北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

大数据概述

Big Data Overview

大数据处理技术

计算机学院

课程提纲

- 大数据时代
- 大数据概念、影响、应用
- 大数据关键技术
- 大数据计算模式
- 大数据与云计算、物联网的关系

大数据时代

Every minute, users send 31.25 million messages and watch 2.77 million videos on Facebook

Walmart handles more than 1 million customer transactions every hour

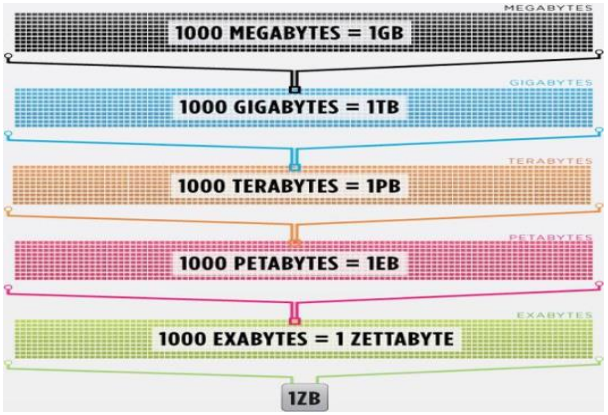
40,000 search queries are performed on Google per second, i.e. 3.46 million searches a day

300 hours of video are uploaded every minute on YouTube

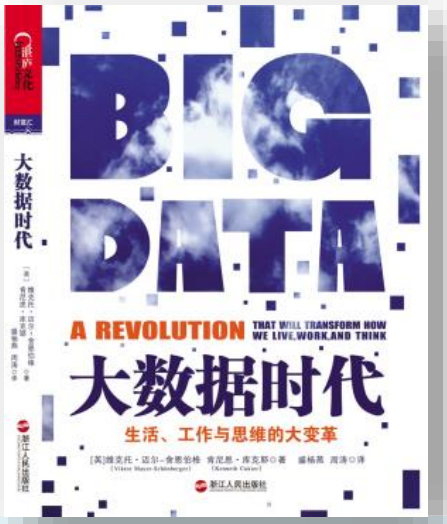
IDC reports that by 2025, real time data will be more than a quarter of all the data

55 billion messages and 4.5 billion photos are sent each day on WhatsApp

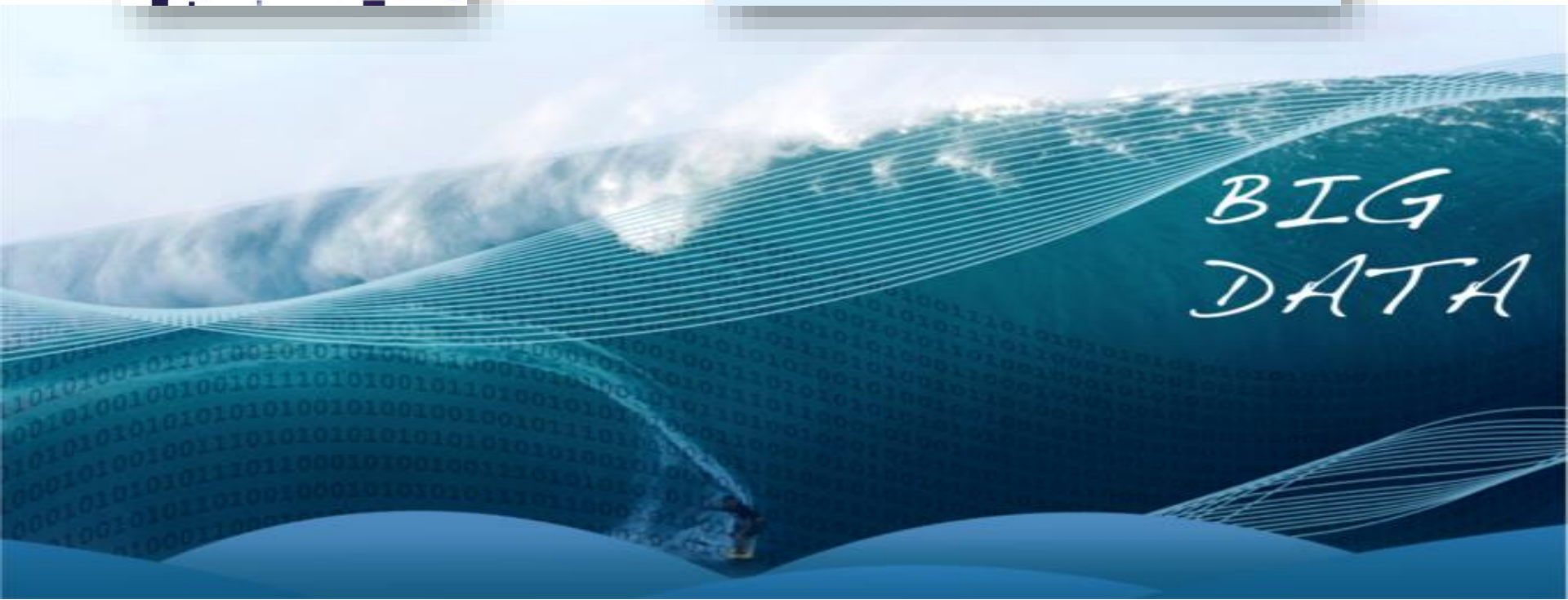
By 2025, the volume of digital data will increase to 163 zettabytes



大数据时代

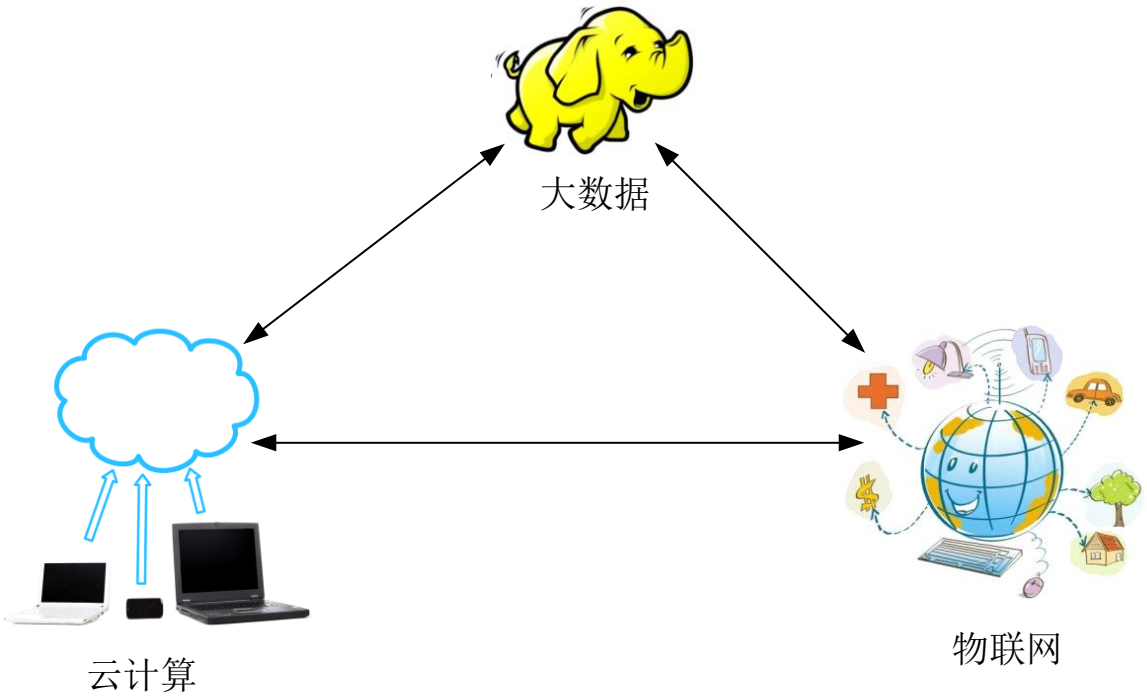

BIG DATA
大数据时代
A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE,WORK,AND THINK
大数据时代
生活、工作与思维的大变革


