物联网架构与技术IOT-6

第 6章 物联网的信息处理技术（软件、云与大数据、数字孪生）

北京交通大学计算机学院网络管理研究中心 刘峰
fliu@bjtu.edu.cn

2023年12月11日

IoT体系结构

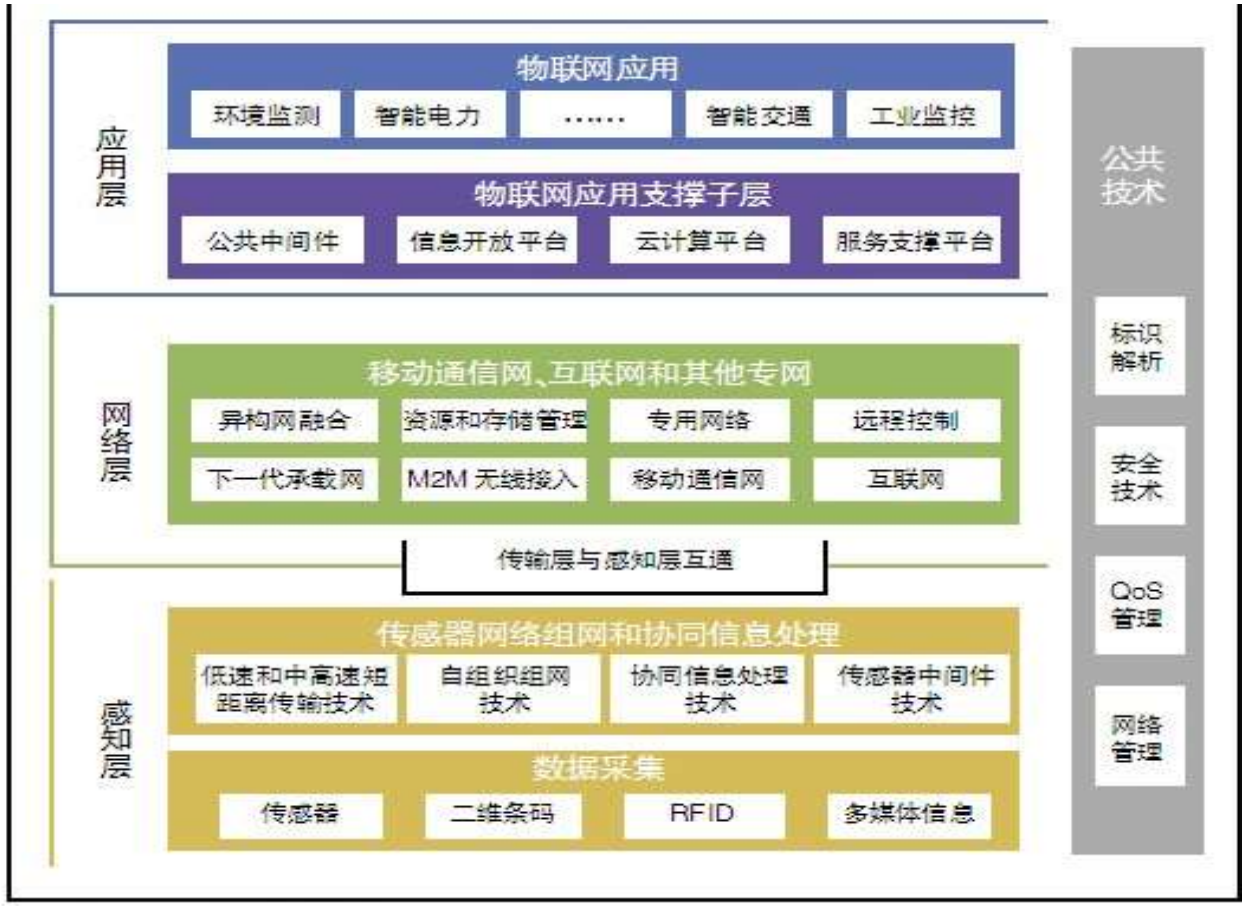


图 2:物联网的技术体系框架

本章概要

- 软件技术
- 云计算技术
- 大数据技术

- 人工智能
- 数字孪生

大数据技术代表了云计算技术的最新发展动态，也是软件技术的最新潮流。

软件技术

物联网软件技术涉及众多内容，可参照教材第六章内容，主要包括：

（1）基于服务器和客户端的应用程序
（2）基于JAVA的WEB应用程序设计
（3）OSGi服务平台
（4）搜索引擎Lucene
（5）基于多智能体的应用程序
（6）规则引擎
（7）移动终端开发技术
（8）Java Card

信息技术发展特征

信息技术是当今世界创新速度最快、通用性最广、渗透性最强的高技术，信息科技领域的创新能力和发展水平是国家创新能力的突出体现。

信息技术的云-网-端领域的创新技术层出不穷

信息科学技术正在进入一个转折期，可能出现重大的技术变革，与以往相比，当代信息科技创新更加活跃，信息技术的云-网-端领域的创新技术层出不穷，物联网、云计算、大数据等新兴技术将深刻地影响未来社会发展的模式。面向2020年，中国必须在战略高度上将信息科技发展作为主战场之一，在新科技革命进程中，加速人-机-物三元世界的融合发展。(李国杰院士)

云计算技术及其架构

- 云计算（cloud computing）是基于互联网的相关服务的增加、使用和交付模式，通常涉及通过互联网来提供动态易扩展且经常是虚拟化的资源。云是网络、互联网的一种比喻说法,表示互联网和底层基础设施的抽象。因此，云计算甚至可以让你体验每秒10万亿次的运算能力，拥有这么强大的计算能力可以模拟核爆炸、预测气候变化和市场发展趋势。用户通过电脑、笔记本、手机等方式接入数据中心，按自己的需求进行运算。
- [
- 美国国家标准与技术研究院（NIST）定义：云计算是一种按使用量付费的模式，这种模式提供可用的、便捷的、按需的网络访问， 进入可配置的计算资源共享池（资源包括网络，服务器，存储，应用软件，服务），这些资源能够被快速提供，只需投入很少的管理工作，或与服务供应商进行很少的交互。

