

2022-2023秋季课程:数据科学与大数据导论

Introduction to Data Science and Big data


Chapter 3: Big Data Analytics Fundamentals

曹劲舟 助理教授

深圳技术大学 大数据与互联网学院

caojinzhou@sztu.edu.cn

2022年9月



Outline

□ Data Types and Sources 数据模型

□ Data Collection 数据采集

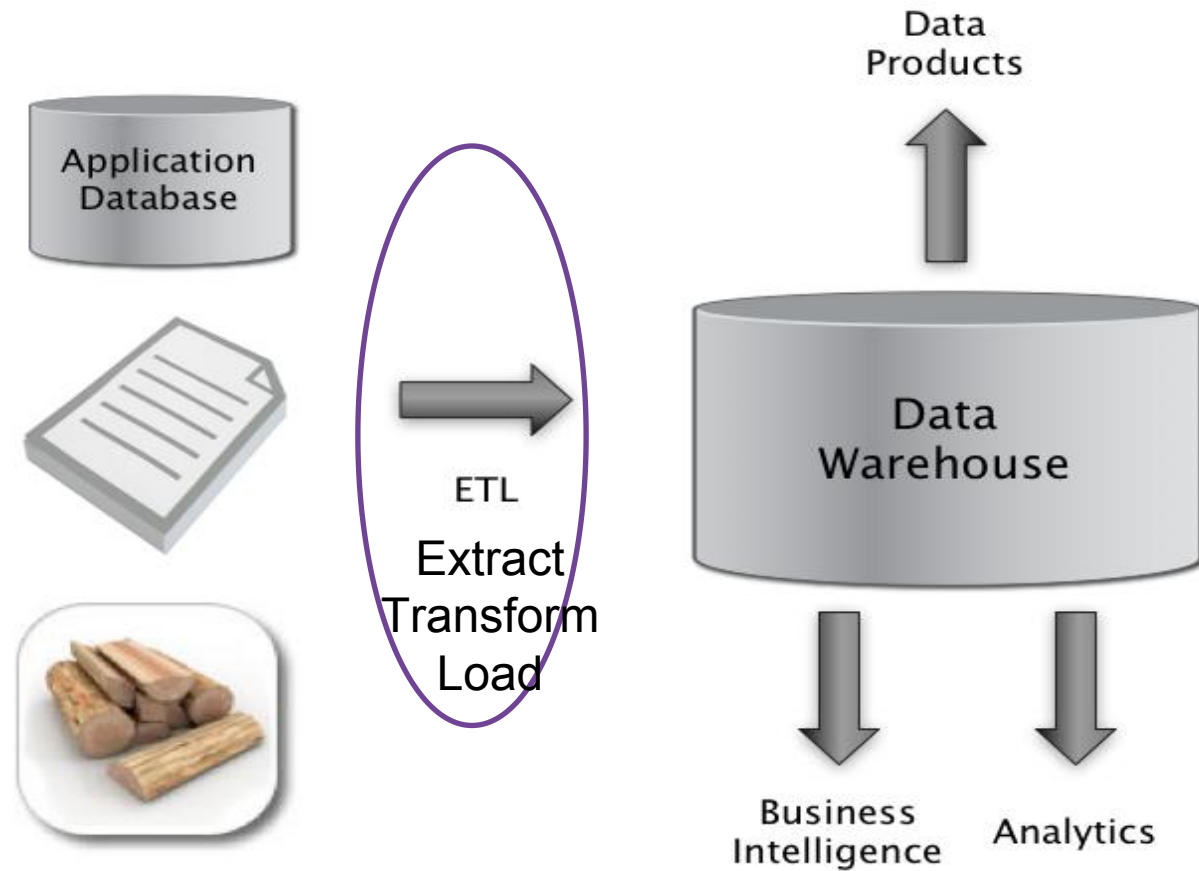
□ Data Preprocessing 数据预处理

□ Exploratory Data Analysis 数据探索性分析

数据科学的工作流程——ETL

□ ETL

- We need to **Extract** data from the **source(s)**
 - Sources: file, database, event log, web site, HDFS...
- We need to **Transform** data at the source, sink, or in a **staging area**
- We need to **Load** data into the **sink**
 - Sinks: Python, R, SQLite, RDBMS, NoSQL store, files, HDFS...



