The Age of Big Data
大数据时代

张文涛/酒已

内容

- 什么是大数据
- 相关技术
- 大数据的来“缘”和影响
- 发展动态及方向

什么是大数据

何为大? —数据度量

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

《红楼梦》含标点87万字（不含标点853509字）

每个汉字占两个字节：1汉字=16bit = 2*8位=2bytes

1GB 约等于 671部红楼梦

1TB 约等于 631,903 部

1PB 约等于 647,068,911部

美国国会图书馆藏书（151,785,778册）（2011年4月：收录数据235TB ）

中国国家图书馆：2631万册

- 1EB = 4000倍 美国国会图书馆存储的信息量
- 600美元的硬盘就可以存储全世界所有的歌曲
- MGI估计,全球企业 2010 年在硬盘上存储了超过 7EB(1EB 等于 10 亿 GB)

的新数据,同时,消费者在 PC 和笔记本等设备上存储了超过 6EB 新数据

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

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）

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

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

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）

大数据的来“缘”和影响

从互联网社会化拉开序幕

- YouTube、 twitter、FaceBook、微博等社交网站出现
  - 海量的视频、图片、文本、短消息以及社会间关系信息数据需求出现

跟随互联网的演进

- 互联网需要更好的理解“消费者”的需求

- 消费者也反作用于互联网

Google的精准化理解用户需求

- 通过免费软件及服务来更精确的理解用户行为和习惯
- 通过对用户的更精确理解来提供精确广告服务

传统企业之殇

- 服装企业调查顾客对商品的购买意愿
- 任正非《让听得见炮火的人来决策》
- 张瑞敏：“一个型号几百万产量”到“几十万个型号”

对软件开发和信息化

- 传统软件开发流程→敏捷开发（快速演进）
- 互联网企业面向海量用户群建立自己的生态圈，吸引用户
- 企业信息化不只是订单系统上线，订单处理也需自动化跟上
- 通过分析师对一系列的数据、行为的分析后才能得到用户需求
- 等等

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

大数据带来的影响

- 麦肯锡评估报告中指出大数据在政府公共服务、医疗服务、零售业、制造业、以及涉及个人位置服务等领域都将带来可观的价值

来源于麦肯锡全球研究院

来源于麦肯锡全球研究院

大数据带来的影响

- 政府等公共职能管理
  - 重视应用大数据技术，盘活各地云计算中心资产：把原来大规模投资产业园、物联网产业园从政绩工程，改造成智慧工程
  - 在安防领域，应用大数据技术，提高应急处置能力和安全防范能力
  - 在民生领域，应用大数据技术，提升服务能力和运作效率，以及个性化的服务，比如医疗、卫生、教育等部门
  - 解决在金融，电信领域等中数据分析的问题：一直得到得极大的重视，但受困于存储能力和计算能力的限制，只局限在交易数型数据的统计分析
  - 政府投入将形成示范效应，大大推动大数据的发展

