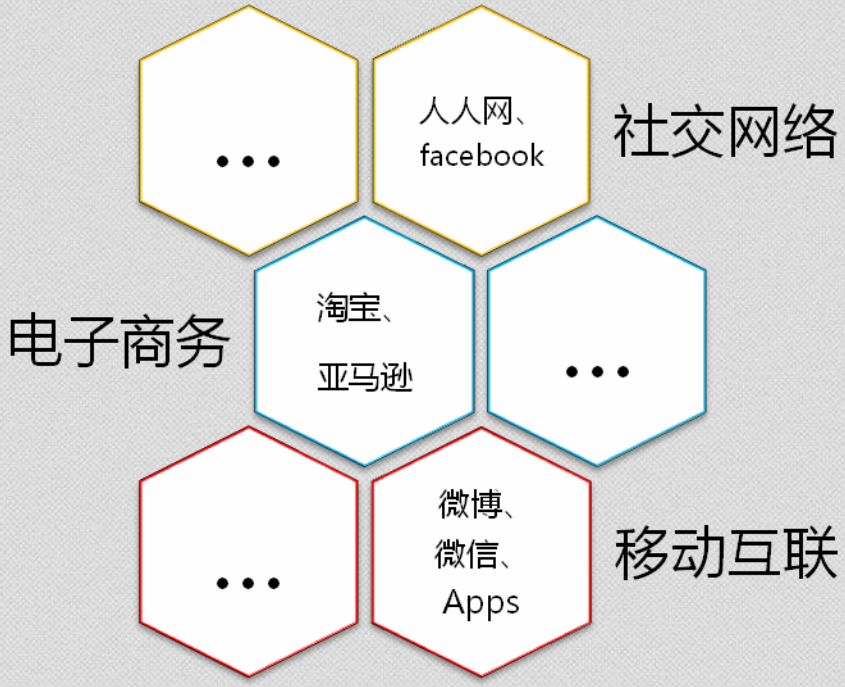
BIG DATA

走近未来的新石油—大数据

——浅谈IT时代到DT时代

背景



半个世纪以来，随着计算机技术全面融入社会生活，信息爆炸已经积累到了一个开始引发变革的程度。它不仅使世界充斥着比以往更多的信息，而且其增长速度也在加快。信息爆炸的学科如天文学和基因学，创造出了“大数据”这个概念。

如今，这个概念几乎应用到了所有人类智力与发展的领域中。

这是一场生活、工作与思维的大变革

大数据到底有多大？一组名为“互联网上一天”的数据告诉我们，一天之中

互联网产生的全部内容可以刻满1.68亿张DVD；

发出的邮件有2940亿封之多（相当于美国两年的纸质信件数量）；

发出的社区帖子达200万个（相当于《时代》杂志770年的文字量）；

卖出的手机为37.8万台，高于全球每天出生的婴儿数量37.1万……

新的时代，人们从信息的被动接受者变成了主动创造者

全球每秒钟发送 2.9 百万封电子邮件，一分钟读一篇的话，足够一个人昼夜不息的读5.5 年...

每天会有 2.88 万个小时的视频上传到Youtube，足够一个人昼夜不息的观看3.3 年...

推特上每天发布 5 千万条消息，假设10 秒钟浏览一条信息，这些消息足够一个人昼夜不息的浏览16 年...

每天亚马逊上将产生 6.3 百万笔订单...

每个月网民在Facebook 上要花费7 千亿分钟，被移动互联网使用者发送和接收的数据高达1.3EB...

Google 上每天需要处理24PB 的数据...

简介

定义？

- 大数据(big data)，或称巨量资料，指的是所涉及的资料量规模巨大到无法透过目前主流软件工具，在合理时间内达到撷取、管理、处理、并整理成为帮助企业经营决策更积极目的的资讯。
- 大数据的核心是*预测*，通过把数学算法运用到海量的数据上来预测事情发生的可能性

过去		现在
随机样本		全体数据
精确性	→	混杂性
因果关系		相关关系

4“V”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 是传统数据仓库的10倍到50倍

多样性Variety

大数据的异构和多样性 很多不同形式（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

大量的不相关信息 对未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传统商务智能(咨询、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