数据科学的工作流程

三个基本任务

- 获取原始数据
- 准备待分析数据
- 针对特定问题进行数据分析

数据采集
数据准备
数据分析

数据采集

	item ₁	item ₂	item ₃	...	item _n
user ₁		5	2		1
user ₂	3				
user ₃	1		3		
...					
user _{m-1}	5		4		2
user _m		4			3

ID	Name	Contact
MA-01	Hello World Tech.	534-55-7478
MA-02	ABC Technologies	283-92-8511

ID	ManufacturerID	Name
PDI-0001	M-01	Tiger T7 Bluetooth Headphones
PDI-0002	M-01	DD-027 In-Ear Headphones, Black
PDI-0003	M-02	Mr. 1022 Deep Bass Earbuds

来源：科技日报

据《新科学家》网站最新发布的消息，超过40%的昆虫物种可能在将来几十年内灭绝。其中蝴蝶、蜜蜂和蜈蚣受到的影响最大。主要原因是栖息地的丧失，这是对过去40年来所有昆虫长期调查得出的令人震惊的结论。

“这种影响对地球生态系统将是灾难性的，因为昆虫是地球上许多生态系统的基石。”论文作者说，他们来自澳大利亚悉尼大学和中国农业科学院。

研究发现，昆虫减少的最大原因是栖息地丧失；其次，寄生虫和疾病也起着重要作用，例如，瓦螨的蔓延导致蜜蜂种群的衰退；最后，气候变化似乎也有影响，热带地区的昆虫可能对温度变化的耐受性较差，其数量可能已经因全球变暖而有所下降。

数据准备

数据分析

特征				标签
...	1
...	0

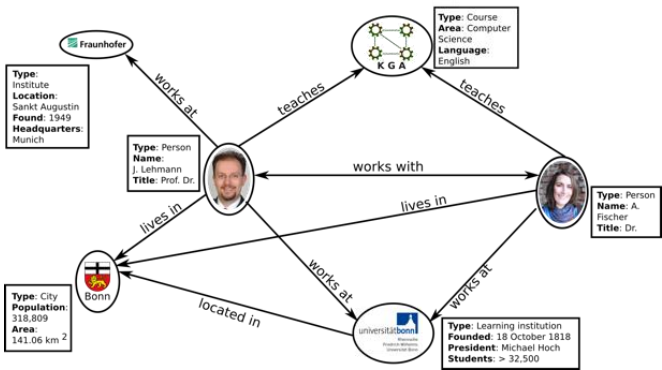
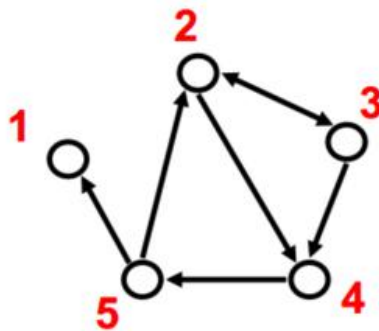
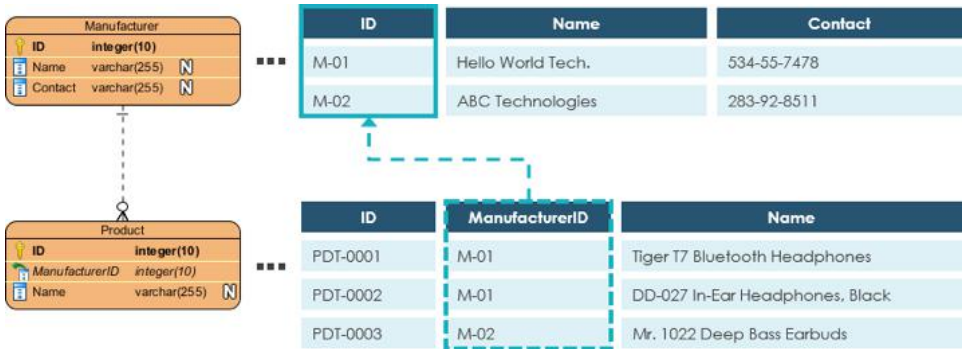
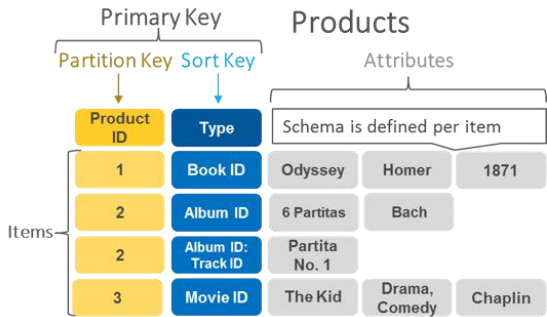
待分析数据