BIG DATA


BIG
DATA

大数据时代

- 2010年前后,以云计算、大数据、物联网的普及为标志迎来第三次信息化浪潮



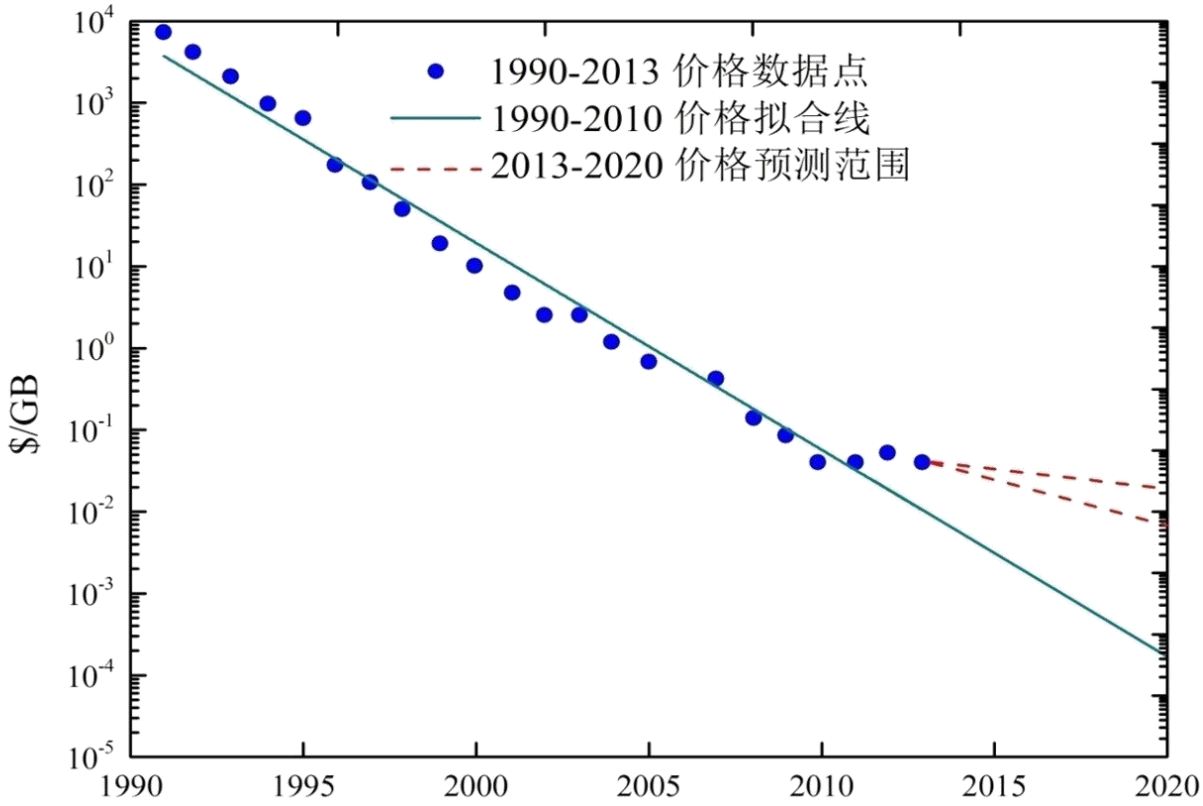
大数据时代

- 根据IBM前首席执行官Louis Gerstner的观点，IT领域每隔十五年就会迎来一次重大变革

信息化浪潮	发生时间	标志	解决问题	代表企业
第一次浪潮	1980年前后	个人计算机	信息处理	Intel、AMD、IBM、苹果、微软、联想、戴尔、惠普等
第二次浪潮	1995年前后	互联网	信息传输	雅虎、谷歌、阿里巴巴、百度、腾讯等
第三次浪潮	2010年前后	物联网、云计算和大数据	信息爆炸	将涌现出一批新的市场标杆企业

大数据时代

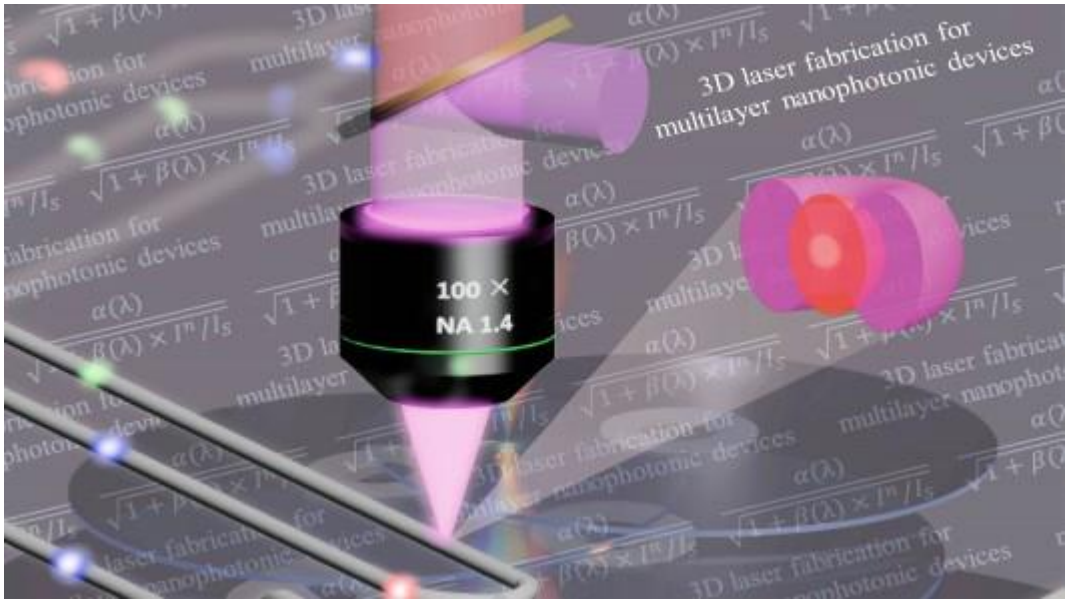
- 信息科技为大数据时代提供技术支撑
  - 存储设备容量不断增加



存储价格随时间变化情况

大数据时代

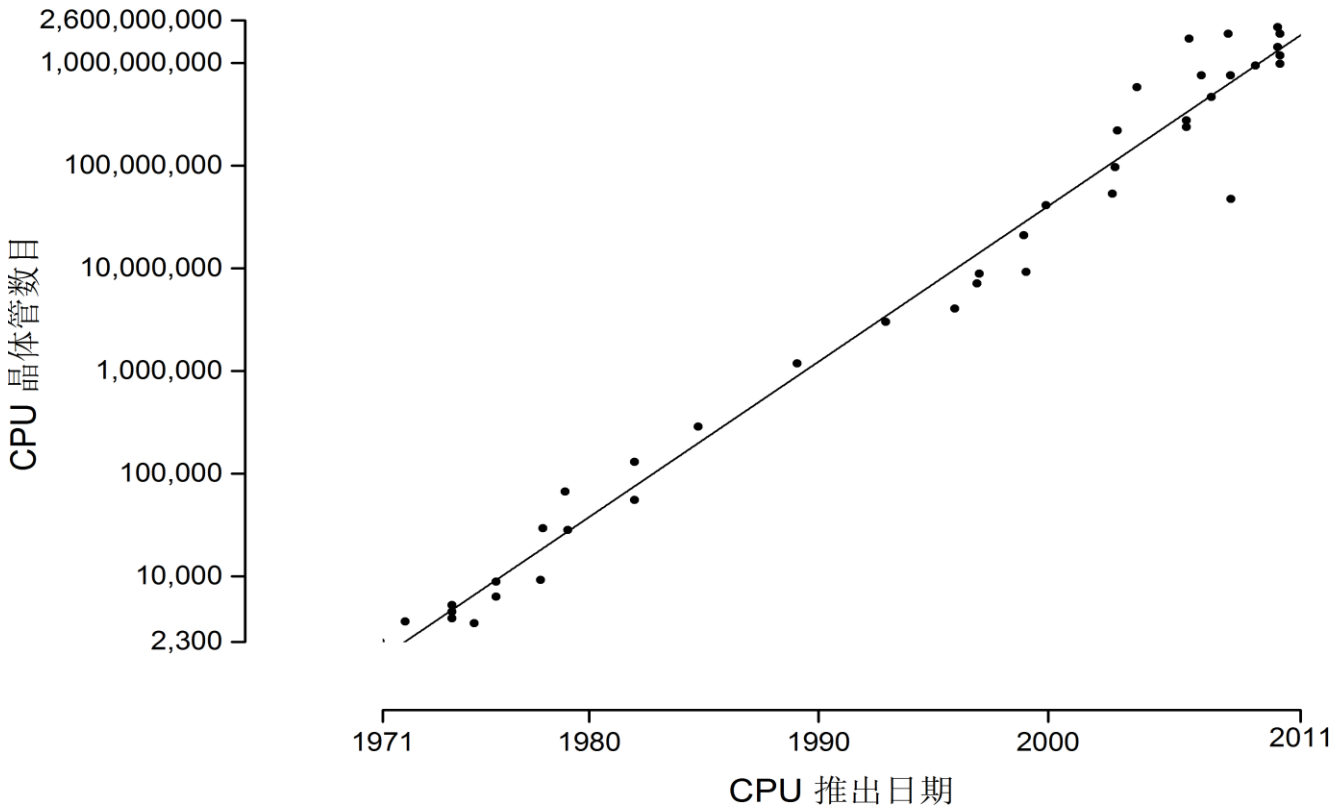
- 信息科技为大数据时代提供技术支撑
  - 存储设备容量不断增加



来自斯威本科技大学（Swinburne University of Technology）的研究团队，在2013年6月29日刊出的《自然通讯（Nature Communications）》杂志的文章中，描述了一种全新的数据存储方式，可将1PB（1024TB）的数据存储到一张仅DVD大小的聚合物碟片上。