云计算发展简史

- 1983年，太阳电脑（Sun Microsystems）提出“网络是电脑”（“The Network is the Computer”）,2006年3月，亚马逊（Amazon）推出弹性计算云（Elastic Compute Cloud；EC2）服务。
- 2006年8月9日，Google首席执行官埃里克·施密特（Eric Schmidt）在搜索引擎大会（SES San Jose 2006）首次提出“云计算”（Cloud Computing）的概念。Google“云端计算”源于Google工程师克里斯托弗·比希利亚所做的“Google 101”项目。
- 2007年10月，Google与IBM开始在美国大学校园，包括卡内基梅隆大学、麻省理工学院、斯坦福大学、加州大学柏克莱分校及马里兰大学等，推广云计算的计划，这项计划希望能降低分布式计算技术在学术研究方面的成本，并为这些大学提供相关的软硬件设备及技术支持（包括数百台个人电脑及BladeCenter与System x服务器，这些计算平台将提供1600个处理器，支持包括Linux、Xen、Hadoop等开放源代码平台）。而学生则可以通过网络开发各项以大规模计算为基础的研究计划。
- 2008年1月30日，Google宣布在台湾启动“云计算学术计划”，将与台湾台大、交大等学校合作，将这种先进的大规模、快速将云计算技术推广到校园。

云计算发展简史

- 2008年2月1日，IBM（NYSE: IBM）宣布将在中国无锡太湖新城科教产业园为中国的软件公司建立全球第一个云计算中心（Cloud Computing Center）。
- 2008年7月29日，雅虎、惠普和英特尔宣布一项涵盖美国、德国和新加坡的联合研究计划，推出云计算研究测试床，推进云计算。该计划要与合作伙伴创建6个数据中心作为研究试验平台，每个数据中心配置1400个至4000个处理器。这些合作伙伴包括新加坡资讯通信发展管理局、德国卡尔斯鲁厄大学Steinbuch计算中心、美国伊利诺伊大学香宾分校、英特尔研究院、惠普实验室和雅虎。
- 2008年8月3日，美国专利商标局网站信息显示，戴尔正在申请“云计算”（Cloud Computing）商标，此举旨在加强对这一未来可能重塑技术架构的术语的控制权。
- 2010年3月5日，Novell与云安全联盟（CSA）共同宣布一项供应商中立计划，名为“可信任云计算计划（Trusted Cloud Initiative）”。
- 2010年7月，美国国家航空航天局和包括Rackspace、AMD、Intel、戴尔等支持厂商共同宣布“OpenStack”开放源代码计划，微软在2010年10月表示支持OpenStack与Windows Server 2008 R2的集成；而Ubuntu已把OpenStack加至11.04版本中。
- 2011年2月，思科系统正式加入OpenStack，重点研制OpenStack的网络服务。

云计算架构


云计算参考架构
客户端
Client
PC
浏览器
移动终端设备
互联网
软件即服务
SaaS
OA
CRM
门户
即时通信
邮件
内容管理
行业应用
企业短信
软件开发
手机音乐
其他
平台即服务
PaaS
身份认证管理
服务总线
工作流
手机支付
报表、数据挖掘
...
基础设施
即服务
IaaS
基础服务
数据存储
计算服务
负载管理
备份
基础设施
虚拟化
主机
存储
网络
其它硬件
管理和业务支撑
开发工具

云计算服务方式

云计算可以认为包括以下几个层次的服务：基础设施即服（IaaS），平台即服务（PaaS）和软件即服务（SaaS）。

IaaS：基础设施即服务

- IaaS(*Infrastructure-as-a- Service*)：基础设施即服务。消费者通过*Internet*可以从完善的计算机基础设施获得服务。例如：硬件服务器租用。

PaaS：平台即服务

- PaaS(*Platform-as-a- Service*)：平台即服务。*PaaS*实际上是指将软件研发的平台作为一种服务，以*SaaS*的模式提交给用户。因此，*PaaS*也是*SaaS*模式的一种应用。但是，*PaaS*的出现可以加快*SaaS*的发展，尤其是加快*SaaS*应用的开发速度。例如：软件的个性化定制开发。

SaaS：软件即服务

- SaaS(*Software-as-a- Service*)：软件即服务。它是一种通过*Internet*提供软件的模式，用户无需购买软件，而是向提供商租用基于*Web*的软件，来管理企业经营活动。例如：阳光云服务器。

云计算特点

(1) 超大规模 Google云计算拥有100多万台服务器， Amazon、IBM、微软、Yahoo等均拥有几十万台服务器。企业私有云一般拥有数百上千台服务器。“云”能赋予用户前所未有的计算能力。

(2) 虚拟化 云计算支持用户在任意位置、使用各种终端获取应用服务。

(3) 高可靠性 数据多副本容错、计算节点同构可互换等措施保障服务高可靠性。

(4) 通用性 云计算不针对特定的应用，可以同时支撑不同的应用运行。

(5) 高可扩展性 规模可以动态伸缩，满足应用和用户规模增长的需要。

(6) 按需服务 是一个庞大的资源池，按需购买；可以像自来水，电，煤气那样计费。

(7) 极其廉价 由于“云”的特殊容错措施可以采用极其廉价的节点来构成云，“云”的自动化集中式管理使大量企业无需负担日益高昂的数据中心管理成本。

(8) 潜在的危险性 云计算服务除了提供计算服务外，还必然提供了存储服务。但是云计算服务当前垄断在私人机构（企业）手中，而他们仅仅能够提供商业信用。对于政府机构、商业机构（特别像银行这样持有敏感数据的商业机构）对于选择云计算服务应保持足够的警惕。