大数据的构成

大数据包括：
交易数据和交互数据集在内的所有数据集

大数据 = 海量数据 + 复杂类型的数据

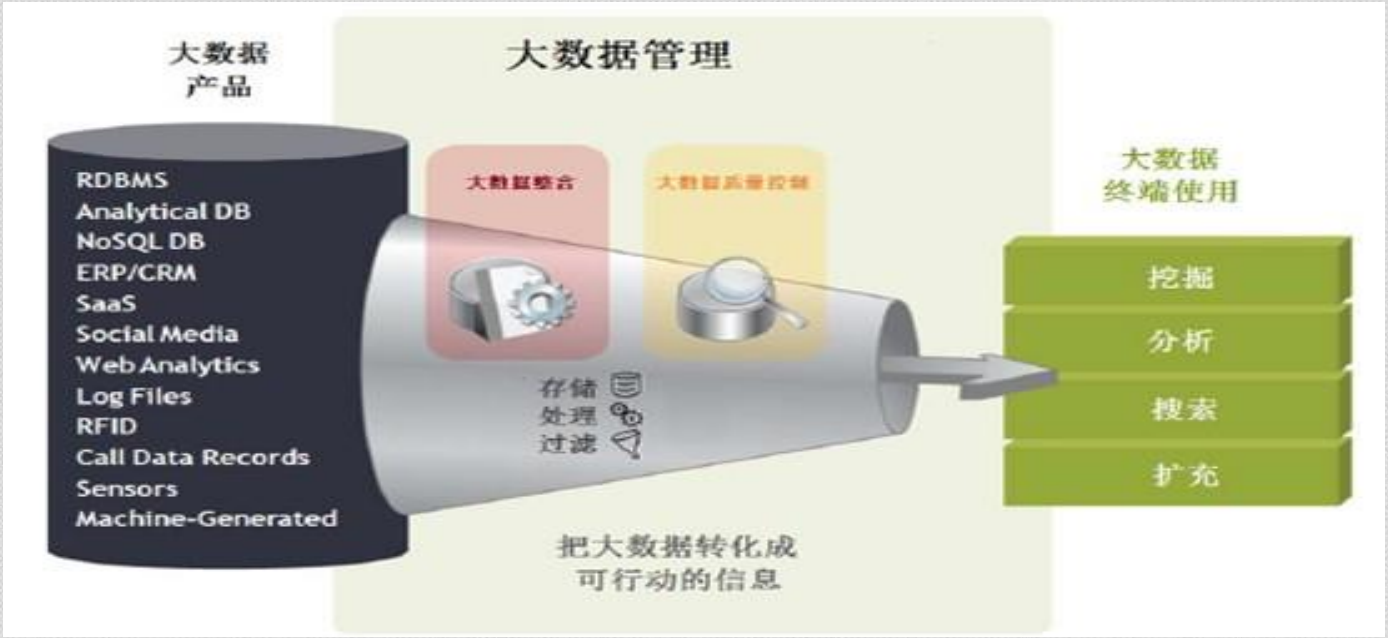
海量交易数据：企业内部的经营交易信息主要包括联机交易数据和联机分析数据，是结构化的、通过关系数据库进行管理和访问的静态、历史数据。通过这些数据，我们能了解过去发生了什么。

海量交互数据：源于Facebook、Twitter、LinkedIn及其他来源的社交媒体数据构成。它包括了呼叫详细记录CDR、设备和传感器信息、GPS和地理定位映射数据、通过管理文件传输Manage File Transfer协议传送的海量图像文件、Web文本和点击流数据、科学信息、电子邮件等等。可以告诉我们未来会发生什么。

海量数据处理：大数据的涌现已经催生出了设计用于数据密集型处理的架构。例如具有开放源码、在商品硬件群中运行的Apache Hadoop。

Big Data技术

企业用以分析的数据越全面，分析的结果就越接近于真实。大数据分析意味着企业能够从这些新的数据中获取新的洞察力，并将其与已知业务的各个细节相融合



大数据技术将被设计用于在成本可承受（economically）的条件下，通过非常快速（velocity）的采集、发现和分析，从大量化（volumes）、多类别（variety）的数据中提取价值（value），将是IT 领域新一代的技术与架构

不得不说的那些人

- 《the economics》称，在大数据领域，他是最受人尊敬的权威发言人之一
- 哈佛大学肯尼迪学院信息监管科研项目负责人、网络监管项目负责人
- 最早洞见大数据时代发展趋势的数据科学家之一
- 大数据商业应用的引路人
- 欧盟互联网官方政策背后真正的制定者与参与者，是众多国家国家政府高层的智囊团成员
- 代表作《大数据时代》、《删除》等