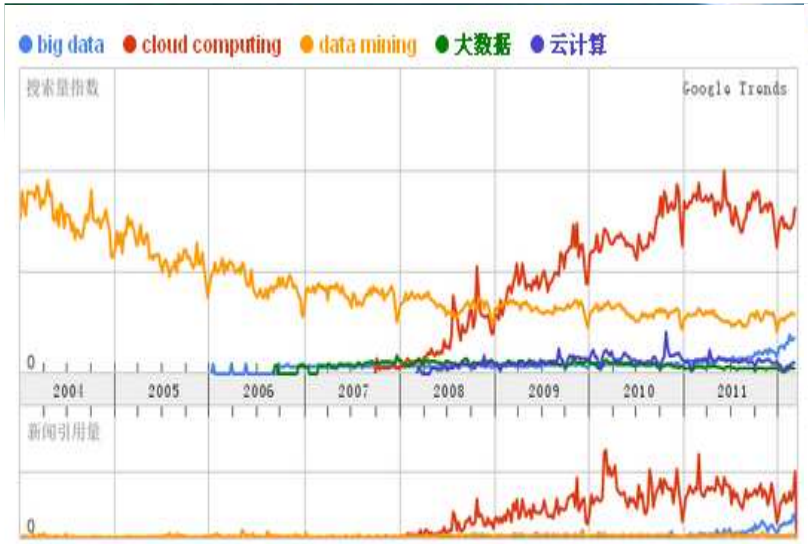
大数据带来的机遇

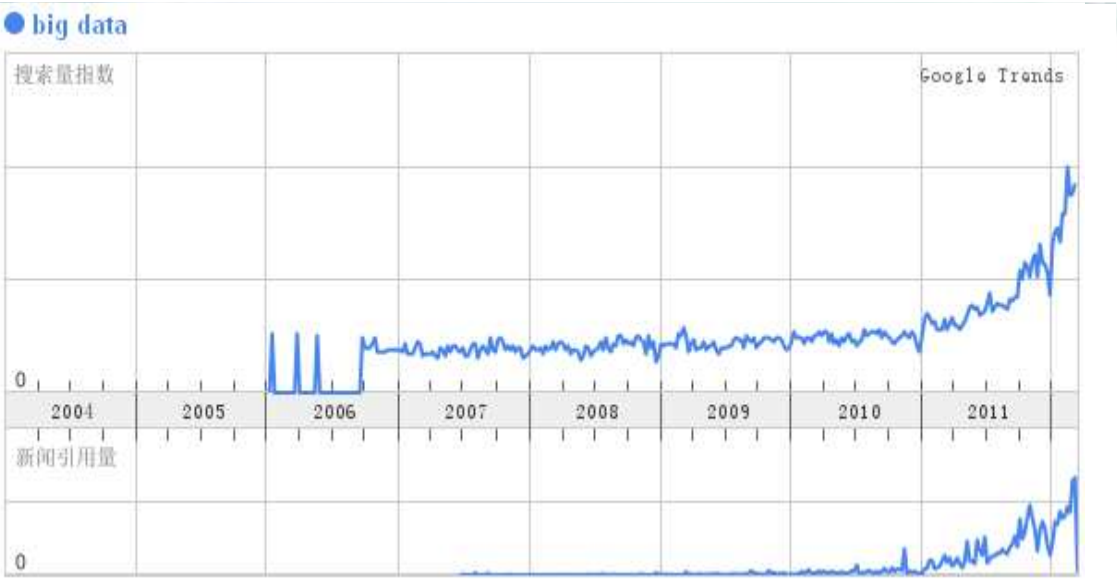
- 大数据赋予我们洞察未来的能力
  - 马云成功预测 2008 年经济危机
  - “2008 年初,阿里巴巴平台上整个买家询盘数急剧下滑,欧美对中国 采购在下滑。海关是卖了货,出去以后再获得数据;而我们提前半年时间 从询盘上推断出世界贸易发生变化了。”
  - 通常而言，买家在采购商品前,会比较多家供应商的产品,反映到阿里巴巴网站统计数据中,就是查询点击的数量和购买点击的数量会保持一个相对的数值,综合各个维度的数据可建立用户行为模型。因为数据样本巨大,保证用户行为模型的准确性。因此在这个案例中,询盘数据的下降,自然导致买盘的下降 。

大数据带来的影响——刀刃的另一面

- 更多的隐私、安全性问题
  - 多少密码和账号是因为“社交网络”流出去的?
  - 2011年4月索尼的系统漏洞导致7700万用户资料失窃
  - 2011年4月，iOS被发现会按照时间顺序记录用户的位置坐标信息
  - 2011年CSDN密码泄露事件

发展动态及方向





排序依据 big data

国家/地区	城市	语言
1. India 印度	1. Bangalore, India 班加罗尔, 印度	1. Bulgarian 保加利亚文
2. South Korea 韩国	2. Chennai, India 清奈, 印度	2. Korean 韩文
3. Bulgaria 保加利亚	3. Mumbai, India 孟买, 印度	3. English 英文
4. Singapore 新加坡	4. New Delhi, India 新德里, 印度	4. Indonesian 印度尼西亚文
5. South Africa 南非	5. San Francisco, CA, USA 旧金山, 美国	5. Swedish 瑞典文
6. Australia 澳大利亚	6. Austin, TX, USA 奥斯汀, 美国	6. Italian 意大利文
7. United States 美国	7. Singapore, Singapore 新加坡, 新加坡	7. Portuguese 葡萄牙文
8. Indonesia 印度尼西亚	8. New York, NY, USA 纽约, 美国	8. Dutch 荷兰文
9. United Kingdom 英国	9. Sydney, Australia 悉尼, 澳大利亚	9. German 德文
10. Canada 加拿大	10. Melbourne, Australia 墨尔本, 澳大利亚	10. Chinese 中文

