发展历程分析

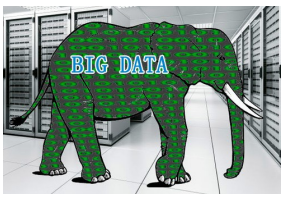
- 大型机时代：Many people One computer
- 小型机时代：One people One computer
- 云计算时代：Many people One Cloud

大型机→小型机→云计算→大数据→AI → ？

大数据的定义

大数据是 …的决策力、洞察发现力和 …率和多样化的信息资产

大数据到底是什么？
大数据有什么特点？



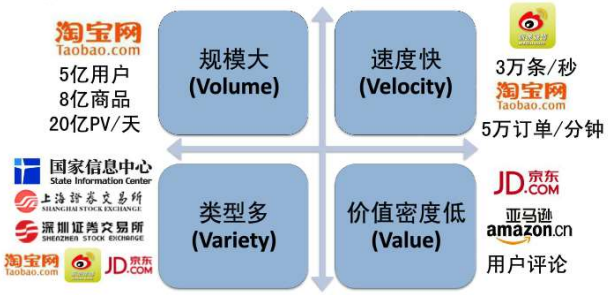
Big?

1Byte = 8 Bit
1KB = 1,024 Bytes
1MB = 1,024 KB = 1,048,576 Bytes
1GB = 1,024 MB = 1,048,576 KB = 1,073,741,824 Bytes
1TB = 1,024 GB = 1,048,576 MB = 1,099,511,627,776 Bytes
1PB = 1,024 TB = 1,048,576 GB =1,125,899,906,842,624 Bytes
1EB = 1,024 PB = 1,048,576 TB = 1,152,921,504,606,846,976 Bytes
1ZB = 1,024 EB = 1,180,591,620,717,411,303,424 Bytes
1YB = 1,024 ZB = 1,208,925,819,614,629,174,706,176 Bytes

大数据定义及特点

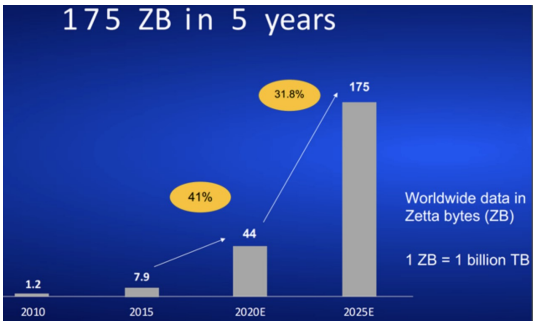
WIKIPEDIA The Free Encyclopedia

大数据是通过传统数据库技术和数据处理工具不能处理的庞大而复杂的数据集合。



大量（Volume）

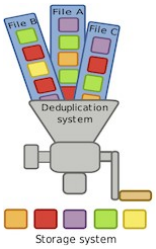
- 2010年，全球数据量已达1.2ZB，到2020年将暴增30倍达35ZB！



大量（Volume）

- 解决大量的数据存储问题：
  - 新型的大容量磁盘；
  - 分布式存储技术；
  - 数据精简技术（压缩和重删技术）；

Seagate Enterprise ST8000NM0065 8 TB 3.5" Internal Hard Drive
by Seagate
$507.78
More Buying Choices
$507.53 new (16 offers)
$425.91 used (1 offer)



快速化（Velocity）

➢从数据的生成到消耗，时间窗口非常小！

✓数据产生的速度很快！

✓数据处理的速度要求很快！

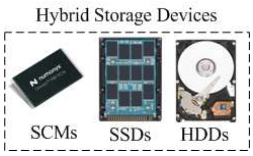
- 每秒钟，人们发送290万封电子邮件
- 每分钟，人们向Youtube上传60个小时的视频
- 每一天，人们在Twitter上发消息1.9亿条微博
- 每一天，人们在Twitter上发出3.44亿条消息
- 每一天，人们在Facebook发出40亿条信息