云计算带来新的影响：
软件开发、架构

- 云计算环境下，软件技术、架构将发生显著变化。首先，所开发的软件必须与云相适应，能够与虚拟化为核心的云平台有机结合，适应运算能力、存储能力的动态变化;二是要能够满足大量用户的使用，包括数据存储结构、处理能力;三是要互联网化，基于互联网提供软件的应用;四是安全性要求更高，可以抗攻击，并能保护私有信息，五是可工作于移动终端、手机、网络计算机等各种环境。
- 云计算环境下，软件开发的环境、工作模式也将发生变化。虽然，传统的软件工程理论不会发生根本性的变革，但基于云平台的开发工具、开发环境、开发平台将为敏捷开发、项目组内协同、异地开发等带来便利。软件开发项目组内可以利用云平台，实现在线开发，并通过云实现知识积累、软件复用。
- 云计算环境下，软件产品的最终表现形式更为丰富多样。在云平台上，软件可以是一种服务，如SAAS，也可以就是一个Web Services，也可能是可以在线下载的应用，如苹果的在线商店中的应用软件，等

云计算带来新的影响：软件测试

- 在云计算环境下，由于软件开发工作的变化，也必然对软件测试带来影响和变化。
- 软件技术、架构发生变化，要求软件测试的关注点也应做出相对应的调整。软件测试在关注传统的软件质量的同时，还应该关注云计算环境所提出的新的质量要求，如软件动态适应能力、大量用户支持能力、安全性、多平台兼容性等。
- 云计算环境下，软件开发工具、环境、工作模式发生了转变，也就要求软件测试的工具、环境、工作模式也应发生相应的转变。通过云构建测试环境;软件测试也应该可以通过云实现协同、知识共享、测试复用。
- 软件产品表现形式的变化，要求软件测试可以对不同形式的产品进行测试，如Web Services的测试，互联网应用的测试，移动智能终端内软件的测试等。

大数据技术及其架构

- **大数据**（big data），指无法在可承受的时间范围内用常规软件工具进行捕捉、管理和处理的数据集合。

- 大数据指不用随机分析法（抽样调查）这样的捷径，而采用所有数据进行分析处理。

- 大数据的5V特点：Volume（大量）、Velocity（高速）、Variety（多样）、Value（价值）Veracity（真实性）

大数据定义

- 研究机构Gartner给出了这样的定义。“大数据”是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。

- 麦肯锡全球研究所给出的定义是：一种规模大到在获取、存储、管理、分析方面大大超出了传统数据库软件工具能力范围的数据集合，具有海量的数据规模、快速的数据流转、多样的数据类型和价值密度低四大特征。

- 大数据技术的战略意义不在于掌握庞大的数据信息，而在于对这些含有意义的数据进行专业化处理。换言之，如果把大数据比作一种产业，那么这种产业实现盈利的关键，在于提高对数据的“加工能力”，通过“加工”实现数据的“增值”。

大数据与云计算的关系

- 从技术上看，大数据与云计算的关系就像一枚硬币的正反面一样密不可分。大数据必然无法用单台的计算机进行处理，必须采用分布式架构。它的特色在于对海量数据进行分布式数据挖掘，但它必须依托云计算的分布式处理、分布式数据库和云存储、虚拟化技术。

- 随着云时代的来临，大数据（Big data）也吸引了越来越多的关注。大数据（Big data）通常用来形容一个公司创造的大量非结构化数据和半结构化数据，这些数据在下载到关系型数据库用于分析时会花费过多时间和金钱。大数据分析常和云计算联系到一起，因为实时的大型数据集分析需要像MapReduce一样的框架来向数十、数百或甚至数千的电脑分配工作。

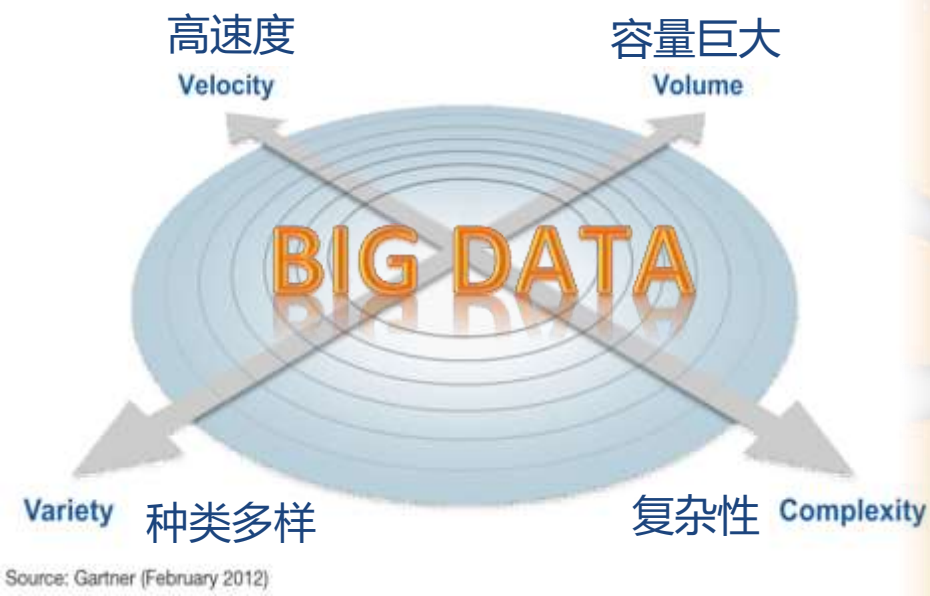
大数据技术

- TERADATA
- CLOUDERA大数据技术架构

TERADATA观点

- 大数据的定义及特征
- 大数据分析的原理及架构
- 如何在行业中适应产业转型运用大数据实现智慧型信息服务

大数据的定义及特征——什么是大数据: 大数据遗失的“V”