大数据时代

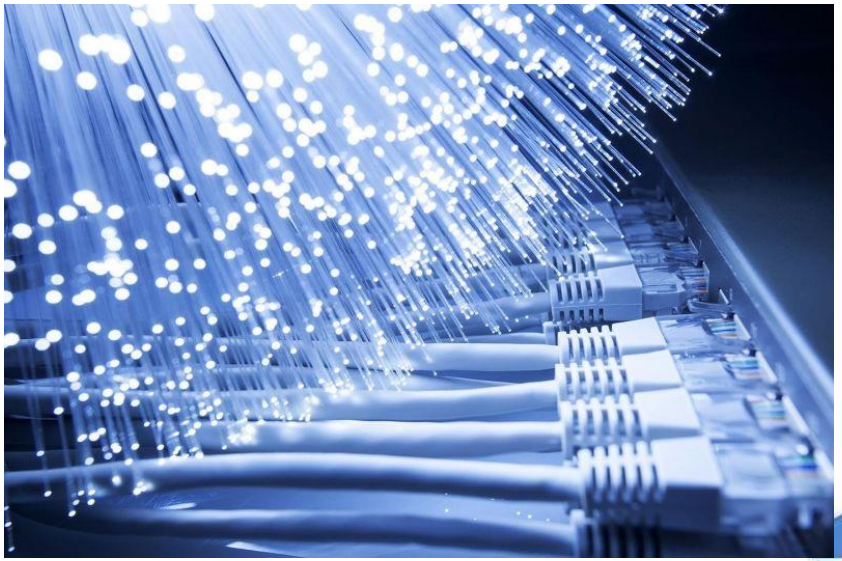
- 信息科技为大数据时代提供技术支撑
  - CPU处理能力大幅提升

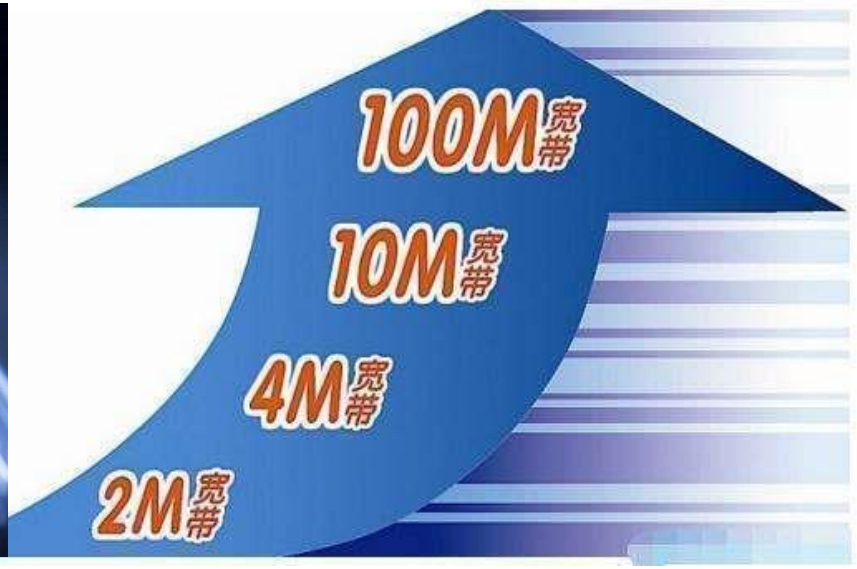


CPU晶体管数目随时间变化情况

大数据时代

- 信息科技为大数据时代提供技术支撑
  - 网络带宽不断增加





大数据时代

- 数据产生方式的变革促成大数据时代的来临

1 运营式系统阶段

•数据库的出现使得数据管理的复杂度大大降低，数据往往伴随着一定的运营活动而产生并记录在数据库中，数据的产生方式是被动的

2 用户原创内容阶段

•数据爆发产生于 Web 2.0 时代，而 Web 2.0 的最重要标志就是用户原创内容
•智能手机等移动设备加速内容产生
•数据产生方式是主动的

3 感知式系统阶段

•感知式系统的广泛使用
•人类社会数据量第三次大的飞跃最终导致了大数据的产生
•物联网中的感知设备每时每刻产生大量数据

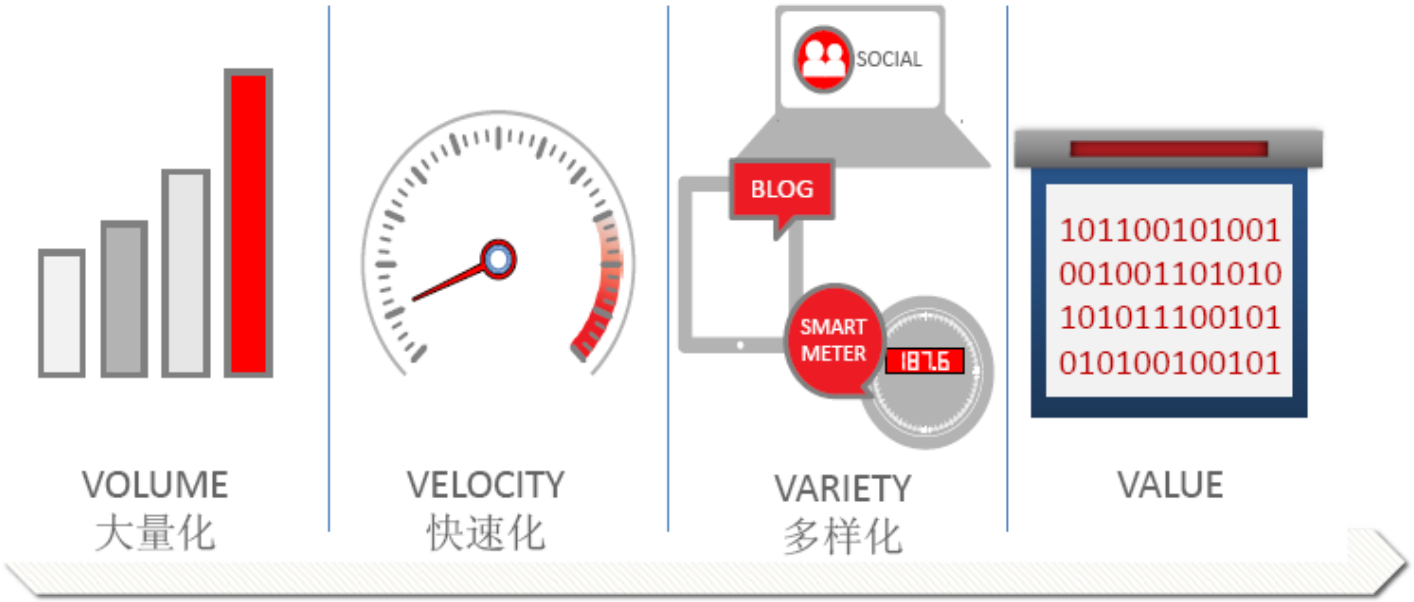
大数据时代

- 大数据的发展历程

阶段	时间	内容
萌芽期	上世纪90年代至本世纪初	随着数据挖掘理论和数据库技术的逐步成熟，一批商业智能工具和知识管理技术开始被应用，如数据仓库、专家系统、知识管理系统等。
成熟期	本世纪前十年	Web2.0应用迅猛发展，非结构化数据大量产生，传统处理方法难以应对，带动了大数据技术的快速突破，大数据解决方案逐渐走向成熟，形成了并行计算与分布式系统两大核心技术，谷歌的GFS和MapReduce等大数据技术受到追捧，Hadoop平台开始大行其道
大规模应用期	2010年以后	大数据应用渗透各行各业，数据驱动决策，信息社会智能化程度大幅提高