数据类型

□ Variety数据的种类繁多

- 数组、矩阵
- 键值对
- 实体-关系表
- 时序数据、流数据
- 图数据
- 文本数据
- 多媒体数据
- ...

	item ₁	item ₂	item ₃	...	item _n
user ₁		5	2		1
user ₂	3				
user ₃	1		3		
...					
user _{m-1}	5		4		2
user _m		4			3



来源：科技日报

据《新科学家》网站最新发布的信息，超过40%的昆虫物种可能在未来几十年内灭绝，其中蝴蝶、蜜蜂和蜚蠊受到的影响最大，主要原因是栖息地的丧失。这是对过去40年来所有昆虫长期调查得出的令人震惊的结论。

“这种影响对地球生态系统将是灾难性的，因为昆虫是世界上许多生态系统的基础。”论文作者说，他们来自澳大利亚悉尼大学和中国农业科学院。

研究发现，昆虫减少的最大原因是栖息地丧失；其次，寄生虫和疾病也起着重要作用，例如，瓦螨的蔓延导致蜜蜂种群的衰退；最后，气候变化似乎也有影响，热带地区的昆虫可能对温度变化的耐受性较差，其数量可能已经因全球变暖而有所下降。



数据模型——数组与矩阵

□ 数据项同类型，可以利用下标访问

■ 例子：NumPy的多维数组（ndarray）

■ 例子：推荐系统中的user-item矩阵

两个用户对三个商品打分：

• $u_1 \rightarrow 1(5); 3(2)$

• $u_2 \rightarrow 2(3); 3(5)$

请用NumPy构造矩阵

	商品				
	$item_1$	$item_2$	$item_3$...	$item_n$
$user_1$		5	2		1
$user_2$	3				
$user_3$	1		3		
.					
.					
.					
$user_{m-1}$	5		4		2
$user_m$		4			3

用户

评分

A. $mat =$

`np.array([[5,0,2],[0,3,5]])`

B. $mat =$

`np.array([[5,np.nan,2],[np.nan,3,5]])`

```
import numpy as np
mat = np.array( [[5,np.nan,2],[np.nan,3,5]] )
mat
array([[ 5., nan,  2.],
       [nan,  3.,  5.]])
```

数据模型——关系数据 (Relational Data)

□简单的关系数据：单表数据

- 行：表示一条记录 (Record)
- 列：表示一个属性 (Attribute)

使用pandas表示单表数据

Team	Win	Loss	Win%
Houston Rockets	20	4	0.83
Golden State Warriors	21	6	0.78
San Antonio Spurs	19	8	0.7
Minnesota Timberwolves	16	11	0.59
Denver Nuggets	14	12	0.54
Portland Trail Blazers	13	12	0.52
New Orleans Pelicans	14	13	0.52
Utah Jazz	13	14	0.48

```
nba_df = pd.DataFrame ({'Team': team_col,  
                        'Win': win_col,  
                        'Loss': loss_col})  
  
print (nba_df)
```

列标签

	Team	Win	Loss
0	Houston Rockets	20	4
1	Golden State Warriors	21	6
2	San Antonio Spurs	19	8
3	Minnesota Timberwolves	16	11
4	Denver Nuggets	14	12
5	Portland Trail Blazers	13	12
6	New Orleans Pelicans	14	13
7	Utah Jazz	13	14