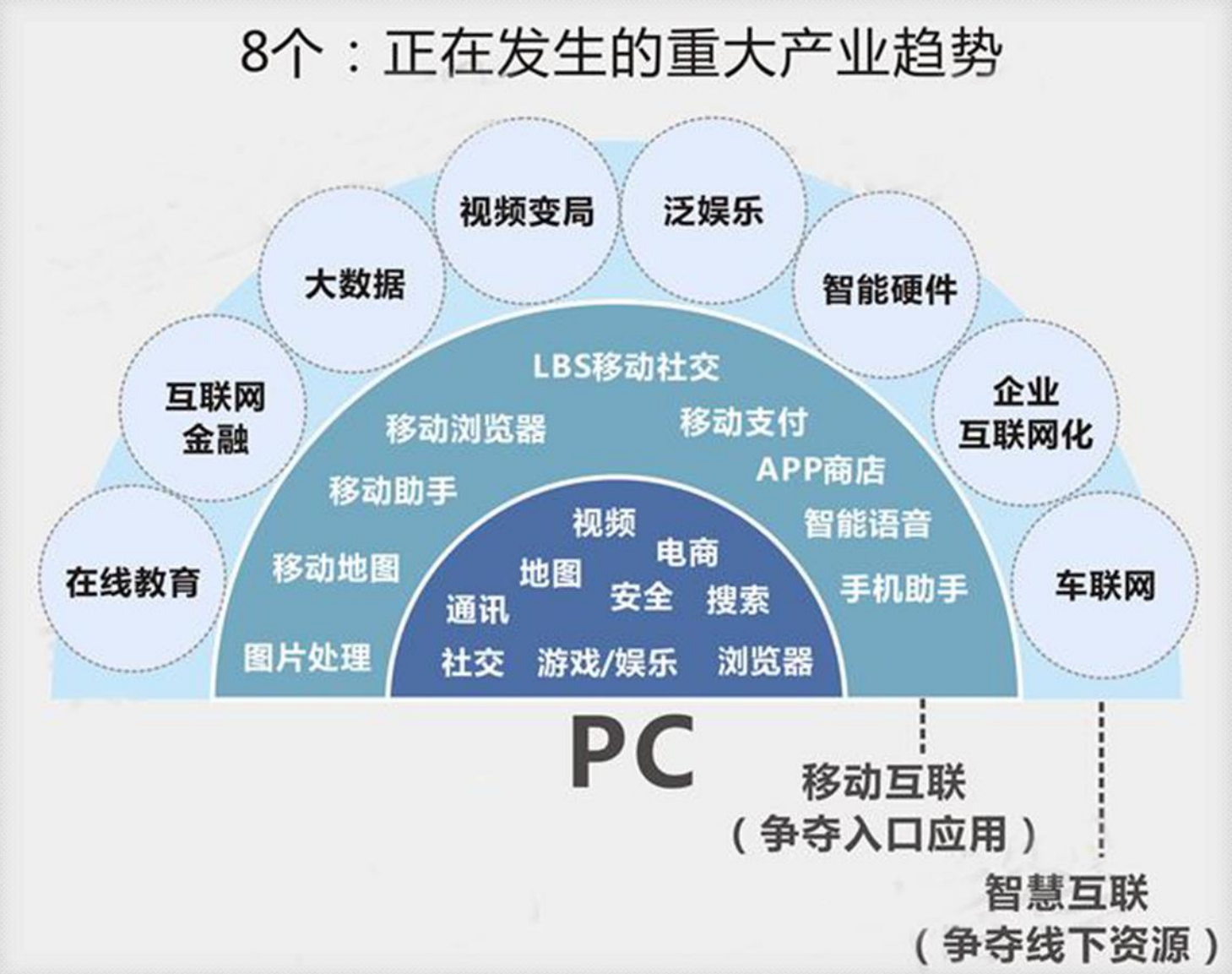
维克托• 迈尔—舍恩伯格

新的思维：互联网+

“**互联网**+”行动计划将重点促进以云计算、物联网、大数据为代表的新一代信息技术与现代制造业、生产性服务业等的融合创新，发展壮大新兴业态，打造新的产业增长点，为大众创业、万众创新提供环境，为产业智能化提供支撑，增强新的经济发展动力，促进国民经济提质增效升级。

马化腾认为：“互联网加一个传统行业，意味着什么呢？其实是代表了一种能力，或者是一种外在资源和环境，对这个行业的一种提升。”

马云说过：现在是鼠标+水泥的新经济时代。企业只有使用互联网思维改变其传统行业的商业模式才能得以发展。


8个：正在发生的重大产业趋势
视频变局
泛娱乐
大数据
智能硬件
互联网
金融
企业
互联网化
LBS移动社交
移动浏览器
移动支付
APP商店
移动助手
智能语音
视频
电商
在线教育
移动地图
地图
安全
搜索
手机助手
车联网
通讯
图片处理
社交
游戏/娱乐
浏览器
PC
移动互联
（争夺入口应用）
智慧互联
（争夺线下资源）