社会管理

欧洲公共事业管理

- 每年创造价值2千5百亿欧元
- 年均生产率提高0.5%

社会生产

制造业

- 产品开发组装成本下降50%
- 运营成本下降7%

社会服务

通信运营商

- 提高客户保有率30%
- 降低手机欺诈比例1%

大数据的最重要的“V”往往被忽视—最关键的“V”是价值（Value）！

KB MB GB Terabytes Petabytes Exabytes Zettabytes Yottabytes

大数据的定义及特征——大数据分析的特征

世间万物的数据化：一切皆可量化

人与人

物与物

人与物

多维度的信息演变

复杂的数据来源

广泛的业务范围

信息的快速响应

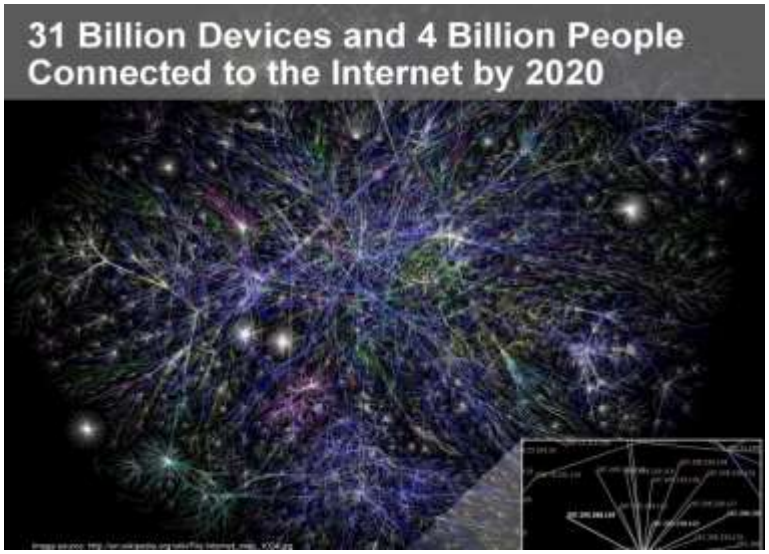
复杂的数据管理

高性能的分析要求

议题

- 大数据的定义及特征
- 大数据分析的原理及架构
- 如何在行业中适应产业转型运用大数据实现智慧型信息服务

大数据分析的原理及架构——大数据分析的原理

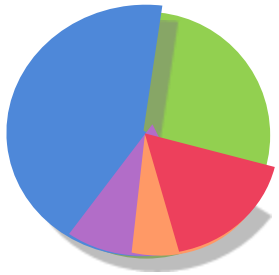


- 推荐 “买了这项产品的人也买了那样的产品”
- 推送 “你可能认识的人”服务: A识B, B识C, AC非常可能认识
- 自动将 “和你一样喜欢B电影的人,也喜欢C电影”推入WishList
- 推送“你是不是也想搜索相关的…”
- 流失率，介绍扩散率， 平均用户贡献度

运营智能

主动事件
我希望发生什么情况?

运营支撑
正在发生什么情况?



加速

互动流程自动化

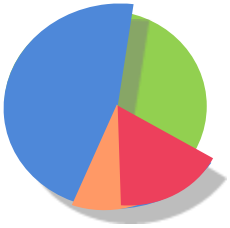
事件式主动触发为主导

打通运营系统

连续更新和流程互动

战术智能

预测
将要发生什么情况?



预测模型

分析
为何发生了这种情况?

即席查询、BI工具

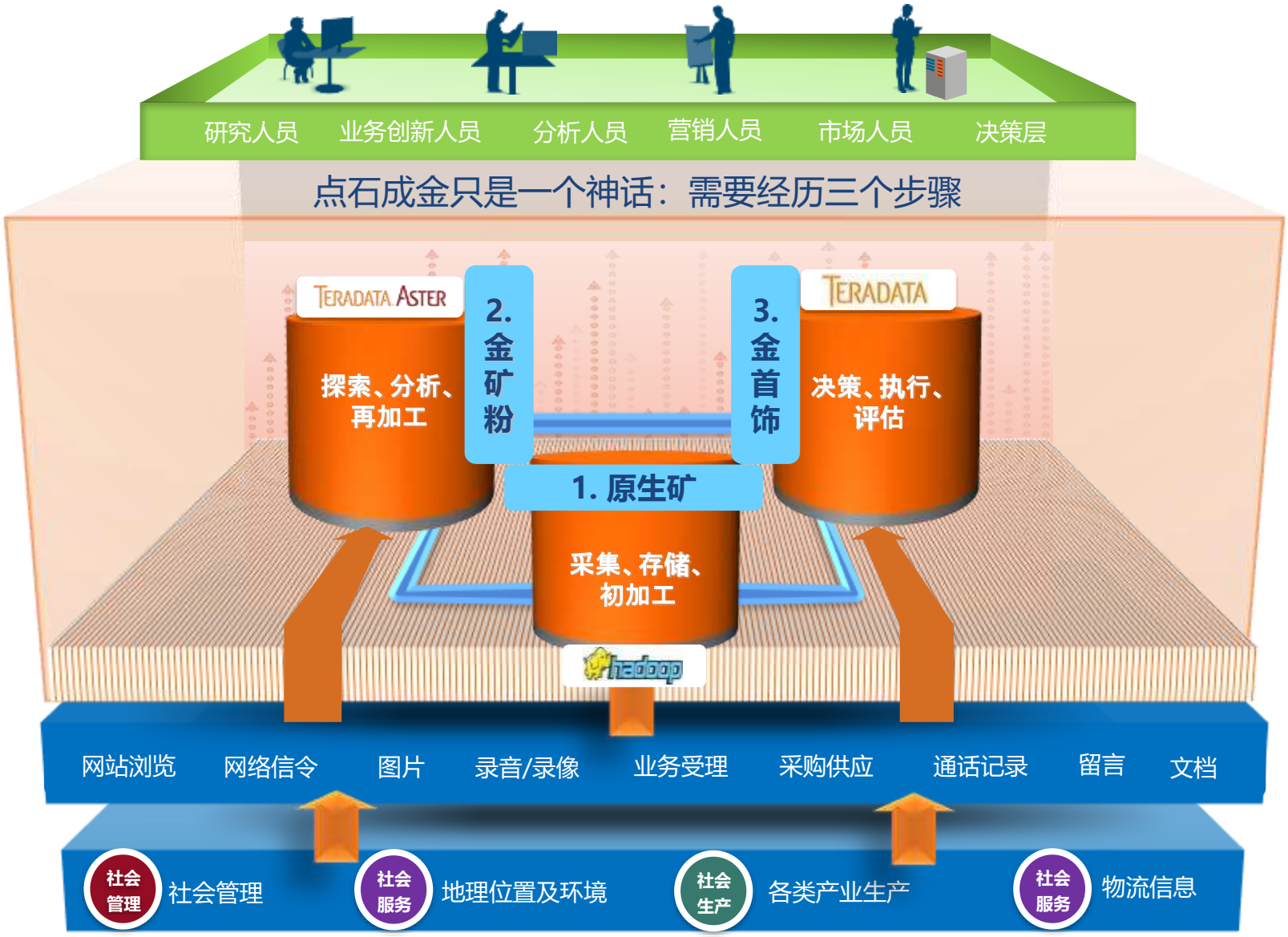
报表
发生了什么情况?

批处理报表

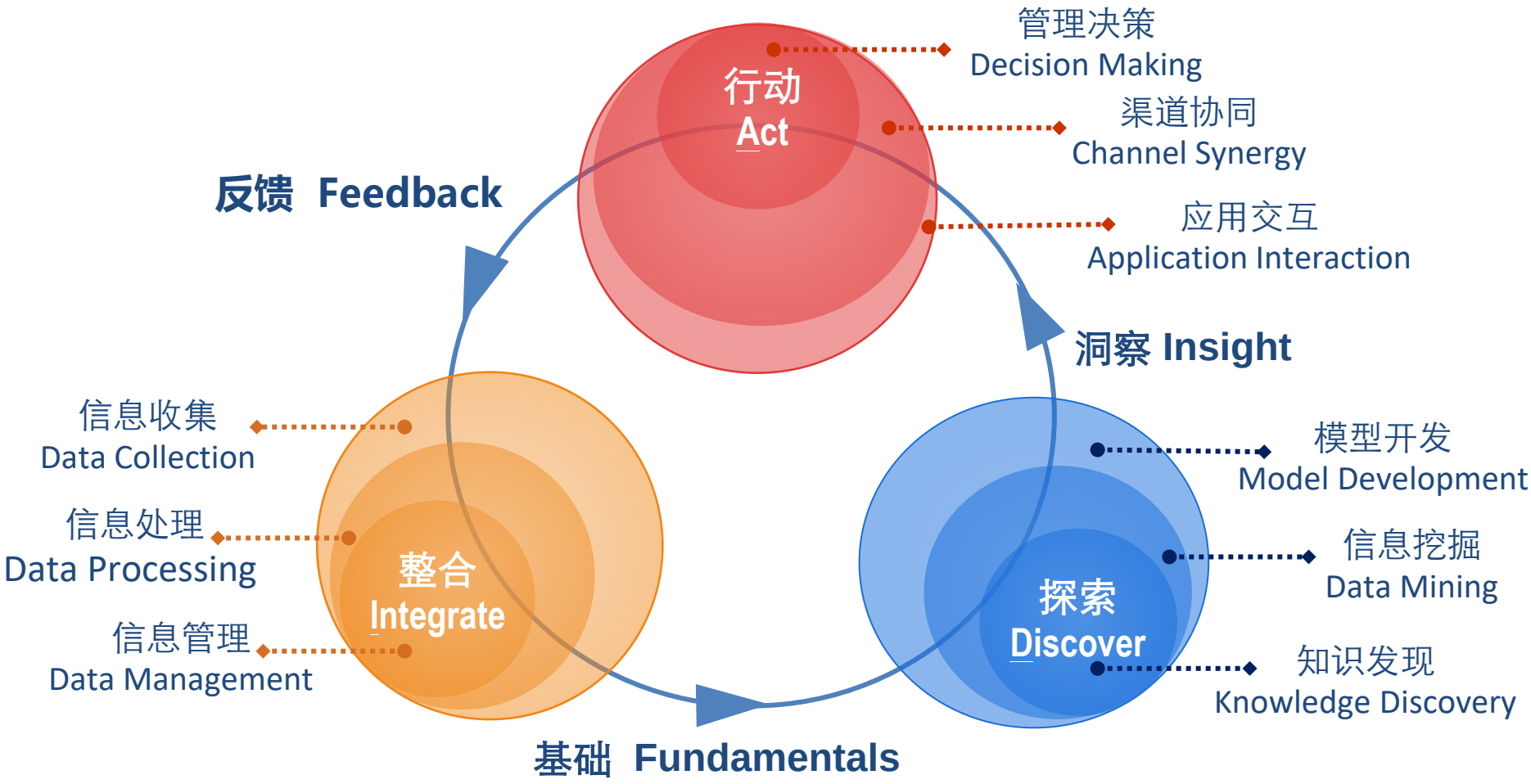
延展

- 事件式触发
- 持续的更新
  简短的战术性查询
- 分析
- 即席查询
- 批处理

大数据分析的原理及架构——大数据分析的架构



大数据分析的原理及架构——如何最大化大数据的价值



大数据分析的原理及架构——如何最大化大数据的价值

1

收集一些大数据，着手进行探索，重点在于快速产生重要价值

2

与其他传统数据结合才能产生更大价值

3

使大数据成为战略决策的一个重要层面

4

统一数据格式、输入及使用，逐步建立共享的大数据标准

大数据时代，分析平台存储成本≠运营成本

谁是企业的关键竞争对手？
如何与互联网企业竞合？

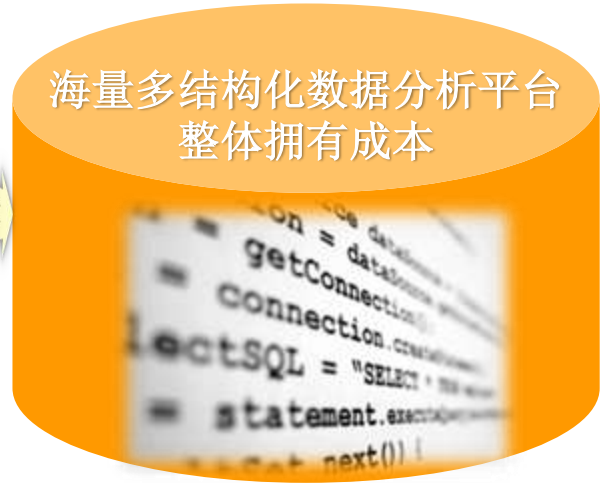
4G LTE会给
市场带来什么变化？

大数据？小时代？
什么才是持之以恒的
数据分析生产力？

大数据时代创新的
数据分析应用模式如何？

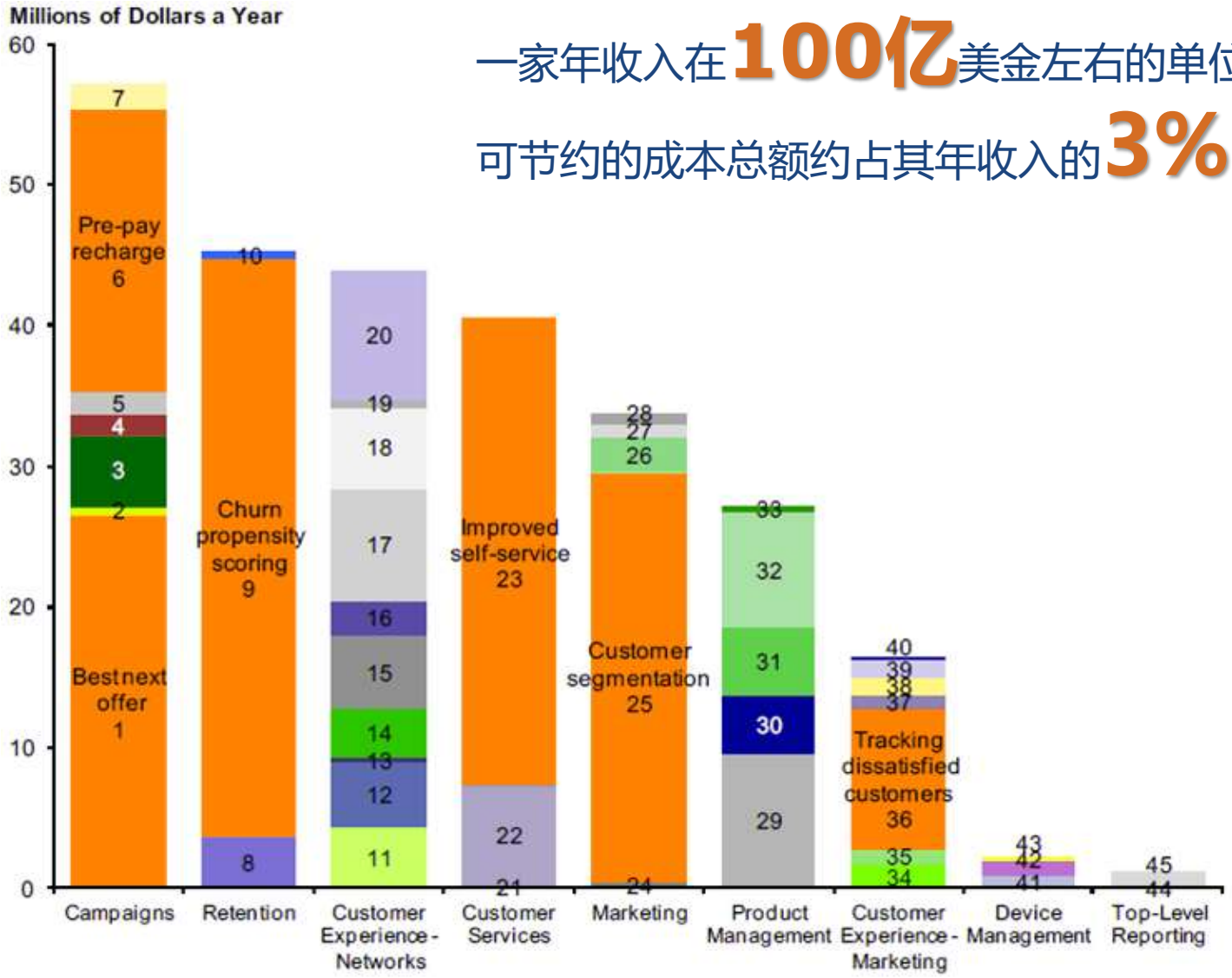
- 平台运行能耗成本
- 平台占地成本
- 人员成本
- 培训成本
- 系统维护成本
- 系统稳定保障成本

运营成本 + 存储成本



海量多结构化数据分析平台的经济衡量指标应该是**“整体拥有成本”**

大数据分析助力政府和企业节约成本



一家年收入在**100亿**美金左右的单位，通过大数据分析，可节约的成本总额约占其年收入的**3%**。

Gartner Group, May-2013

议题

- 大数据的定义及特征
- 大数据分析的原理及架构
- 如何在行业中适应产业转型运用大数据实现智慧型信息服务

大数据分析助力产业结构调整和转型

- 基于大数据，逐步开放公共数据，打造透明政府
- 监控重大社会事件，构建科学预警监控体系

1. 城市规划
2. 税收
3. 公共安全
4. 电子政务 ……

社会管理

社会生产

- 基于大数据，促进社会应用创新，推动中国制造走向中国创造

1. 一产：农业，包括种植和养殖
2. 二产：工业，包括制造，加工

大数据大网络

社会服务

- 实施“循数管理”，提高科学决策能力

1. 教育
2. 医疗
3. 通信
4. 公共交通
5. 餐饮
6. 养老
7. 社区
8. 零售 ……

FedEx选择适应其企业发展需求的数据分析

企业战略目标

- FedEx提供隔夜快递、地面快递、重型货物运送、文件复印及物流服务，总部设于美国田纳西州。提供涵盖运输、电子商务和商业运作等。
- 全球约14万名员工，每个工作日约330万件包裹，654架飞机，大致有44,000辆专用货车。
- 原有数据库已经无法满足企业的数据分析需求，选用Teradata构建其企业级数据仓库。