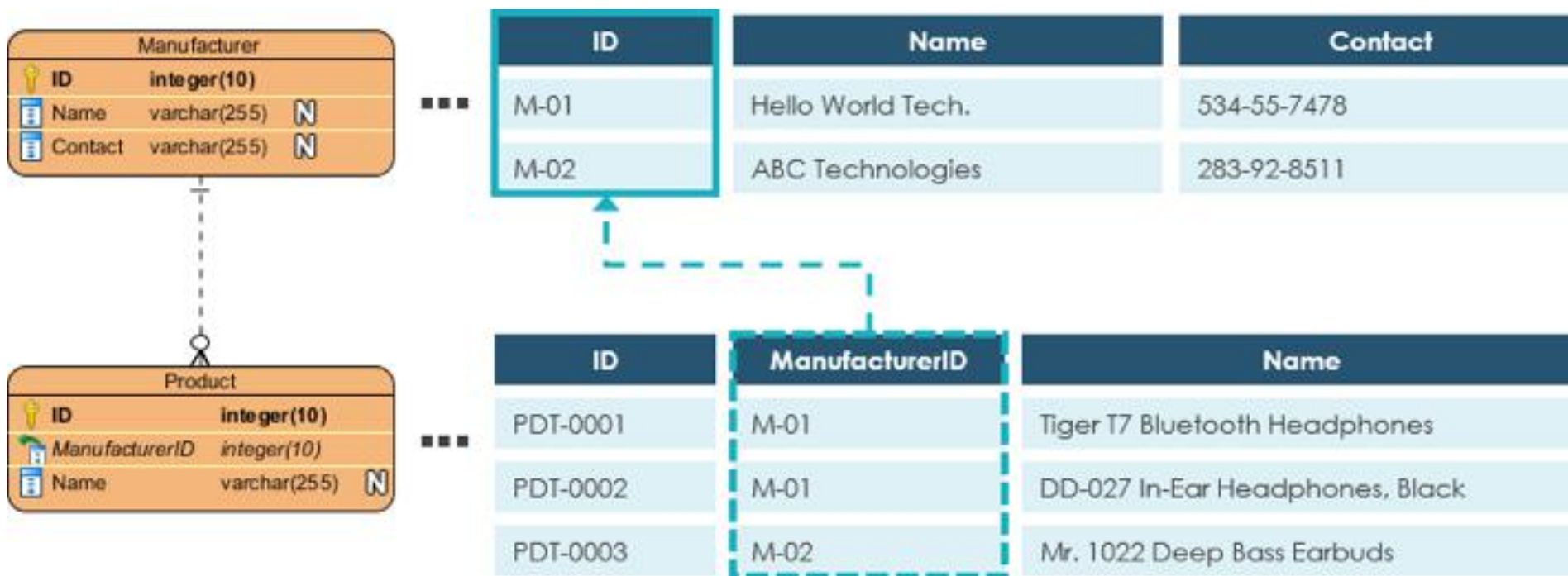
行标签

数据模型——关系数据（Relational Data）

□ 关系数据库：将数据表示为多个彼此可关联的表格

■ ER模型组织数据

■ 表格、属性、主外键



数据模型——文本数据

□ 自然语言是人们交流信息最为自然的表达方式

■ 互联网网页、论坛评论等

■ 企业文档

■ 聊天记录

来源：科技日报

据《新科学家》网站最新发布的消息，超过40%的昆虫物种可能在未来几十年内灭绝，其中蝴蝶、蜜蜂和蜉蝣受到的影响最大，主要原因是栖息地的丧失。这是对过去40年来所有昆虫长期调查得出的令人震惊的结论。

“这种影响对地球生态系统将是灾难性的，因为昆虫是世界上许多生态系统的基础。”论文作者说，他们来自澳大利亚悉尼大学和中国农业科学院。

研究发现，昆虫减少的最大原因是栖息地丧失；其次，寄生虫和疾病也起着重要作用，例如，瓦螨的蔓延导致蜜蜂种群的衰退；最后，气候变化似乎也有影响，热带地区的昆虫可能对温度变化的耐受性较差，其数量可能已经因全球变暖而有所下降。

- 非结构化，给文本分析处理带来巨大挑战
- 理解词语、实体、句子、关系等
- 自然语言的语义鸿沟

数据模型——图数据

□ 顶点一般表示实体或者属性值

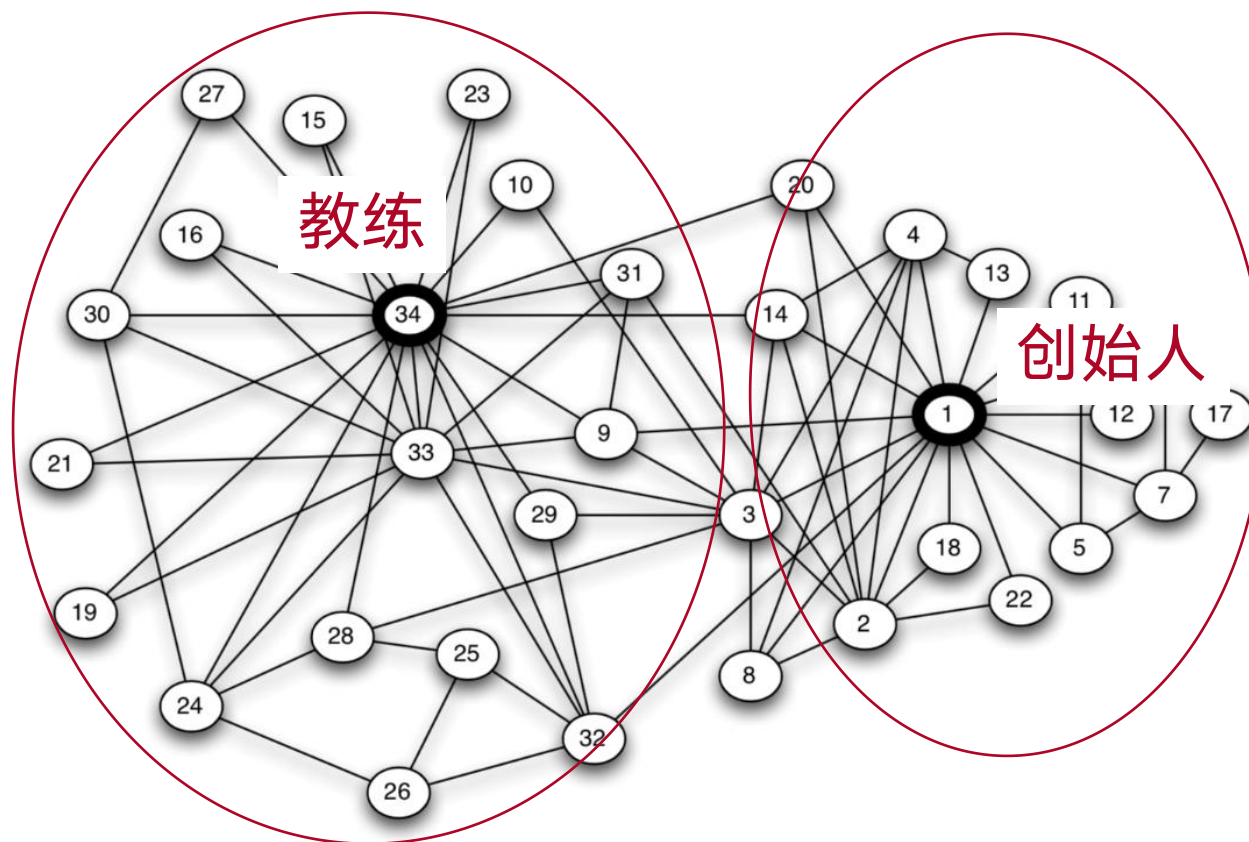
■ 顶点之间的边，表示被连接的两个顶点间的关系

■ 实例

- 社交网络
- 知识图谱

请你预言该俱乐部在不久的将来会：

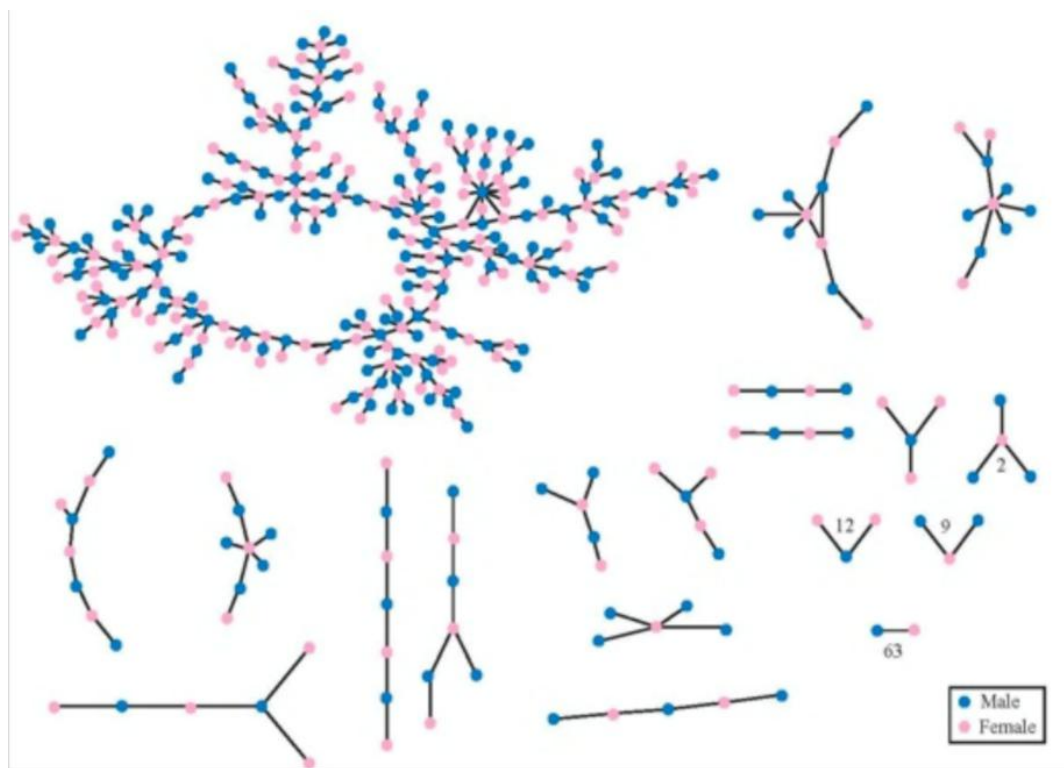
- A. 分裂为两个俱乐部
- B. 团结在创始人的周围



数据模型——图数据

□ 图数据：直观地理解群体的行为

■ 例子：高中生恋爱关系图（边代表二人在18个月内恋爱过）



Chains of Affection: The Structure of Adolescent Romantic and Sexual Networks

Peter S. Bearman
Columbia University

James Moody
Ohio State University

Katherine Stovel
University of Washington

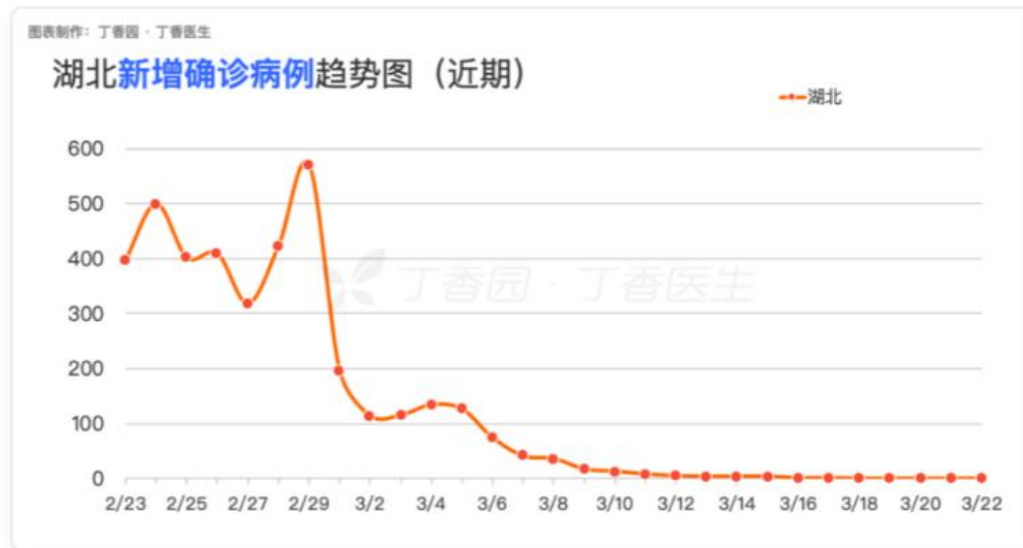
July 2004 · [American Journal of Sociology](#). 110(1)

DOI:[10.1086/386272](#)