互联网化加速产业发展

随着大数据、移动、云计算等技术地规模应用，以BAT为代表的中国互联网企业正逐级深入地将业务向传统行业延伸，使得传统企业不断将营销、渠道、产品、运营等层面的商务活动依赖于互联网。

DT时代：迎来新的探索

马云说 ：人类正从IT时代走向DT时代 。

DT(Data technology)时代，它是以服务大众、激发生产力为主的技术。

未来的竞争拼的是人才和创新价值的能力，拼的是你的数据能够给社会创造多少价值，用数据挣钱才是未来真正核心所在，靠控制成本做生意。

大数据的应用——政府

政府职能变革:

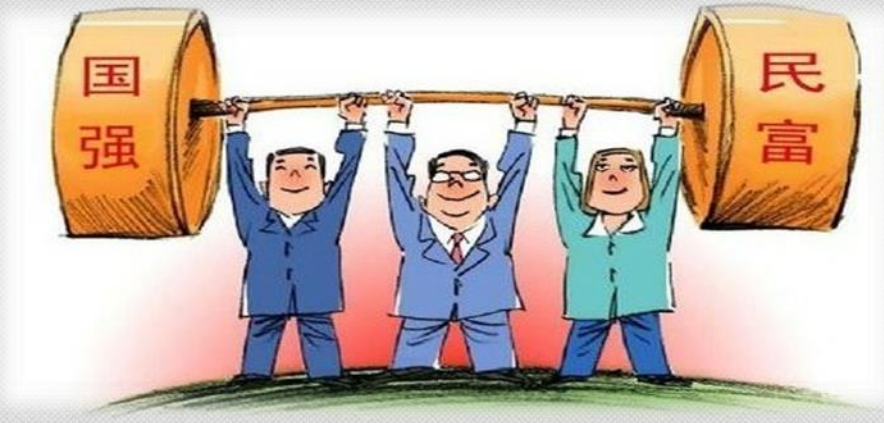
重视应用大数据技术，盘活各地云计算中心资产：把原来大规模投资产业园、物联网产业园从政绩工程，改造成智慧工程；

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；

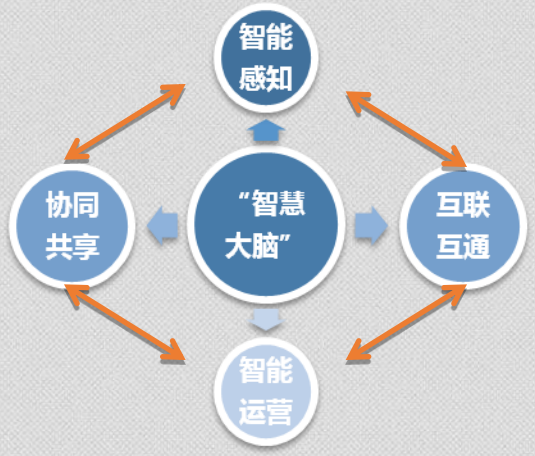
解决在金融，电信领域等中数据分析的问题：一直得到得极大的重视，但受困于存储能力和计算能力的限制，只局限在交易数型数据的统计分析；

政府投入将形成示范效应，大大推动大数据的发展



大数据案列分析1

对城市电力供应和调控



大数据的应用——热点：智慧城市

美国奥巴马政府在白宫网站发布《大数据研究和发展倡议》，提出“通过收集、处理庞大而复杂的数据信息，从中获得知识和洞见，提升能力，加快科学、工程领域的创新步伐，强化美国国土安全，转变教育和学习模式”；

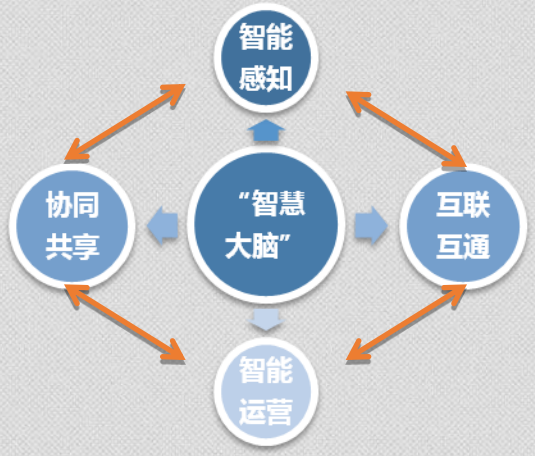
中国工程院院士邬贺铨说道，“智慧城市是使用智能计算技术使得城市的关键基础设施的组成和服务更智能、互联和有效，随着智慧城市的建设，社会将步入“大数据”时代。”

难点：

1、在最初就合理规划智慧城市（深度思考哪些领域能够运用）；

2、在城市发展基础设施和“云产业”的同时，更多重视“数据”的价值；

3、在大数据处理领域的核心技术不足，需要政府更大的投入。



大数据案列分析2

大数据对高铁线路站点供票的配置

大数据的应用——企业在投入

IBM：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

微软：2011年1月与惠普（具体而言是HP数据库综合应用部门）合作目标是开发了一系列能够提升生产力和提高决策速度的设备。

EMC：EMC 斩获了纽交所和Nasdaq；大数据解决方案已包括40多个产品。

Oracle：Oracle大数据机与Oracle Exalogic中间件云服务器、Oracle Exadata数据库云服务器以及Oracle Exalytics商务智能云服务器一起组成了甲骨文最广泛、高度集成化系统产品组合。

大数据案列分析3

大数据对广告行业精准投放案列

大数据的应用——未来，改变一切

数据的再利用：由于在信息价值链中的特殊位置，有些公司可能会收集到大量的数据，但他们并不急需使用也不擅长再次利用这些数据。例如，移动电话运营商手机用户的位置信息来传输电话信号，这对以他们来说，数据只有狭窄的技术用途。但当它被一些发布个性化位置广告服务和促销活动的公司再次利用时，则变得更有价值。

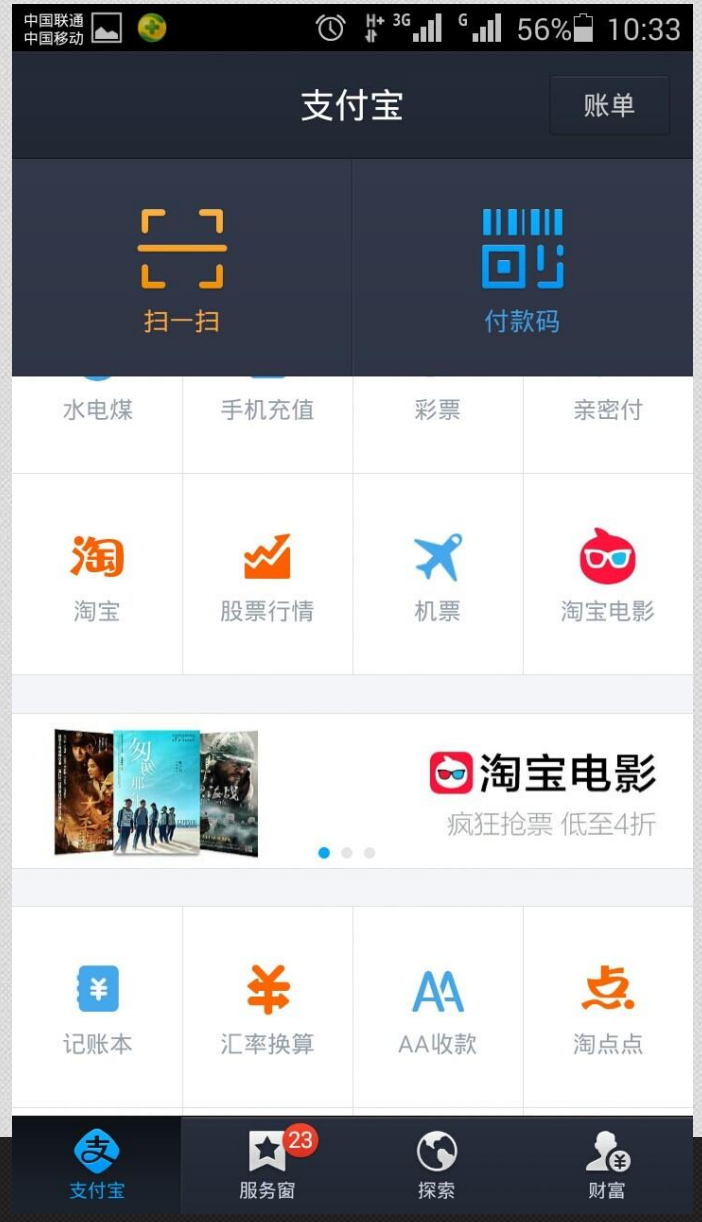
大数据价值链的3大构成：数据本身、技能与思维：其中三者兼具的谷歌公司，谷歌在刚开始收集数据的时候就已经有多次使用数据的想法。比方说，它的街景采集车手机全球定位系统数据不光是为了创建谷歌地图，也是为了制成全自动汽车以及谷歌眼镜等与实景交汇的产品。

