数据库系统概论新技术篇

大数据概述

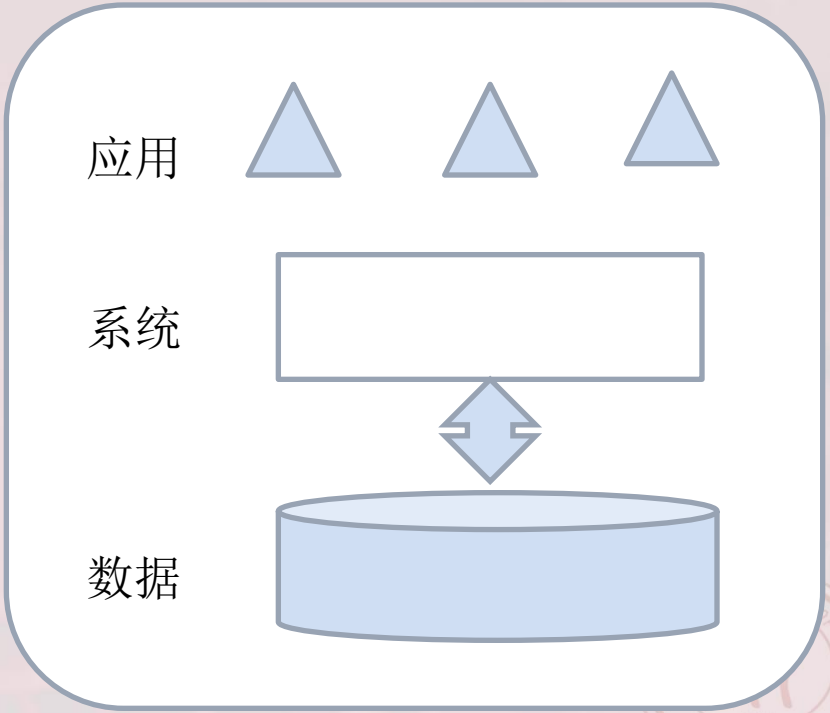
杜小勇

中国人民大学信息学院

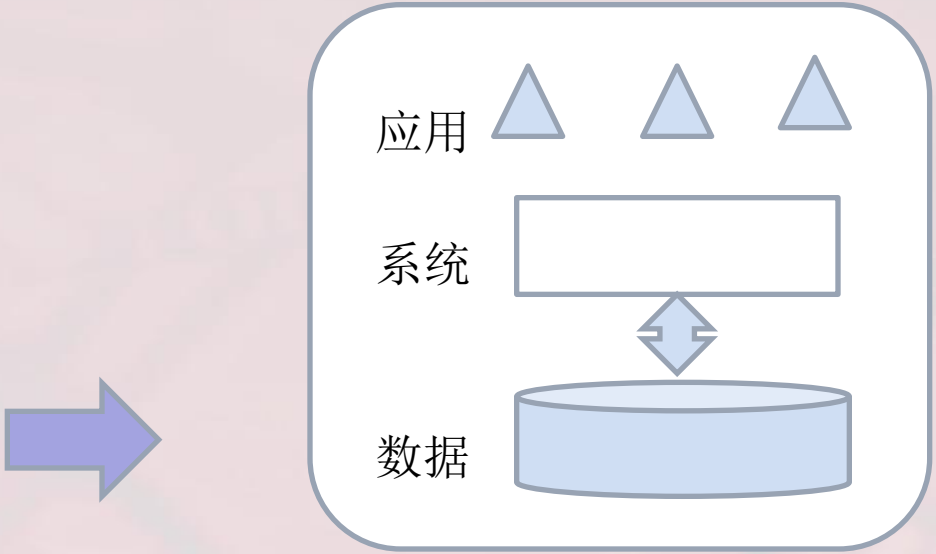
2017年6月

目录

- ❖1 大数据的数据特征
- ❖2 大数据的系统特征
- ❖3 大数据的应用特征

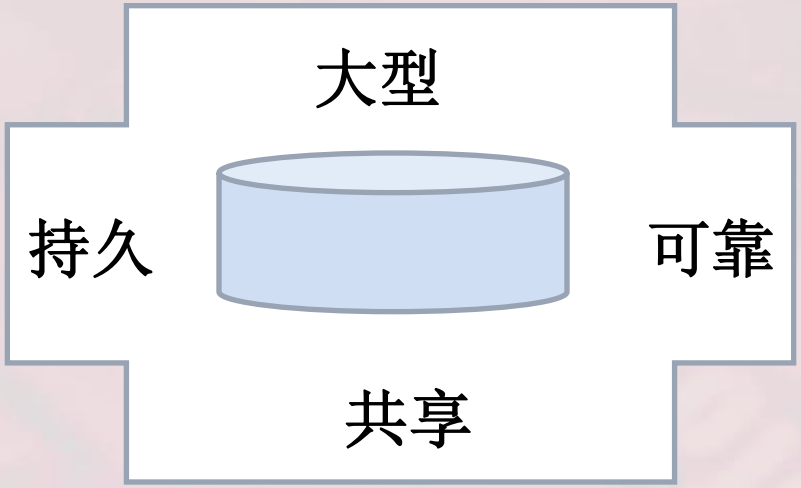


1、大数据的数据特征



关系数据库的定义

- 是大型、共享、持久、可靠数据的集合



大数据的数据特征

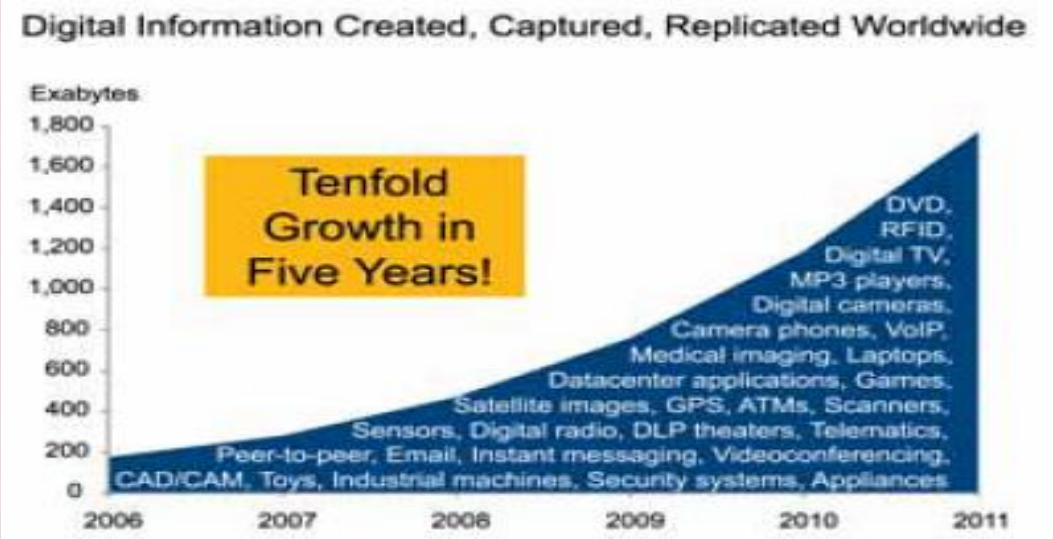
容量大（Volume）

- 观点：“大”是相对的，是和当时的计算机处理能力相关的，超过了现有技术的能力。
- 但是，“大规模”又是大数据的基本要求。
  - 80年代，百万条记录就是VERY LARGE DATA
  - 00年代，TB级别就是DATA INTENSIVE 应用
  - 10年代，100T以上，甚至PB级才能够算得上是大数据

新摩尔定理

- 从现在起，每18个月,新增的存储量等于有史以来存储量之和!

——1998年图灵奖获得者Jim Gray



多样性(Variety)

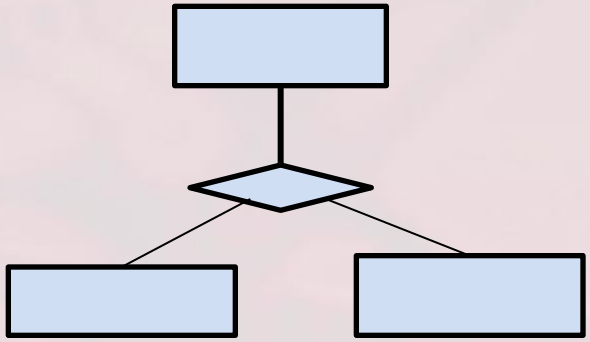
- 数据种类的多样性：文字、语音、图片、视频等，不再是单一的“关系”数据了；
- 数据来源的多样性：同一个对象的数据来自不同的数据源，数据需要集成；

电商平台

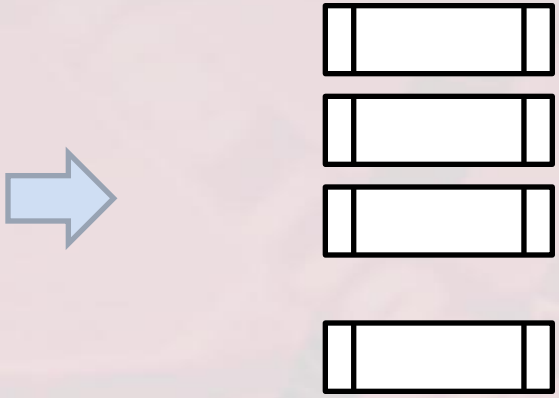
- 商品广告平台