大数据概念、影响、应用

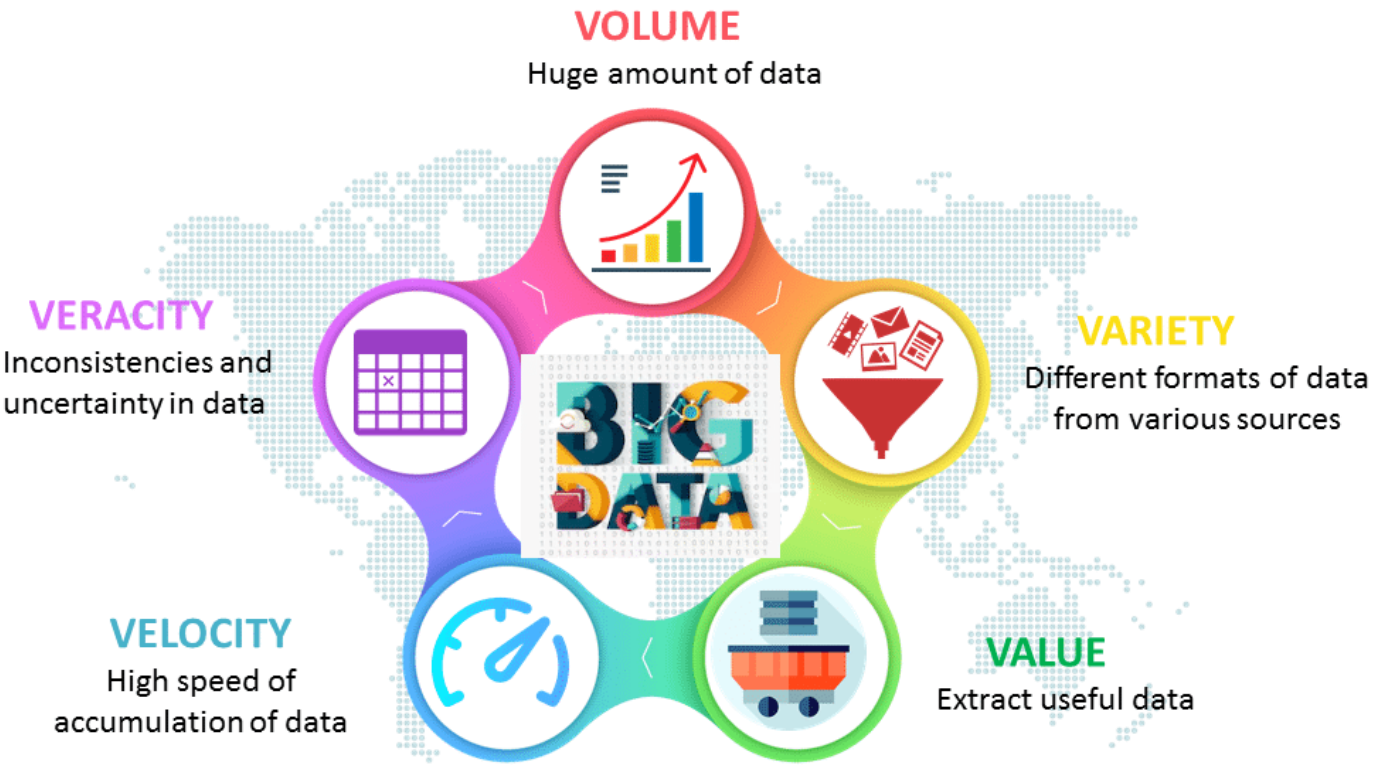
- 大数据的特征 (V)



“大量化(Volume)、多样化(Variety)、快速化(Velocity)、价值密度低(Value)”就是“大数据”的显著特征，或者说，只有具备这些特点的数据，才是大数据。

大数据概念、影响、应用

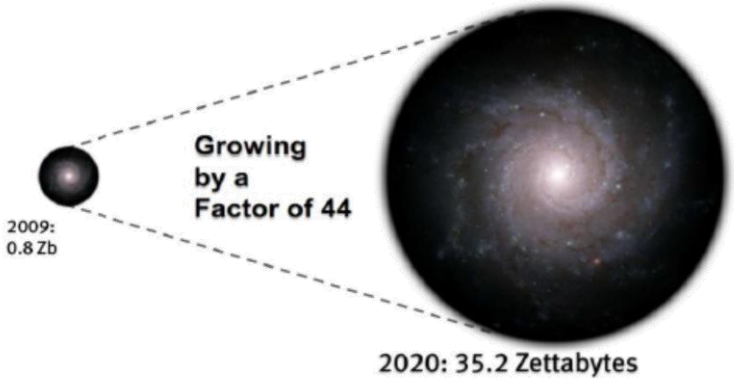
- 大数据的特征 (V)

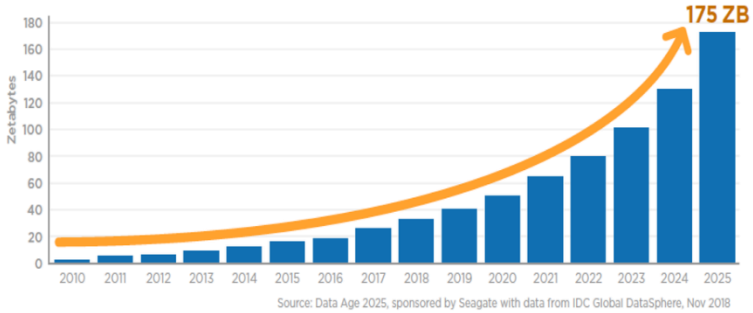


大数据概念、影响、应用

- 数据量大 (Volume)

  - 根据IDC作出的估测，数据一直都在以每年50%的速度增长，也就是说每两年就增长一倍（大数据摩尔定律）
  - 人类在最近两年产生的数据量相当于之前产生的全部数据量
  - 预计到2020年，全球将总共拥有35ZB的数据量，相较于2010年，数据量将增长近30倍

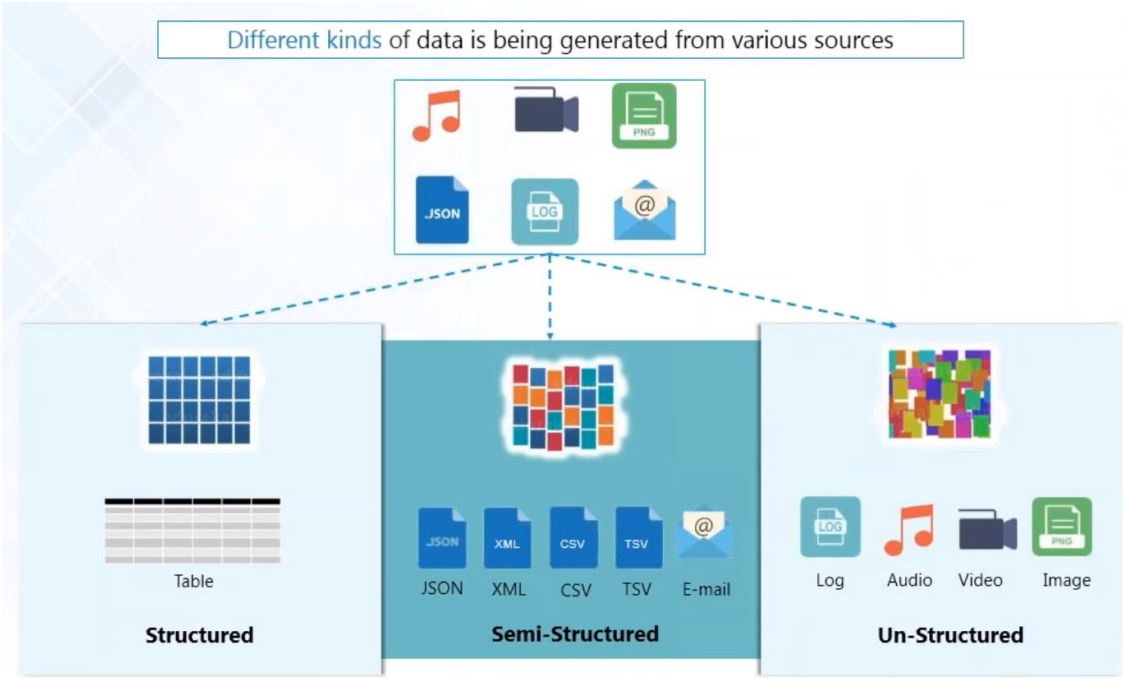




TERABYTE	10 的 12 次方	一块 1TB 硬盘		200,000 照片或 mp3 歌曲
PETABYTE	10 的 15 次方	两个数据中心机柜		16 个 Blackblaze pod 存储单元
EXABYTE	10 的 18 次方	2,000 个机柜		占据一个街区的 4 层数据中心
ZETTABYTE	10 的 21 次方	1000 个数据中心		纽约曼哈顿的 1/5 区域
YOTTABYTE	10 的 24 次方	一百万个数据中心		特拉华州和罗德岛州

大数据概念、影响、应用

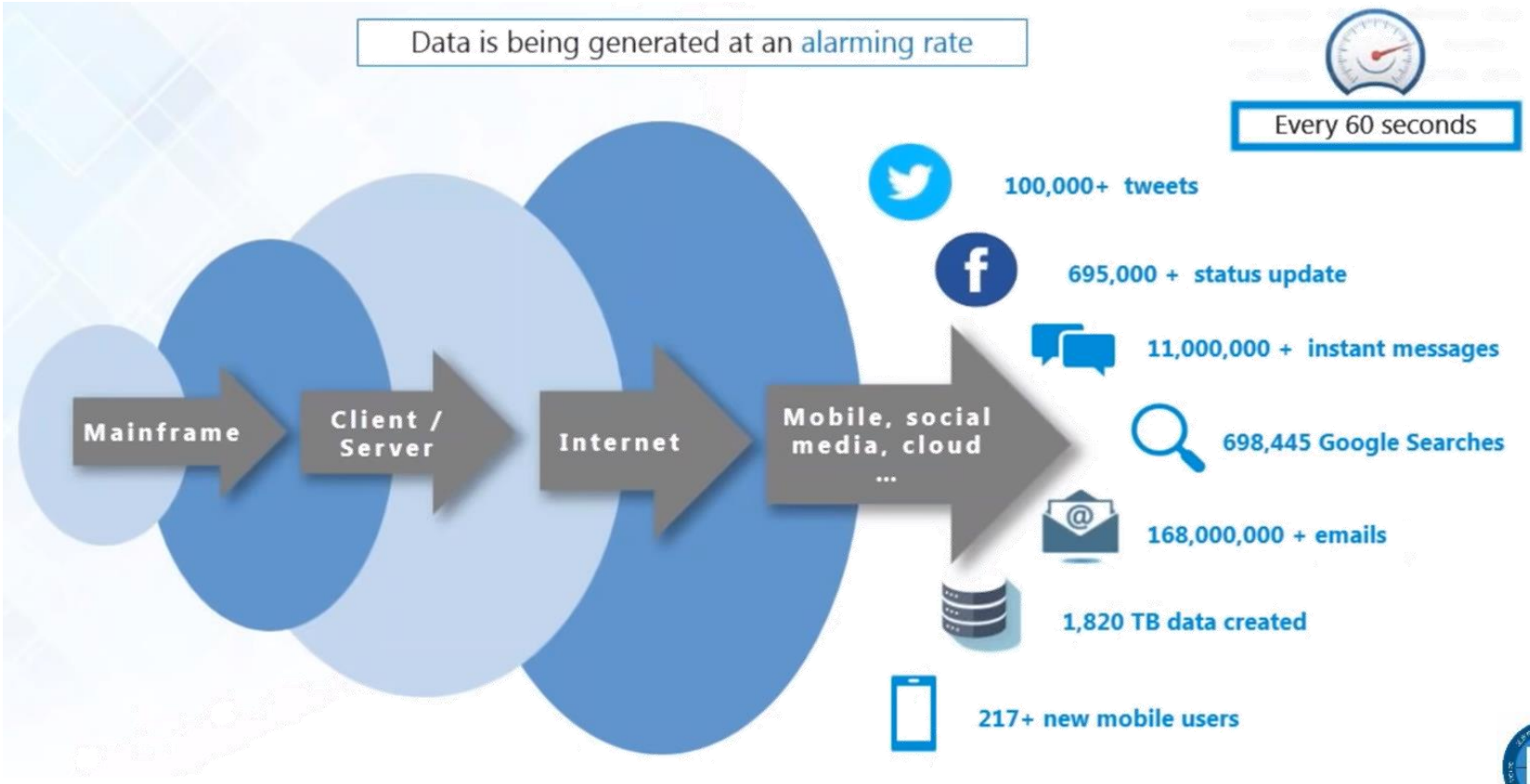
- 数据类型繁多 (Variety)
  - 大数据是由结构化和半/非结构化数据组成的
    - 10%的结构化数据，存储在数据库中
    - 90%的半/非结构化数据，它们与人类信息密切相关



- 科学研究
  - –基因组
  - –LHC 加速器
  - –地球与空间探测
- 企业应用
  - –Email、文档、文件
  - –应用日志
  - –交易记录
- Web 1.0数据
  - –文本
  - –图像
  - –视频
- Web 2.0数据
  - –查询日志/点击流
  - –Twitter/ Blog / SNS
  - –Wiki

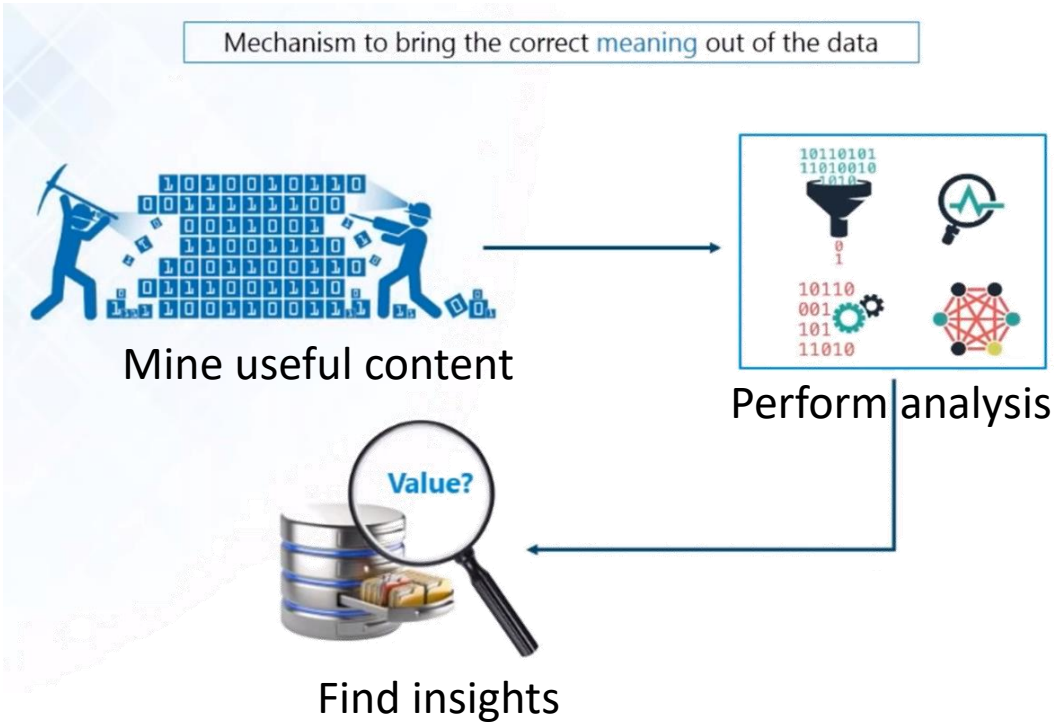
大数据概念、影响、应用

- 处理速度快 (Velocity)
  - 从数据的生成到消耗，时间窗口非常小，可用于生成决策的时间非常少
  - 1秒定律（秒级定律）：和传统的数据挖掘技术有着本质的不同



大数据概念、影响、应用

- 价值密度低 (Value)
  - 大数据虽然拥有海量的信息，但是真正可用的数据可能只有很小一部分
  - 以视频为例，连续不间断监控过程中，可能有用的数据仅仅有一两秒，但是具有很高的商业价值



大数据概念、影响、应用

- 真实性/准确性 (Veracity)
  - Veracity关注数据的质量
  - 指数据的可信赖性。我们可以信赖这些数据代表的事实吗？

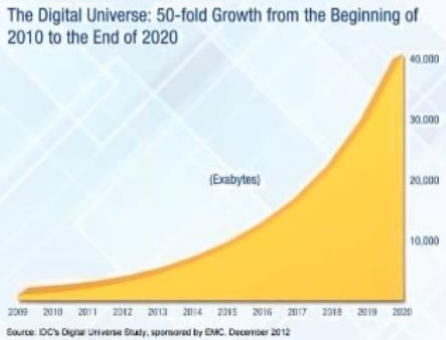
Uncertainty and inconsistencies in the data

Min	Max	Mean	SD
4.3	?	5.84	0.83
2.0	4.4	3.05	50000000
15000	7.9	1.20	0.43
0.1	2.5	?	0.76

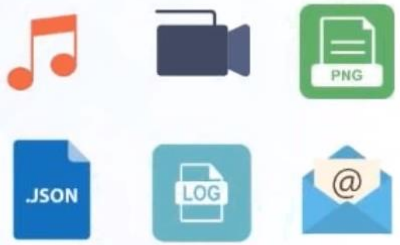
Mining is not useful if the data is messy and poor in quality, and is hard to analyze messy data

大数据概念、影响、应用

- **V's associated with Big Data may grow with time…**



Volume



Different kinds of data is being generated from various sources

Variety

Data is being generated at an alarming rate

Velocity

Mechanism to bring the correct meaning out of the data

Value

Min	Max	Mean	SD
4.3	?	5.84	0.83
2.0	4.4	3.05	50000000
15000	7.9	1.20	0.43
0.1	2.5	?	0.76

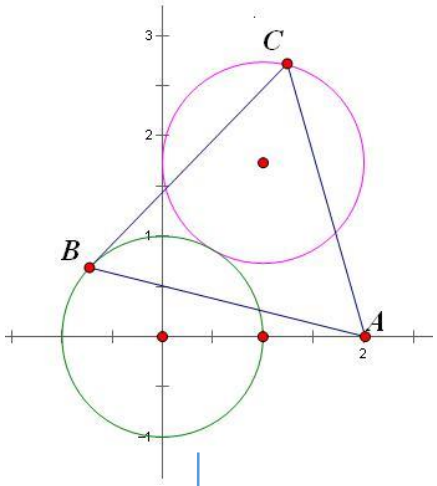
Uncertainty and inconsistencies in the data

Veracity

- Validity: correctness of data
- Variability: dynamic behavior
- Volatility: tendency to change in time
- Vulnerability: vulnerable to breach or attacks
- Visualization: visualizing meaningful usage of data

大数据概念、影响、应用

- 图灵奖获得者、著名数据库专家Jim Gray 博士观察并总结在科学研究上，先后历经了实验科学、理论科学、计算科学和数据密集型科学四种范式



实验　　理论　　计算　　数据

大数据概念、影响、应用

- 在思维方式方面，大数据完全颠覆了传统的思维方式：
  - 全样而非抽样
  - 效率而非精确
  - 相关而非因果

作　者
[奥地利] 维克托·迈尔-舍恩伯格

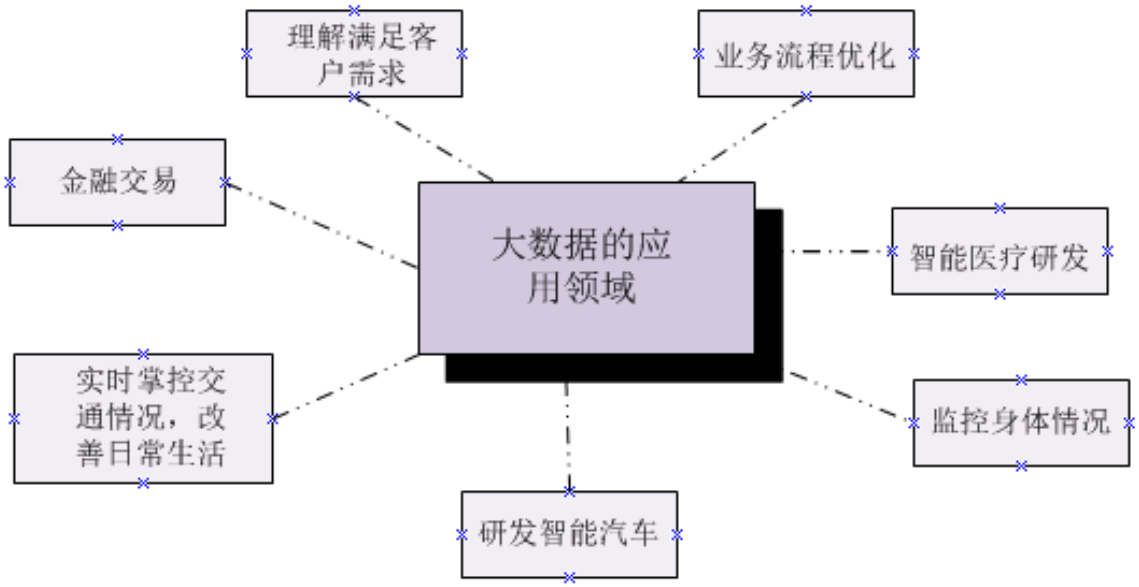
原版名称
Big Data:A Revolution That Will Transform How We Live, Work, and Think

大数据概念、影响、应用

- 在社会发展方面，大数据决策逐渐成为一种新的决策方式，大数据应用有力促进了信息技术与各行业的深度融合，大数据开发大大推动了新技术和新应用的不断涌现
- 在就业市场方面，大数据的兴起使得数据科学家成为热门职业
- 在人才培养方面，大数据的兴起，将在很大程度上改变中国高校信息技术相关专业的现有教学和科研体制

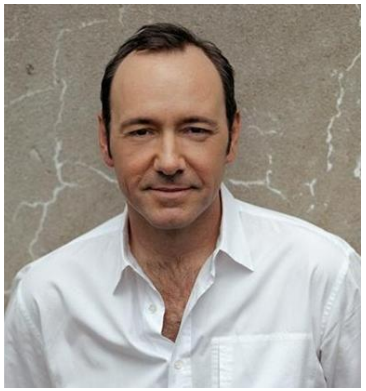
大数据概念、影响、应用

- 大数据无处不在，包括金融、汽车、零售、餐饮、电信、能源、政务、医疗、体育、娱乐等在内的社会各行各业都已经融入了大数据的印迹



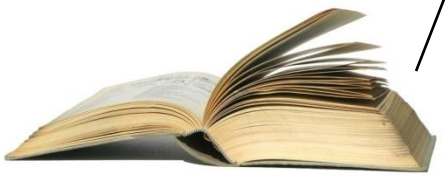
大数据概念、影响、应用

大数据概念、影响、应用

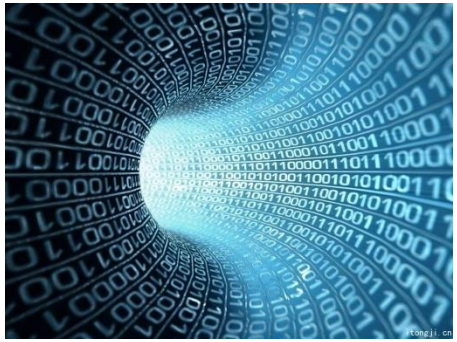


Kevin Spacey

David Fincher



英国同名小说《纸牌屋》

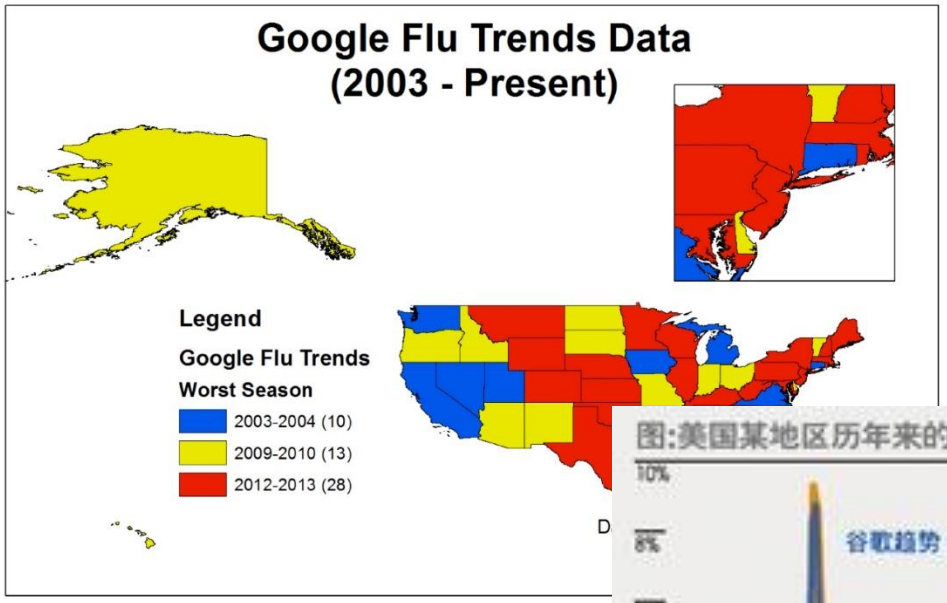


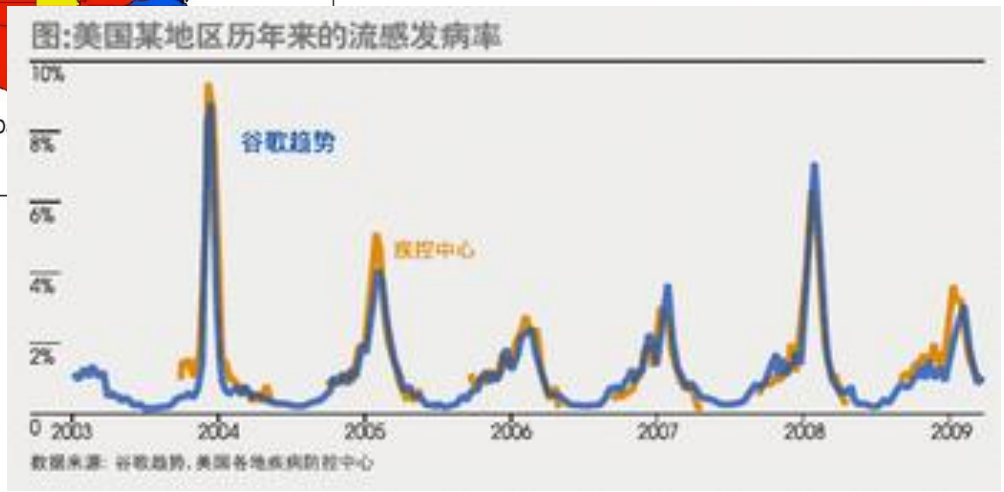
大数据分析

风靡全球的美剧《纸牌屋》

大数据概念、影响、应用

- “谷歌流感趋势”，通过跟踪搜索词相关数据来判断全美地区的流感情况





大数据概念、影响、应用

- 其他典型应用？

大数据关键技术

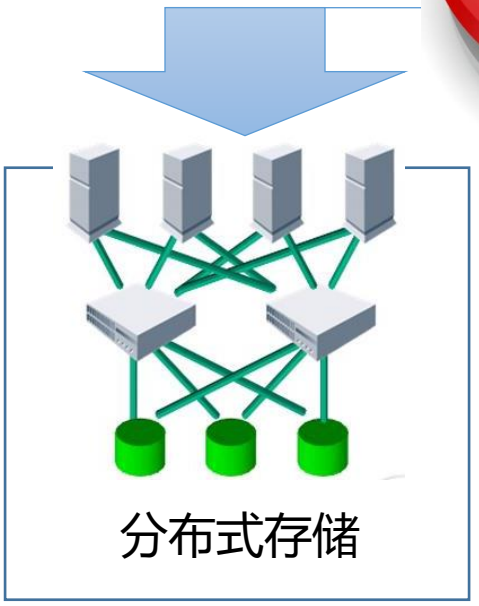
- 大数据技术层次

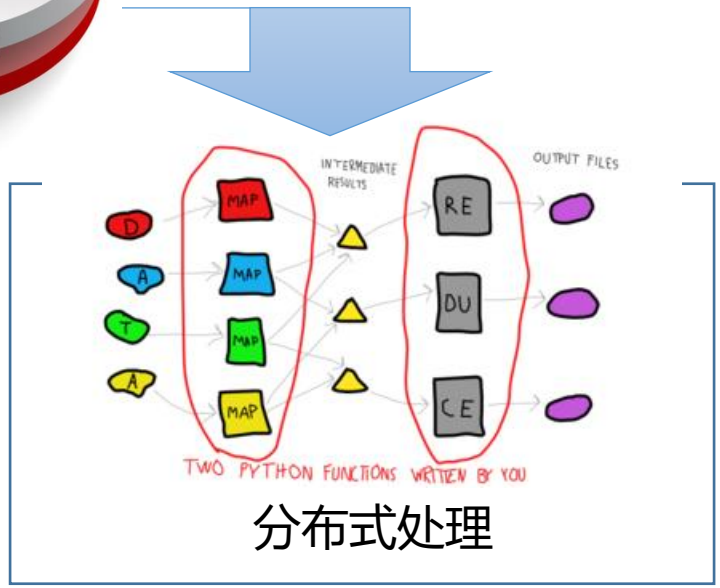
大数据关键技术

技术层面	功能
数据采集	利用ETL工具将分布的、异构数据源中的数据如关系数据、平面数据文件等，抽取到临时中间层后进行清洗、转换、集成，最后加载到数据仓库或数据集市中，成为联机分析处理、数据挖掘的基础；或者也可以把实时采集的数据作为流计算系统的输入，进行实时处理分析
数据存储和管理	利用分布式文件系统、数据仓库、关系数据库、NoSQL数据库、云数据库等，实现对结构化、半结构化和非结构化海量数据的存储和管理
数据处理与分析	利用分布式并行编程模型和计算框架，结合机器学习和数据挖掘算法，实现对海量数据的处理和分析；对分析结果进行可视化呈现，帮助人们更好地理解数据、分析数据
数据隐私和安全	在从大数据中挖掘潜在的巨大商业价值和学术价值的同时，构建隐私数据保护体系和数据安全体系，有效保护个人隐私和数据安全

大数据关键技术

- 两大核心技术





解决海量数据的存储问题

解决海量数据的处理问题

大数据计算模式

- 不同的计算模式需要使用不同的产品

企业中不同的应用场景属于不同的计算模式，需要使用不同的大数据技术

大数据计算模式

- 典型的计算模式

批处理计算

流计算

图计算

查询分析计算

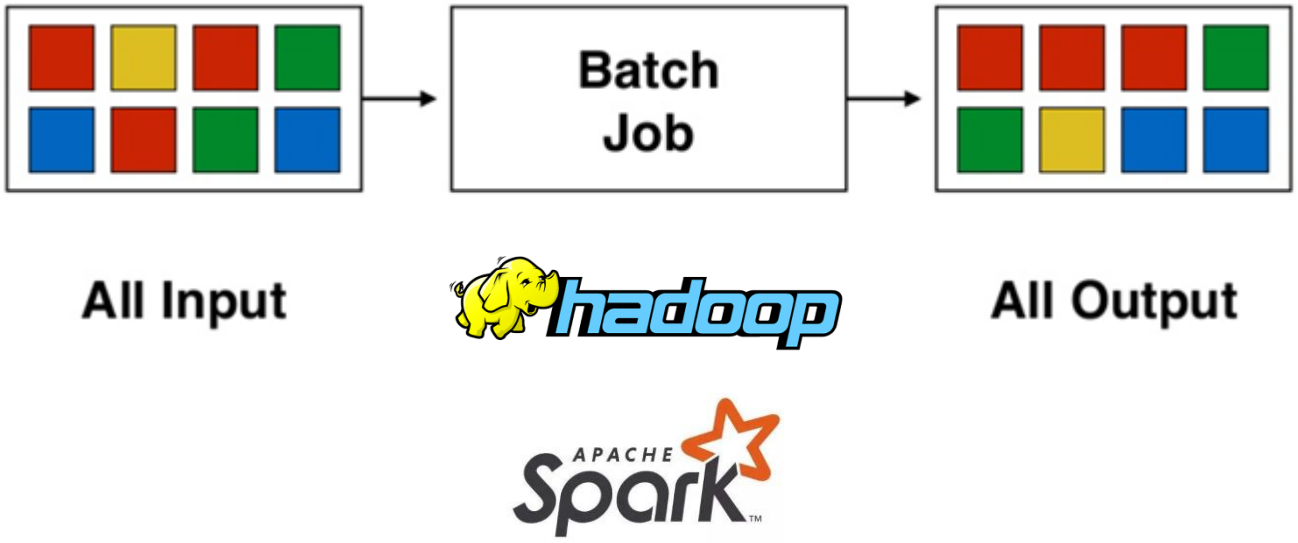
大数据计算模式

- 批处理 vs. 流处理

Batch processing	Stream processing
Data is collected over time	Data streams continuously.
Once data is collected, it's sent for processing	Data is processed piece-by-piece.
Batch processing is lengthy and is meant for large quantities of information that aren't time-sensitive.	Stream processing is fast and is meant for information that's needed immediately.

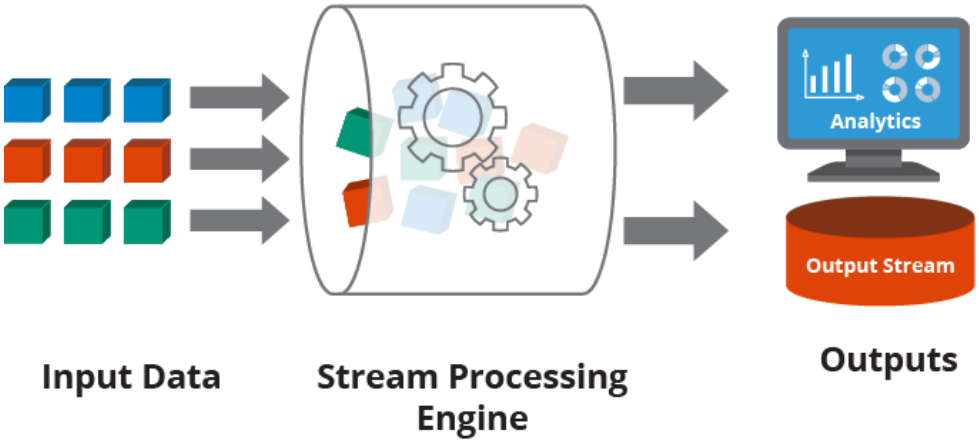
大数据计算模式

- 批处理计算：针对大规模数据的批量处理
  - 代表产品：MapReduce、Spark等



大数据计算模式

- 流计算：针对流数据的实时计算、给出实时响应
  - 代表产品：Spark Streaming、Storm、S4等



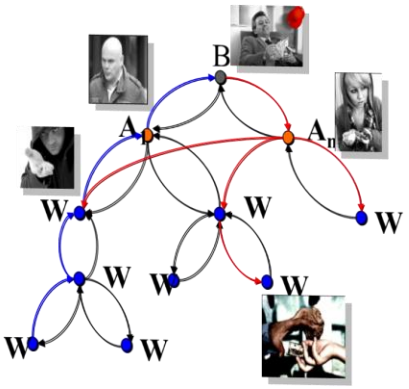
- Fraud detection
- Social media sentiment analysis
- Log monitoring
- Analyzing customer behavior