企业商业智能（BI）

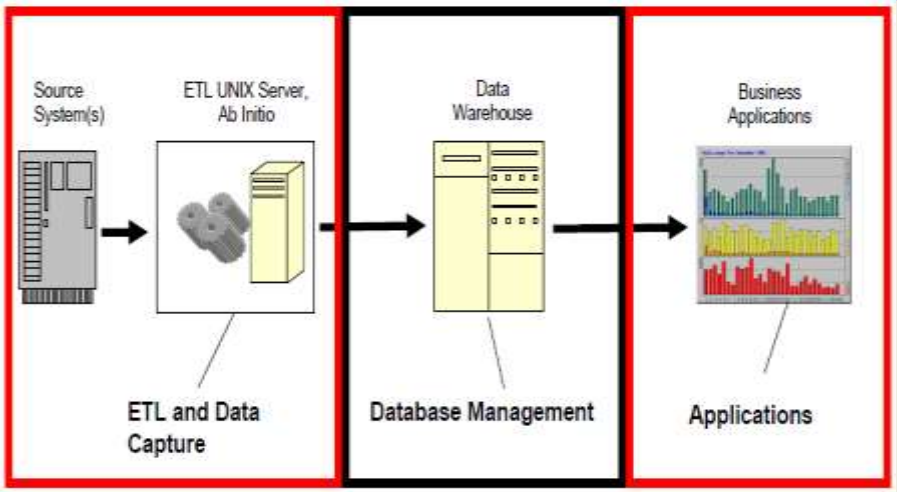
FedEx集成所有的明细数据到Teradata数据仓库中，提供行包流转的复杂流程的的完整视图，通过集成电子投递记录和产品扫描信息。
FedEx借助分析提高以下业务能力：路径计划优化，提高客户贡献度，提高服务效率，提高产能和收益。

- 7000多个数据仓库用户：40多个VP，7个运营公司，5个国际区域。
- 每月160万即席查询。
- 加载数十个数据源的数据。

商业智能实践效果

- 为客户提供集成的业务方案库，提供物流、仓储、后勤和供应链分析方案的单一访问入口。
- 提供在线行包跟踪和服务信息的单一Web门户
- 每日加载400万行包运输信息 。
- 提供360°客户统一视图，集成了所有客户的接触点，所有客户接触点信息共享。

Teradata解决方案



DHL利用数据分析从企业内部注入活力

企业战略目标

- DHL是全球物流行业的市场领导者。DHL致力于在国际包裹、快递、空运及海运、公路和铁路运输,合同物流及国际邮政服务方面提供专业化的服务。
- DHL的业务遍布全球220的国家和地区，是全球国际化程度最高的公司。DHL在全球的员工人数超过285,000人，DHL几乎能够为各种物流需求提供完美的解决方案。

企业商业智能（BI）

- **商业应用**： 收入调节、市场营销分析、销售绩效分析、 趋势（客户、产品和销售业绩）、成本分析（回扣、折扣及其他可调节收入）、收益率分析、按需提供的客户报告、索赔管理。
- **市场与营销**： 收入与效益、业务量管理、销售机会管理、活动管理（活动、电话访问、登门拜访）
- **业务运营**： 全方位跟踪查询、事件分析、可视化服务绩效分析、货运趋势分析、假设分析等。

商业智能实践效果

Teradata解决方案

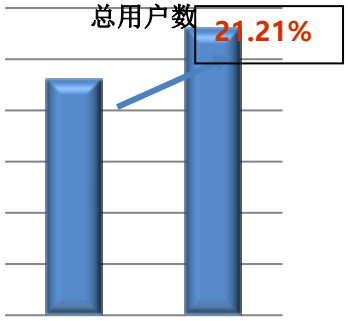
美国诺福克南方铁路公司Norfolk Southern

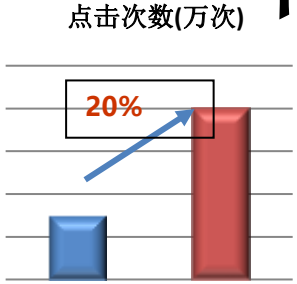
客户简介	业务过程	应用	效益	配置
• NS公司是1982年诺福克铁路公司与南方铁路公司合并组成的一级铁路公司，公司经营范围跨越美国东部22个州、哥伦比亚区以及安大略湖，线路连接22个港口，提供完善的物流和多式联运等服务。	• 运营管理 • 客户管理 • 财务管理 • 数据管理	• 资源管理和利用 • 实时运营支持 • 成本管理 • 利润分析 • 人员排班 • 货流监控	• 客户电话呼入降低35% • 通过提供更多的“car hire” 决策支持，为公司节省300万美金 • 超过1500个外部客户访问Teradata中的货物运输信息 • 1000多个内部用户和管理者得到了运营决策支撑，包括交通流量和减少拥堵等方面 • 非常高的系统访问速度和效率—用户访问一张3千万的表在几秒钟内返回结果	• 34TB用户数据 • 约1000多内部用户

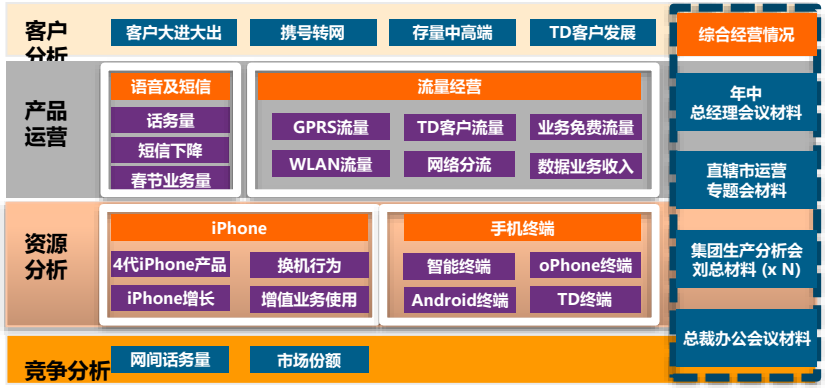
美国联合太平洋铁路公司（Union Pacific）

客户简介	业务过程	应用	效益	配置
• 太平洋联合铁路公司是现今全美国最大的铁路网络之一，横跨美国中部及西部地区，太平洋联合铁路公司领导着美国的运输业，在美国的西部和加拿大、墨西哥长达36000英里的铁路线上运输煤、化学制品。	• 运营管理 • 客户管理 • 财务管理 • 数据管理	• 定价管理 • 机车利用 • 计费对账 • 联合运输管理 • 车厢管理和预测 • 法律诉讼支撑 • 个人信息管理 • 业务分析 • 收入变化分析 • 通道流量分析 • 成本和利润评估 • 市场细分和客户管理	• 平均每年增加1亿美金收入 • 新服务内容 • 更加有竞争力的定价机制 • 增加收入 • 更好的资源利用 • 缩短收账周期 • 挖掘新业务	• 60TB用户空间 • 约2000多张业务表 • 每月有4000多个不同用户 • 每月350万用户查询 • 绝大数据数据每日夜间更新