发展动态

- 在2009年中，美国政府通过启动Data.gov网站的方式进一步开放了数据的大门，这个网站向公众提供各种各样的政府数据
- 在2009年，欧洲一些领先的研究型图书馆和科技信息研究机构建立了伙伴关系致力于改善在互联网上获取科学数据的简易性
- 2011年5月：肯锡全球研究院（MGI）发布了一份报告——《大数据：创新、竞争和生产力的下一个新领域》，大数据开始备受关注
- 在2011年12 月8 日工信部发布的物联网十二五规划上，把信息处理技术作为4 项关键技术创新工程之一被提出来，其中包括了海量数据存储、数据挖掘、图像视频智能分析，这都是大数据的重要组成部分
- 2012年1月份：瑞士达沃斯召开的世界经济论坛上，大数据是主题之一，会上发布的报告《大数据，大影响》(Big Data, Big Impact) 宣称，数据已经成为一种新的经济资产类别，就像货币或黄金一样

发展动态

图表22：巨头在大数据时代的业务布局

公司	时间	收购
EMC	2010 年 7 月	收购数据库软件供应商 Greenplum，花费 3 亿美元
EMC	2009 年 7 月	收购数据复制解决方案提供商 Data Domain，花费 24 亿美元
EMC	2009 年起	陆续收购 Archer Technologies，SourceLabs，FastScale Technology，Configuresoft，and Varonis Systems
IBM	2010 年 9 月	收购数据库分析供应商 Netezza 公司，花费 17 亿美元
IBM	2010 年 10 月	收购网络分析软件供应商 Coremetrics
IBM	2009 年 10 月	收购数据分析和统计软件提供商 SPSS，花费 12 亿美元
IBM	2009 年 1 月	收购业务规则管理软件供应商 ILOG，花费 3 亿 4 千万美元
IBM	2007 年	花费 20 亿美元收购商务智能软件供应商 Cognos
SAP	2007 年	68 亿美元收购了全球商业智能软件老大 Business Object
Oracle	2008 年 11 月	33 亿美元收购商业智能解决方案提供商海波龙(Hyperion)公司
Oracle	2009 年 7 月	收购专注于数据复制和实时数据集成解决方案的私人企业 GoldenGate Software
HP	2011 年	花费 100 亿美元收购英国软件公司 Autonomy
微软	2006 年	收购商业智能软件公司 ProClarity
微软	2008 年	收购数据仓库产品厂商 DATAllegro

来源：国金证券研究所

图表26： IBM 依然保持银行、电信行业的强势地位

背景

- IBM 的策略是提供一个全面的方法来解决前所未有的信息爆炸提出的挑战，因为信息量无论在流量、种类、速度还是活力上都是爆炸式增长
- IBM 一直致力于扩大对包括数据仓库中的大数据、信息流和结构化数据的分析
-

收购

在过去四年中，IBM 已经投入超过 120 亿美元进行了 23 项相关并购，其中包括：

- 2010 年 9 月收购数据库分析供应商 Netezza 公司，花费 17 亿美元
- 2010 年 10 月收购网络分析软件供应商 Coremetrics
- 2009 年 10 月收购数据分析和统计软件提供商 SPSS, 花费 12 亿美元
- 2009 年 1 月收购业务规则管理软件供应商 ILOG, 花费 3 亿 4 千万美元
- 2007 年花费 20 亿美元收购商务智能软件供应商 Cognos

提供的产品和服务

- IBM 大数据提供的服务包括数据分析，文本分析，蓝色云杉（混搭供电合作的网络平台）;业务事件处理; IBM Mashup Center 的计量，监测，和商业化服务（MMMS）
- IBM 的大数据产品组合中的最新系列产品的 InfoSphere bigInsights，基于 Apache Hadoop。
- 该产品组合包括
- 打包的 Apache Hadoop 的软件和服务,代号是 bigInsights 核心,用于开始大数据分析
- 软件被称为 bigsheet，软件目的是帮助从大量数据中轻松、简单、直观的提取、批注相关信息
- 为金融，风险管理，媒体和娱乐等行业量身定做的行业解决方案