- 购物交易平台
- 支付平台
- 社交平台
- 。。。。

比较：关系数据库

- 世界上的一切对象和关系，都用单一的概念“关系”来建模！



概念世界模型



“实体”关系

“关系”关系

关系世界模型

变化快(Velocity)

- 数据快速增长。这是原来所没有或者说强调不够的特征。
  - 数据到达或者产生的速度太快，对系统处理造成巨大的压力。
  - 例如，入库速度要求：**100GB/S**。

2016年“双十一”的天猫交易额

时间	交易额	描述
0时0分52秒	超过10亿	一分钟超10亿的交易额
0时14分16秒	超191亿	超过**2012年**双十一全天成交额
1时	突破353亿	超过**2013年**双十一全天成交额
6时54分53秒	超571亿	超越**2014年**双十一全天成交额
15时19分13秒	912亿	超越**2015年**双十一全天交易额
24时	超1207亿	交易额翻了一番
		无线交易额占比81.87%，覆盖235个国家和地区

- “现象级”应用：在某一个时期，对系统的压力突然暴增，极易导致系统的崩溃。

质量弱(Veracity)

- 大数据天然就带有噪音。由于进入系统的数据缺乏控制，数据质量不高。
- 如何处理弱质的数据？从中获得有用的信息，是大数据处理需要面对的挑战。

问题数据的存在是常态

- 不完整数据（incomplete）
- 不正确数据（incorrectness）
- 不一致数据（inconsistency）
- 不精确数据（unprecision）

小结

关系数据库

- 大型
- 共享
- 持久
- 可靠

大数据系统

- 海量
- 多类型
- 快速变化
- 弱质量