传统行业最终都会转变为大数据行业，无论是金融服务也、医药还是制造业。

《大数据时代》

大数据案列分析4

大数据对用户行为的分析用于电子商务成交



大数据应用成功的几大案例

一、大数据与乔布斯癌症治疗：乔布斯是世界上第一个对自身所有DNA和肿瘤DNA进行排序的人。为此，他支付了高达几十万美元的费用。他得到的不是样本，而是包括整个基因的数据文档。医生按照所有基因按需下药，最终这种方式帮助乔布斯延长了好几年的生命。

二、Google成功预测冬季流感：Google成功预测冬季流感 2009年，Google通过分析5000万条美国人最频繁检索的词汇，将之和美国疾病中心在2003年到2008年间季节性流感传播时期的数据进行比较，并建立一个特定的数学模型。最终google成功预测了2009冬季流感的传播甚至可以具体到特定的地区和州。

三、**啤酒与尿布：**全球零售业巨头沃尔玛在对消费者购物行为分析时发现，男性顾客在购买婴儿尿片时，常常会顺便搭配几瓶啤酒来犒劳自己，于是尝试推出了将啤酒和尿布摆在一起的促销手段。没想到这个举措居然使尿布和啤酒的销量都大幅增加了。如今，“啤酒＋尿布”的数据分析成果早已成了大数据技术应用的经典案例，被人津津乐道。

四、**奥巴马大选连任成功：**奥巴马大选连任成功 2012年11月奥巴马大选连任成功的胜利果实也被归功于大数据，因为他的竞选团队进行了大规模与深入的数据挖掘。时代杂志更是断言，依靠直觉与经验进行决策的优势急剧下降，在政治领域，大数据的时代已经到来；各色媒体、论坛、专家铺天盖地的宣传让人们对大数据时代的来临兴奋不已，无数公司和创业者都纷纷跳进了这个狂欢队伍。

五、**超市预知高中生顾客怀孕：**明尼苏达州一家塔吉特门店被客户投诉，一位中年男子指控塔吉特将婴儿产品优惠券寄给他的女儿——一个高中生。但没多久他却来电道歉，因为女儿经他逼问后坦承自己真的怀孕了。塔吉特百货就是靠着分析用户所有的购物数据，然后通过相关关系分析得出事情的真实状况。

六、**好巧网预知哪个酒店更适合你：**去境外旅行，人生地不熟的，对于住哪里更合适，旅客往往一头雾水。好巧网的技术专家，通过结合酒店、景点和旅客等多项大数据，快速帮用户找到最适合自己的酒店。只需输入目的地名称，就可轻松感知结果。

七、**微软大数据成功预测奥斯卡21项大奖：**2013年，微软纽约研究院的经济学家大卫•罗斯柴尔德（David Rothschild）利用大数据成功预测24个奥斯卡奖项中的19个，成为人们津津乐道的话题。今年罗斯柴尔德再接再厉，成功预测第86届奥斯卡金像奖颁奖典礼24个奖项中的21个，继续向人们展示现代科技的神奇魔力。

八、**QQ圈子把前女友推荐给未婚妻：** 2012年3月腾讯推出QQ圈子，按共同好友的连锁反应摊开用户的人际关系网，把用户的前女友推荐给未婚妻，把同学同事朋友圈子分门别类，利用大数据处理能力给人带来“震撼”。

九、胸部最大的是新疆妹子：淘宝数据平台显示，购买最多的文胸尺码为B罩杯。B罩杯占比达41.45%，其中又以75B的销量最好。其次是A罩杯，购买占比达25.26%，C罩杯只有8.96%。在文胸颜色中，黑色最为畅销。以省市排名，胸部最大的是新疆妹子。

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

大数据存储技术

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。

解决方案：Hadoop（MapReduce技术）、流计算（twitter的storm和yahoo！的S4）

技术领域技术架构的挑战：

1、对现有数据库管理技术的挑战：传统的数据库部署不能处理数TB 级别的数据，也不能很好的支持高级别的数据分析。急速膨胀的数据体量即将超越传统数据库的管理能力。如何构建全球级的分布式数据库(Globally-Distributed Database) ，可以扩展到数百万的机器，数已百计的数据中心，上万亿的行数据。

2、经典数据库技术并没有考虑数据的多类别（variety）

SQL（结构化数据查询语言），在设计的一开始是没有考虑非结构化数据的。

3、实时性的技术挑战：

一般而言，像数据仓库系统、BI应用，对处理时间的要求并不高。因此这类应用往往运行1、2天获得结果依然可行的。但实时处理的要求，是区别大数据应用和传统数据仓库技术、BI技术的关键差别之一。

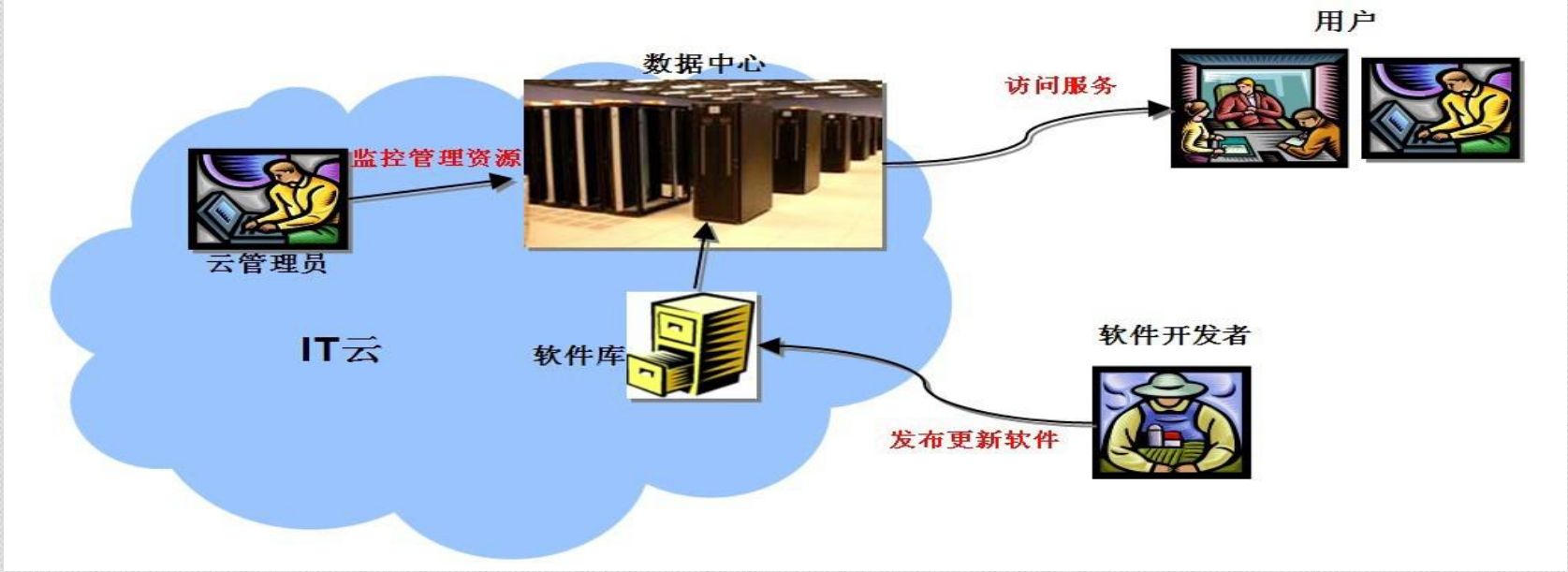
网络架构、数据中心、运维的挑战：

人们每天创建的数据量正呈爆炸式增长，但就数据保存来说，我们的技术改进不大，而数据丢失的可能性却不断增加。

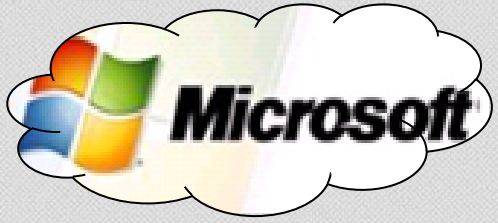
如此庞大的数据量首先在存储上就会是一个非常严重的问题，硬件的更新速度将是大数据发展的基石。

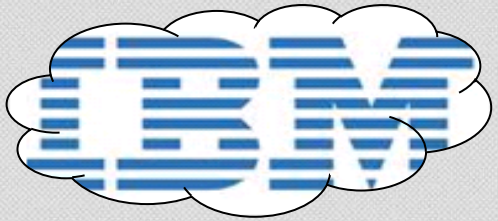
大数据与云计算

- 云计算的模式是业务模式，本质是数据处理技术。
- 数据是资产，云为数据资产提供存储、访问和计算。
- 当前云计算更偏重海量存储和计算，以及提供的云服务，运行云应用，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，是大数据核心议题，也是云计算的最终方向。