合作伙伴

- 2010 年 5 月，选择与使用开源技术设计海量数据分析软件的 Apache Hadoop 合作

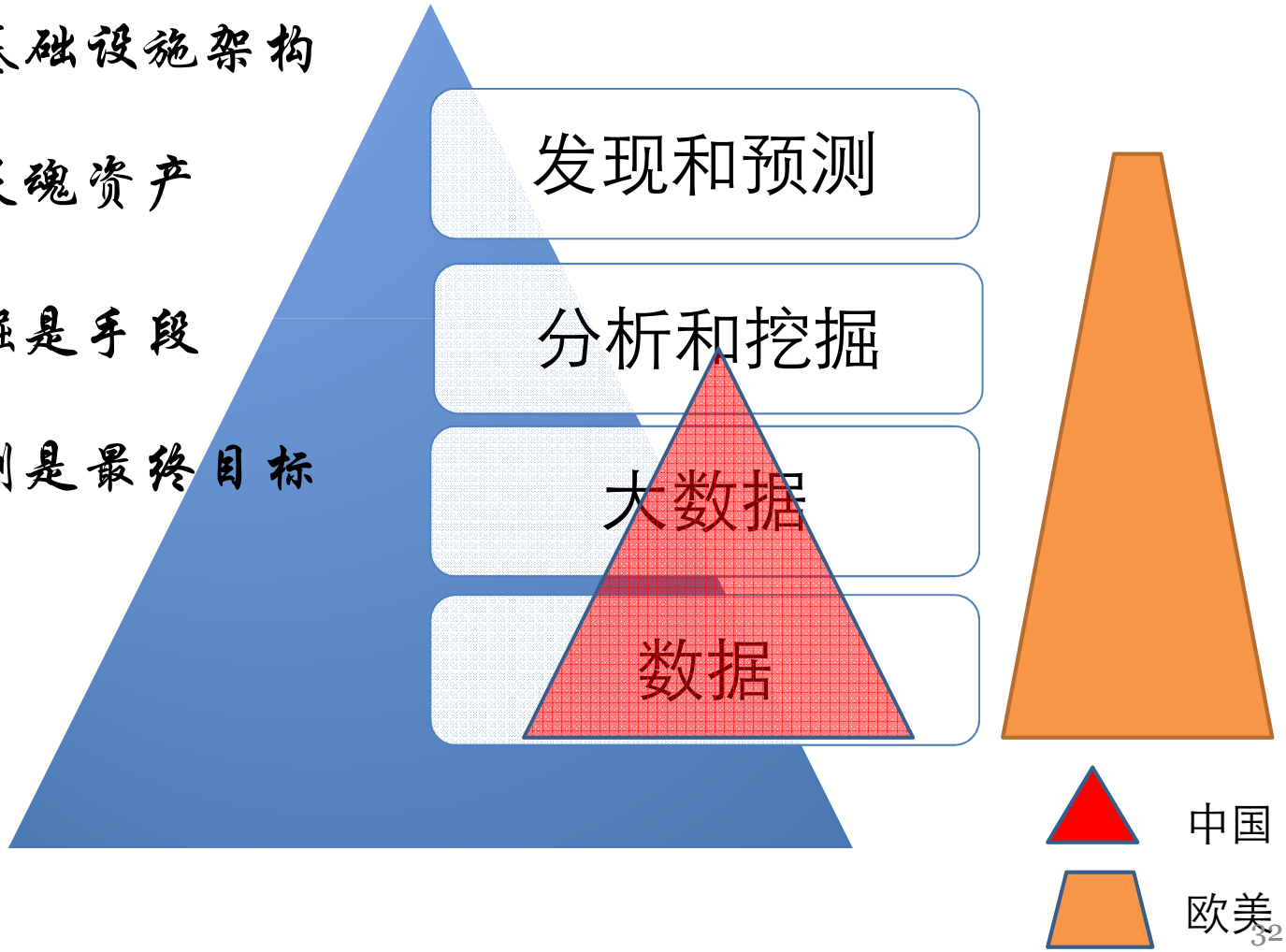
客户

- 全球 25000 多分析客户，其中包括 22 个全球前 24 大商业银行，18 个前 22 电信运营商和 11 个美国排名前 12 位的专业零售商
- 赢得新客户，包括 Avis, FUJIFILM Imaging Colorants,西班牙的社会服务机构，and Hildebrand to extract insights from data
- 2010 年 8 月，北卡罗来纳州使用 IBM 大数据技术，以简化搜索和匹配潜在的大学研究项目的投资和合作机会的过程
- 其他客户：Hertz，大英图书馆

来源： MGI 国金证券研究所

大数据的方向

- 云计算是基础设施架构
- 大数据是灵魂资产
- 分析、挖掘是手段
- 发现和预测是最终目标



谢谢！

参考文献

- 1. http://en.wikipedia.org/wiki/Big_data
- 2. http://www-01.ibm.com/software/data/bigdata/
- 3. http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation