



大趨勢

—Big Data



国内大数据

马云对将来的预测，是建立在对顾客行文分析的基础上。“2023年初,阿里巴巴平台上整个买家询盘数急剧下滑，欧美对中国采购在下滑。海关是卖了货,出去后来再取得数据;而我们提前六个月时间从询盘上推断出世界贸易发生了变化了。”

腾讯在天津投资建立亚洲最大的数据中心；
百度也在投资建立大数据处理中心；





美国的大数据战略

2023年3月，美国奥巴马政府宣告投资2亿美元开启“大数据研发计划”，旨在提升和改善从海量和复杂数据中获取知识的能力，加速美国在科学和工程领域发明的步伐，增强国家安全。

这是继1993年美国宣告“信息高速公路”计划后的又一次重大科技发展布署，由美国国家科学基金会、能源部等6个联邦部门共同投资。



目录

大数据的定义

了解大数据

有关技术与应用



大数据时代的背景

“大数据”的诞生：

半个世纪以来，伴随计算机技术全方面融入社会生活，信息爆炸已经积累到了一种开始引起变革的程度。它不但使世界充斥着比以往更多的信息，而且其增长速度也在加紧。信息爆炸的学科如天文学和基因学，发明出了“大数据”这个概念*。如今，这个概念几乎应用到了全部人类智力与发展的领域中。



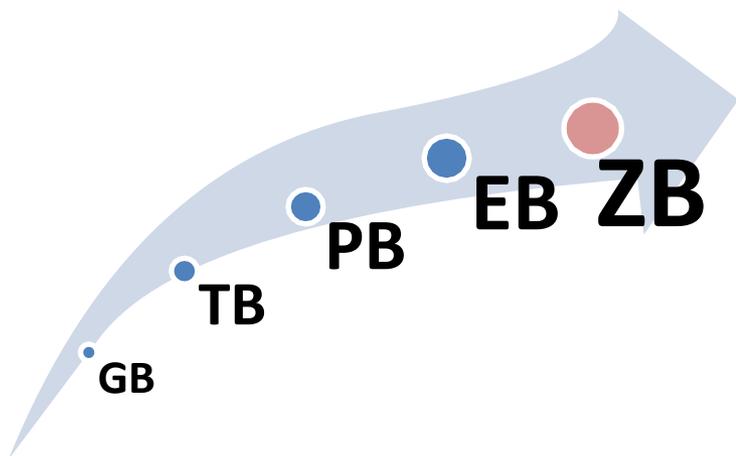
二十一世纪是数据信息大发展的时代，移动互联、社交网络、电子商务等极大拓展了互联网的边界和应用范围，各种数据正在迅速膨胀并变大。

互联网（社交、搜索、电商）、移动互联网（微博）、物联网（传感器，智慧地球）、车联网、GPS、医学影像、安全监控、金融（银行、股市、保险）、电信（通话、短信）都在疯狂产生着数据。

*



数据大爆炸



1PB = 2^{50} 字节

1EB = 2^{60} 字节

1ZB = 2^{70} 字节

地球上至今总共的数据量：

在**2023**年，个人顾客才刚刚迈进**TB**时代，全球一共新产生了约**180EB**的数据；

在**2023**年，这个数字到达了**1.8ZB**。

而有市场研究机构预测：

到**2023**年，整个世界的**数据总量**将会增长**44**倍，到达**35.2ZB**（**1ZB=10**亿**TB**）！

想驾驭这庞大的数据，我们必须了解大数据的特征。



大数据的4V特征

体量Volume

非结构化数据的超大规模和增长
总数据量的80~90%
比结构化数据增长快10倍到50倍
是老式数据仓库的10倍到50倍

多样性Variety

大数据的异构和多样性
诸多不同形式（文本、图像、视频、机器数据）
无模式或者模式不明显
不连贯的语法或句义

价值密度Value

大量的不有关信息
对将来趋势与模式的可预测分析
深度复杂分析（机器学习、人工智能Vs老式商务智能(征询、报告等)

速度Velocity

实时分析而非批量式分析
数据输入、处理与丢弃
立竿见影而非事后见效



目录

大数据的定义

了解大数据

有关技术与应用

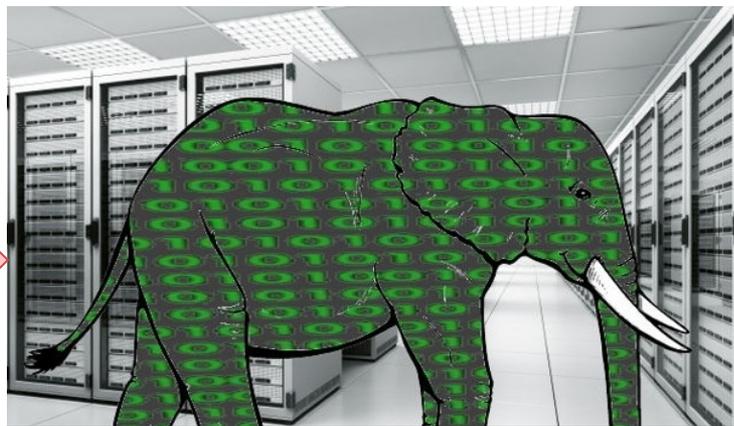
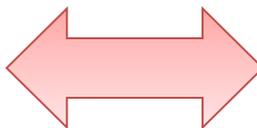


1、密不可分的大数据与云计算

大数据是落地的云



商业模式驱动



应用需求驱动

云计算本身也是大数据的一种业务模式

- 云计算的模式是业务模式，本质是数据处理技术。
- 数据是资产，云为数据资产提供存储、访问和计算。
- 目前云计算更偏重海量存储和计算，以及提供的云服务，运营云应用，但是缺乏盘活数据资产的能力，挖掘价值性信息和预测性分析，为国家、企业、个人提供决策和服务，是大数据关键议题，也是云计算的最终方向。



2、大数据不但仅是“大”

多大？
至少PB
级

比大更主要的是
数据的复杂性，
有时甚至大数据
中的小数据如一
条微博就具有颠
覆性的价值



4、大数据的应用不但是精确营销

- 经过顾客行为分析实现精确营销是大数据的经典应用，但是大数据在各行各业尤其是公共服务领域具有广阔的应用前景





目录

大数据的定义

了解大数据

有关技术与应用



某些有关技术

Ø 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

Ø 大数据技术：

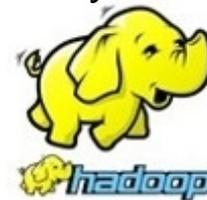
- 数据采集：ETL工具
- 数据存取：关系数据库；NoSQL；SQL等
- 基础架构支持：云存储；分布式文件系统等
- 计算成果呈现：云计算；标签云；关系图等

Ø 存储

- 构造化数据：
 - ρ 海量数据的查询、统计、更新等操作效率低
- 非构造化数据
 - ρ 图片、视频、word、pdf、ppt等文件存储
 - ρ 不利于检索、查询和存储
- 半构造化数据
 - ρ 转换为构造化存储
 - ρ 按照非构造化存储

Ø 处理方案：

- Hadoop (MapReduce技术)
- 流计算 (twitter的storm和yahoo! 的S4)



以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/355042114303011330>