数据模型——时序数据

□ 随时间不断变化或累积的数据

- 每个数据项有时间戳
- 关注一段时间内的数据值变化、关注异常值
- 新的数据价值更高
- 多用于监控传感等场景



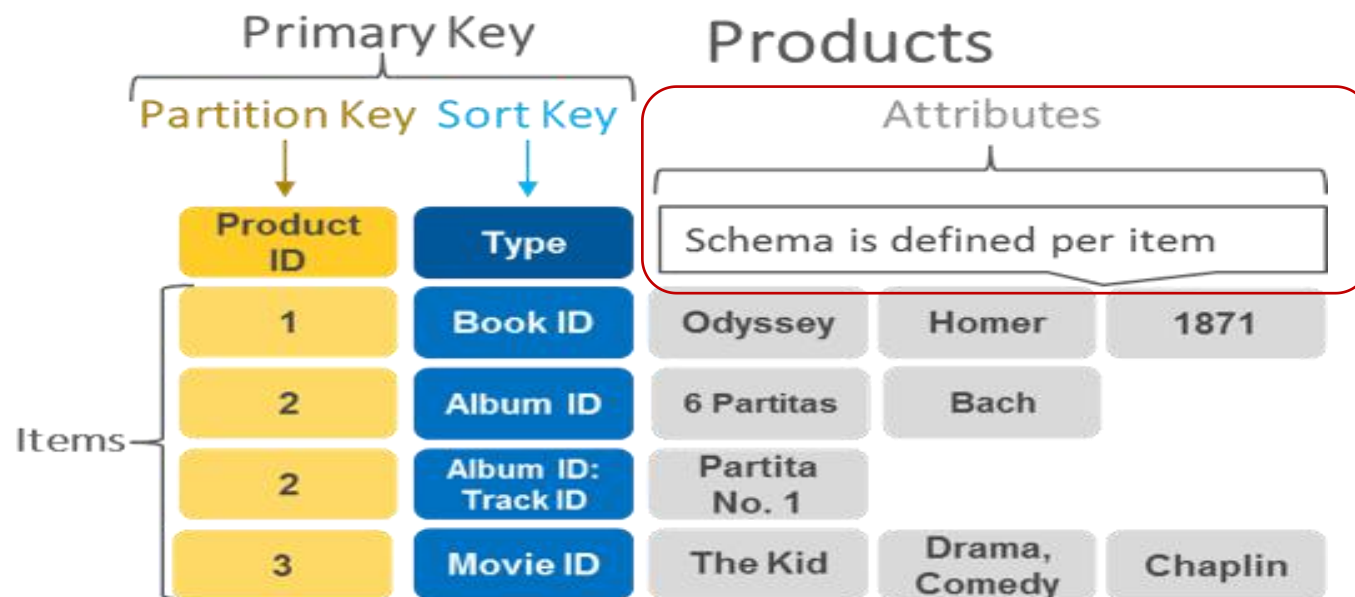
数据模型——键值对

□ 键值对灵活定义属性，每行可以有多个不同的属性

■ 例子：用户画像

■ 通过键直接访问值

■ 简单的如Hash table，Map等数据结构



数据模型——多媒体数据

- 图像、视频、音频等
 - 多种媒体类型的混合
 - 更关注语义
 - 处理复杂，计算代价高
 - 数据量相对更大
 - 在自媒体应用中普遍存在



【简介】比尔及梅琳达·盖茨基金会联席主席比尔·盖茨12日在通过新华社独家发布的视频里说，过去一年里中国在促进全球发展方面继续作出重要贡献。具体聊了哪些贡献？快戳视频看看吧！



ISSN: 1077-3142

Computer Vision and Image Understanding

Editor-in-Chief: [N. Paragios](#)