Spark Streaming

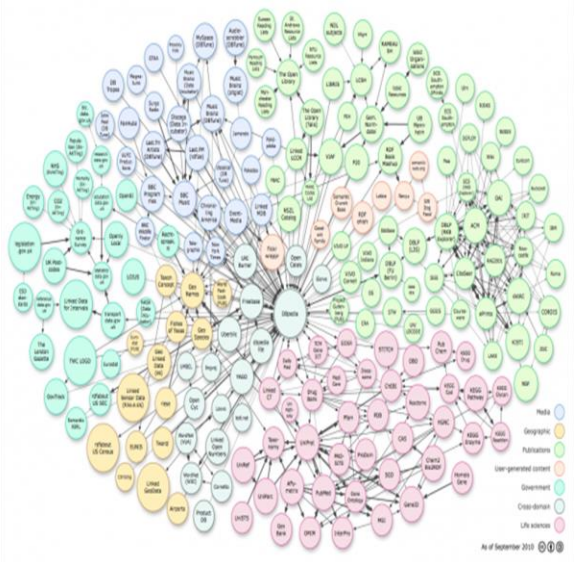
STORM

大数据计算模式

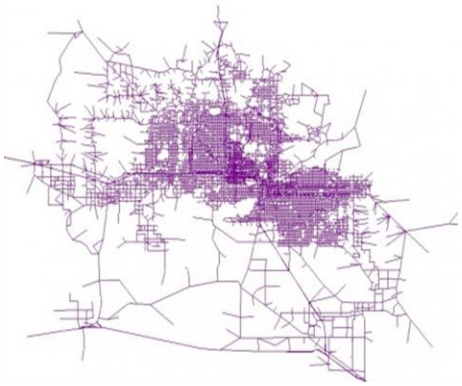
- 图计算：针对大规模图结构数据的处理
  - 代表产品：Pregel、GraphX、Giraph等



社交网络



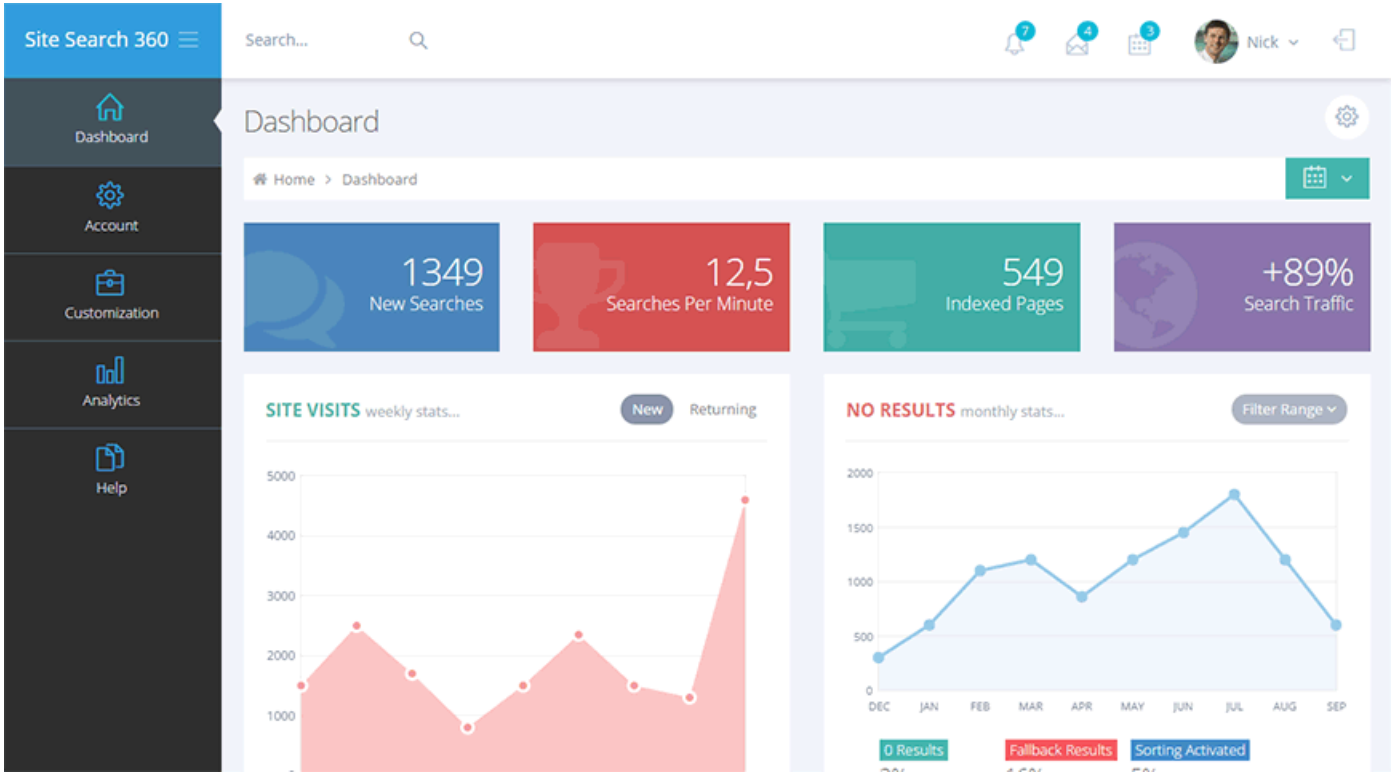
知识图谱



交通路网

大数据计算模式

- 查询分析计算：大规模数据的存储管理和查询分析
  - 代表产品：Dremel、Cassandra等

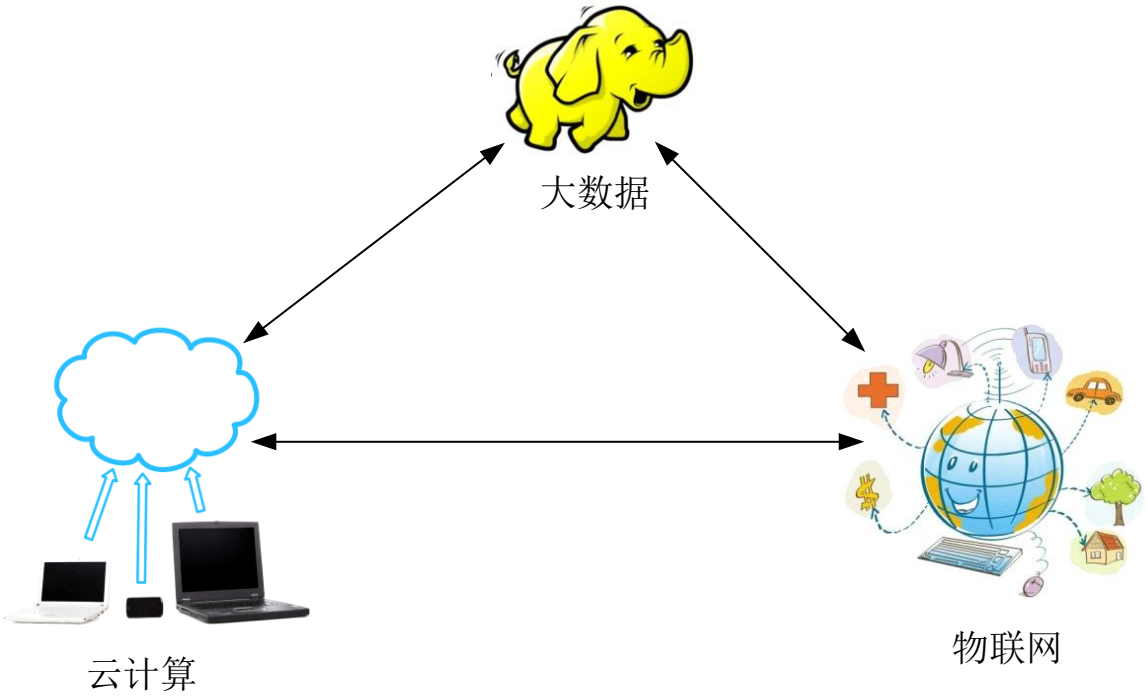


大数据计算模式

大数据计算模式	解决问题	代表产品
批处理计算	针对大规模数据的批量处理	MapReduce、Spark等
流计算	针对流数据的实时计算	Storm、S4、Flume、Streams、Puma、DStream、Super Mario、银河流数据处理平台等
图计算	针对大规模图结构数据的处理	Pregel、GraphX、Giraph、PowerGraph、Hama、GoldenOrb等
查询分析计算	大规模数据的存储管理和查询分析	Dremel、Hive、Cassandra、Impala等

讨论

- 云计算、大数据和物联网代表了IT领域最新的技术发展趋势，三者相辅相成，有着密不可分的关联



讨论

大数据、云计算、物联网与 AI的关系?

小结