大数据与云计算





蓝蓝的天上白云飘

白云下面数据跑

如果数据是财富，那么大数据就是宝藏，而云计算就是挖掘和利用宝藏的利器！没有强大的计算能力，数据宝藏终究是镜中花；没有大数据的积淀，云计算也只能是杀鸡用的宰牛刀！

大数据行业发展前景：机遇和挑战

企业和政府机构越来越看重通过数据驱动决策，这导致他们对分析和信息管理专业的需求不断增加。以下是大数据领域从业人员的几个趋势。

趋势一：薪金将继续增长

趋势二：大数据人才供不应求

趋势三：雇佣外包

趋势四：大数据专业人士需要不断进步

趋势五：精通大数据的专业人才将成为最重要的业务角色

趋势六：大数据领域需要数据科学家

趋势七：高校回应大数据人才缺口

趋势八：数据驱动的工作令人满意并充满挑战

趋势九：大数据专业人士将拥抱未来

大数据国内外现状

美国：美国政府在2012年3月29日宣布投资两亿美元拉动大数据相关产业发展，将“大数据战略”上升为国家意志。

中国：中国商业联合会：副会长刘建沪介绍说，随着互联网的快速发展，中国的电子商务企业纷纷组建了数据分析部门。

2011年10月，工信部确认京沪深杭等5城市为“云计算中心”试点城市。而真正的问题或许不在于怎样建设“云计算中心”。国家信息中心常务副主任杜平直言不讳：“应对大数据的到来，需要不断建基础设施，但是建了干什么，有些数据需要存储，也有很多数据可能不需要储存。”

大数据的市场有多大？中央财经大学中国经济管理研究院博士张永力说，国外大数据行业约有1000亿美元的市场，而且每年都以10%的速度在增长，增速是软件行业的两倍。

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；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。

人类从依靠自身判断做决定到依靠数据做决定的转变，也是大数据作出的最大贡献之一。

——《大数据时代》

挑战

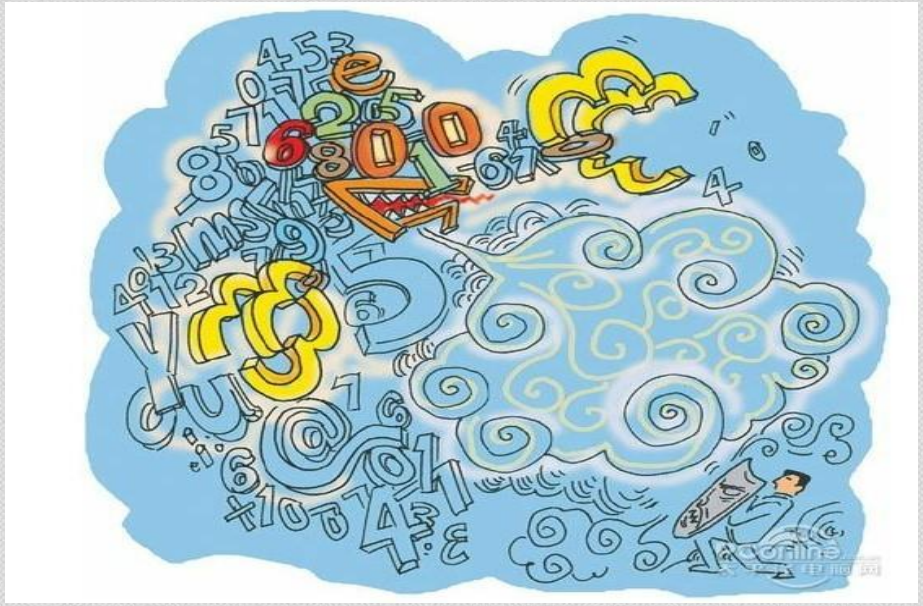
诸多领域的问题亟待解决，最重要的是每个人的信息都被互联网所记录和保留了下来，并且进行加工和利用，为人所用，而这正是我们所担忧的信息安全隐患！

更多的隐私、安全性问题：我们的隐私被二次利用了

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

…



眼下中国互联网热门的话题之一就是互联网实名制问题，我愿意相信这是个好事。毕竟我们如果明着亮出自己的身份，互联网才能对我们的隐私给予更好保护。

“大数据为新财富，价值堪比石油”

大数据时代！你准备好了没有