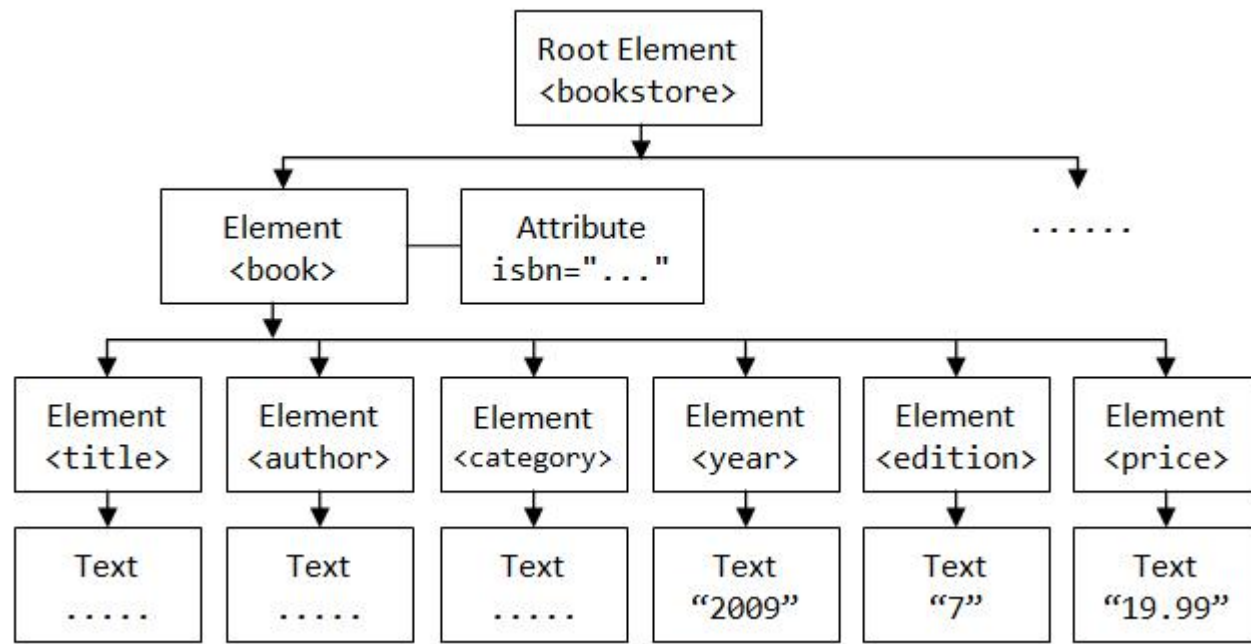
[View Editorial Board](#)

[CiteScore: 8.7](#) ^① [Impact Factor: 3.121](#) ^①

数据模型——XML and DOM

XML 是一种对 DOM（文档对象模型）进行编码的文本格式的数据结构，常用于网页。

DOM树状结构:



XML 编码数据

```
<location>  
  <latitude>37.78333</latitude>  
  <longitude>122.4167</longitude>  
</location>
```

An XML schema for this element:

...

```
<xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema"  
  elementFormDefault="unqualified">  
  <xsd:complexType name="location">  
    <xsd:sequence>  
      <xsd:element name="latitude" type="xsd:decimal"/>  
      <xsd:element name="longitude" type="xsd:decimal"/>  
    </xsd:sequence>  
  </xsd:complexType name="location">
```

Event-Driven Parsing: SAX

<?xml version="1.0" encoding="UTF-8"?>

➡ 文档头

<!-- bookstore.xml -->

➡ 注释

<bookstore>

➡ 开始元素 “bookstore”

<book ISBN="0123456001">

➡ 开始元素 “book”

<title>Java For Dummies</title>

➡ 开始元素 “title”
结束元素 “title”

<author>Tan Ah Teck</author>

<category>Programming</category>

<year>2009</year>

<edition>7</edition>

<price>19.99</price>

</book>

➡ 结束元素 “book”

JSON

JSON (Javascript Object Notation) :

```
{  
  "firstName": "John",  
  "lastName": "Smith",  
  "age": 25,  
  "address": {  
    "streetAddress": "21 2nd Street",  
    "city": "New York",  
    "state": "NY",  
    "postalCode": "10021-3100" },  
  "phoneNumbers": [  
    { "type": "home",  
      "number": "212 555-1234" },  
    { "type": "office",  
      "number": "646 555-4567" } ],  
  "children": [],  
  "spouse": null  
}
```

数据模型

□ 大数据时代：多模态数据并存

■ 以关系数据为代表的结构化数据

- 数据量占比低于20%
- 数据价值相对高

■ 以文本、图数据为代表的非结构化数据

- 数据量占比高于80%
- 数据价值相对低

■ 需要融合结构化数据和非结构化数据

- 信息抽取
- 实体链接与数据融合

数据模型

□数据模型小结

- 不同类型的数据与数据模型
- 人们如何理解与表达数据
- 计算机如何存储与处理数据

关系数据库里使用的 数据模型三要素

数据结构：描述数据有由什么元素构成，是什么类型，有什么关系等

数据操作：可以施加于数据对象的操作以及相关规则

数据完整性约束条件：指在给定的数据模型中，数据及其联系所遵守的一组通用的规则，以保证数据的正确性和一致性