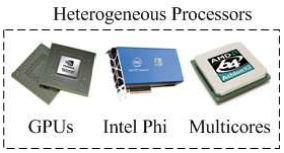
快速化（Velocity）

➢提高数据存储和处理的速度：

✓新型的数据存储器件；

✓多核处理器和GPU并行处理技术！

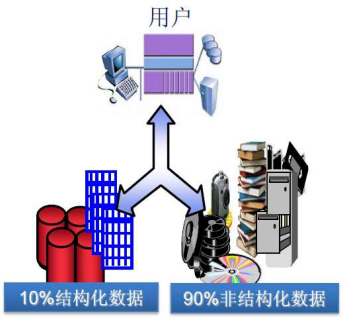




多样性（Variety）

➢数据来源多样：拍摄、语音、点击、传感器等；

➢数据格式多样：邮件、语音、图片、视频等！



多样性（Variety）

SQL：
结构化存储，固定Schema
索引
标准化查询语言
ACID
扩展性弱

NoSQL：
Schema不固定，可以动态改变
没有固定查询语言
BASE （Basically Available， Soft State，Eventually Consistency）
最终一致性
可以扩展到很大规模
高容错性

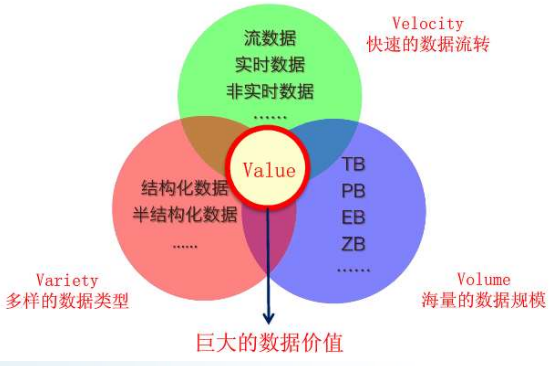
Not Only SQL

价值（Value）

➢浪里淘沙却又弥足珍贵！

➢大数据助力商业模式和应用创新！

大数据的性质（4V）



4 “V”s

大数据

- 大数据 4V
  - 大量（Volume）
    - 存储大；
    - 计算量大；
  - 多样（Variety）
    - 来源多；
    - 格式多；
  - 快速（Velocity）
    - 增长速度快
    - 处理速度要求快
  - 价值（Value）
    - 浪里淘沙却又弥足珍贵

数据没有办法在可容忍的时间下使用常规软件方法完成存储、管理和处理任务

相关技术

大数据相关技术

- 分析技术
  - 数据处理：自然语言处理技术
  - 统计和分析：A/B test; top N排行榜；地域占比；文本情感分析
  - 数据挖掘：关联规则分析；分类；聚类
  - 模型预测：预测模型；机器学习；建模仿真
- 大数据技术
  - 数据采集：ETL工具
  - 数据存取：关系数据库；NoSQL；SQL等
  - 基础架构支持：云存储；分布式文件系统等
  - 计算结果展现：云计算；标签云；关系图等

相关技术

大数据相关技术

- 存储
  - 结构化数据：
    - 海量数据的查询、统计、更新等操作效率低
  - 非结构化数据
    - 图片、视频、word、pdf、ppt等文件存储
    - 不利于检索、查询和存储
  - 半结构化数据
    - 转换为结构化存储
    - 按照非结构化存储
- 存储问题解决方案
  - 在CAP理论指导下数据库技术适当“退化”
    - NoSQL技术： HDFS, HBASE, OceanBase, MongoDB等

相关技术

大数据相关技术

- 计算
  - 因结构变化为导致计算模式变更
  - 需求模式变化带来的计算碰到瓶颈
- 解决方案
  - Hadoop（MapReduce技术）
  - 流计算（twitter的storm和yahoo！的S4）

数据来源

数据来源

- 互联网企业：SNS、微博、视频网站、电子商务网站
- 物联网、移动设备、终端中的商品、个人位置、传感器采集的数据
- 联通、移动、电信等通信和互联网运营商
- 天文望远镜拍摄的图像、视频数据、气象学里面的卫星云图数据等

大数据来“缘”与影响

来“缘”及发展影响

- 来“缘”
  - 互联网大发展，特别是社交化网络的出现
  - 信息化工作效果的积累
  - 信息社会的基础设施建设积累
- 影响
  - 传统企业与互联网进行融合
  - 对大数据进行精准化分析和挖掘，大势所趋

大数据与云计算

大数据

- 大数据与云计算
  - 云计算的模式是业务模式，本质是数据处理技术。（肉体+灵魂）
  - 数据是资产，云为数据资产提供存储、访问和计算。
  - 盘活资产，使其为国家治理、企业决策、个人生活服务，是大数据核心议题，也是云计算的最终方向
- 海量数据： 两个V（volume和value）