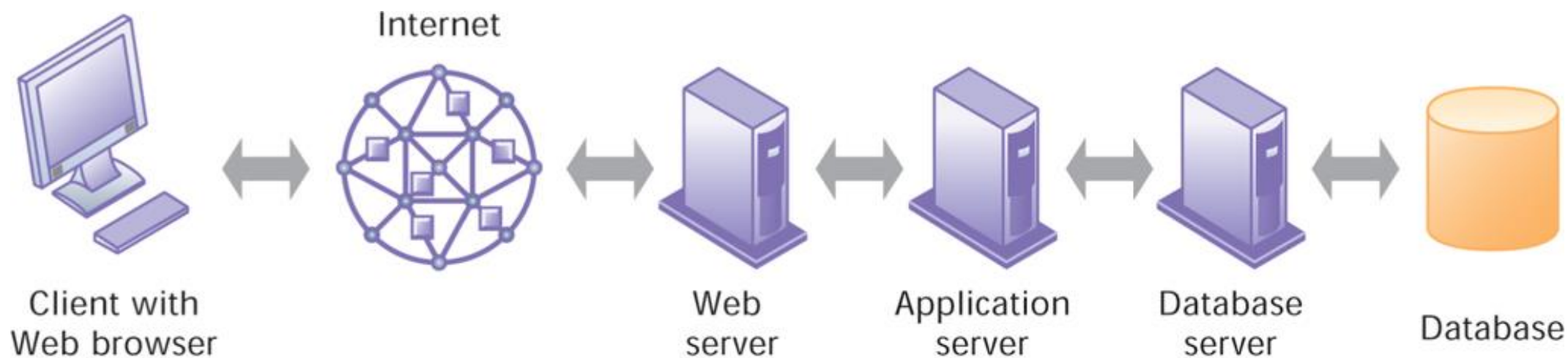
- 大数据几乎无法使用大多数的数据库管理系统处理，而必须使用“在数十、数百甚至数千台服务器上同时平行运行的软件”，即需要借助云计算的处理能力
- 大数据的定义取决于持有数据组的机构之能力，以及其平常用来处理分析数据的软件之能力，即由云计算的能力决定大数据能力

* 数据集大小增长的部分原因来自于信息持续从各种来源被广泛收集，这些来源包括搭载传感设备的移动设备、高空传感科技（遥感）、软件记录、相机、麦克风、无线射频辨识（RFID）和无线传感网络

6.5 大数据与云计算

* 网络与企业内部数据库的连接

- 用户利用个人电脑上的浏览器浏览，通过互联网访问企业内部数据



6.5 大数据与云计算

* 大数据应用

- 大数据的应用示例包括了大科学、RFID、传感设备网络、天文学、大气学、基因组学、生物学、大社会数据分析、互联网文件处理、制作互联网搜索引擎索引、通信记录明细、军事侦查、社交网络、通勤时间预测、医疗记录、照片图像和图像封存、大规模的电子商务等

6.5 大数据与云计算

* 大数据公司的类型

- 数据的拥有者（数据源）。比如电信公司、电商、银行等拥有海量用户数据的公司
- 大数据咨询。专门提供技术服务的公司，从基础设施规划、建设、维护，到软件开发、数据分析等业务
- 大数据分析工具开发。进行大数据利用和分析工具的开发
- 整合应用。搜集或购买数据，结合人工智能AI解决实际问题

* 大数据行业最有价值的部分在于如何利用机器去处理数据得到洞见，影响组织和个人的行为，从而改变世界

6.5 大数据与云计算

* 大数据业务类型

- 分析用户行为，改进产品和营销，比如Linkedin的推荐系统和iBeacon实现店内营销
- 统筹大量分散个体，利用大数据实现精确有效的预测和规划，比如Uber和Amazon Fresh及Grub Market
- 分析识别各种类型的数据，开发更智能的设备和程序，比如Google大脑及无人车，Nest的智能设备等

6.5 大数据与云计算

* 大数据应用举例-社会学

- 大数据产生的背景离不开脸书、微博等社交网络的兴起，人们每天通过这种自媒体传播信息或者沟通交流，由此产生的信息被网络记录下来，社会学家可以在这些数据的基础上分析人类的行为模式、交往方式等。美国的涂尔干计划就是依据个人在社交网络上的数据分析其自杀倾向，该计划从美军退役士兵中拣选受试者，透过脸书的行动App收集资料，并将用户的活动数据传送到一个医疗资料库。收集完成的数据会接受人工智能系统分析，接着利用预测程序来即时监视受测者是否出现一般认为具伤害性的行为

6.5 大数据与云计算

* 大数据应用举例-商业领域

- 人类目前生活在一个密布各类信息设施的环境中，人类每日行为中的很大一部分通过这些设施建立与外部环境的联系和信息交换，比如邮件、购物、通信、娱乐、新闻、阅读，等等，此类数据的长期累积，形成了个人特征和行为模式的信息记录，对此类数据进行分析，就可以获得商务领域所需要的有价值的客户特征信息，从而可以定向性地开展针对性营销，降低商业成本，提高营销效率



THE END