中国移动集团锻造其企业“数据文化”







数据分析系统现状

- 中国移动集团于2004年建设一经系统，经6期工程建设，裸数据容量将达到1107TB。
- 面向全网运营管理相关构建应用支撑，并向市场部、数据部和终端部等业务部门提供分析服务。
- 目前一经系统经日接口数据量1100GB， 小时级作业13个，日作业7,615个，月接口数据量量4,144GB，作业17,759个，所有作业均按时完成，满足分析时效要求。

数据分析实践效果

- 目前一经系统建设了31个应用专区，共1083个具体应用；目前服务用户数1227个，平均日点击量354次。
- 根据各级领导需求，共完成117份分析报告，并针对各业务及管理部门实现621项专题。
- 企业运营效率明显得到提升，提升率可到200%。
- 运营成本得到明显的下降，约下降50%。
- 助力企业发展，利润增长7%，数据服务增长17%，数据业务中无线上网业务增长51%，移动用户增长3%。

Teradata助力国内通信运营商建设大数据平台项目—应用场景

序号	应用项目	应用项目具体场景	选择理由
1	客户交往圈应用	客户交往圈分析 校园市场识别模型 家庭V网识别模型 流动客户识别模型	利用Teradata高效数据处理技术和交往圈分析模型，实现精确识别，同时缓解主库压力，实现经分整体提速。
2	流量经营应用	流量数据处理（WAP话单等） 中文文本分词 信息分类 流量监控与分析、内容监控与分析等	Teradata易用易开发的相关的非结构化数据处理函数，支持中文分词、支持大规模流量数据处理和分析。
3	电子渠道及互联网信息 处理与分析	电子渠道日志等数据处理 网站点击行为分析、黄金路径分析、产品优化等	Teradata易用易开发的相关的非结构化数据处理函数、路径分析函数等，针对本类应用处理性能尤为突出。
4	产品关联分析	如：产品关联分析、购物篮分析	Teradata易用易开发的产品关联函数针对本类应用处理性能尤为突出。

大数据助力汽车制造行业，改进汽车质量

提升运营效率与客户满意度

探索分析

- 早期预警
- 根源分析
- 欺诈检测
- 存货分析
- 成本分析
- 政策分析

决策执行

- 法规遵从
- 召回管理
- 客户满意度
- 成本回收和供应商绩效

大数据处理平台

决策树　　生存模型　　关联分析

产品数据	产品质量数据	市场研究数据	用户交互与服务数据
物料清单	投诉	第三方市场研究	在线论坛
流程和质量	场地反馈	NCB调查	电子邮件电话记录
操作员记录	保修记录	产品报告	服务日志

大数据助力航空业，提升客户忠诚度和个性化营销体验

- 客户体验的多维度
- 以客户为中心的整体的360度视角

- 个人价值评分
  - 频次、贡献度
  - 盈利产品、社交网络

- 整体质量的提升
- 基于事件的沟通、互动与管理

- 市场归属统计
- 行为预测与路径分析

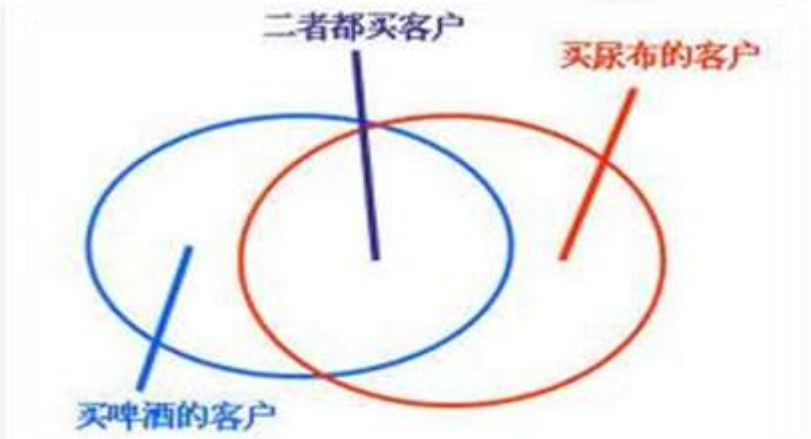
1 关注客户体验　2 衡量客户价值　3 相关联的客户价值及体验　4 驱动价值成长

生产　财务　常旅客 FFP　全市场数据　客户关系管理

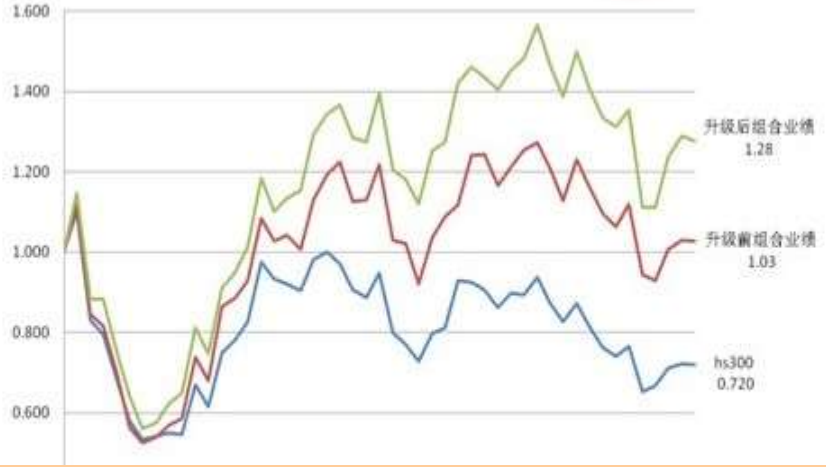
电子商务　社交媒体　QAR监控　传感器

大数据分析的发现： 手电筒与草莓夹心饼干

WAL★MART®

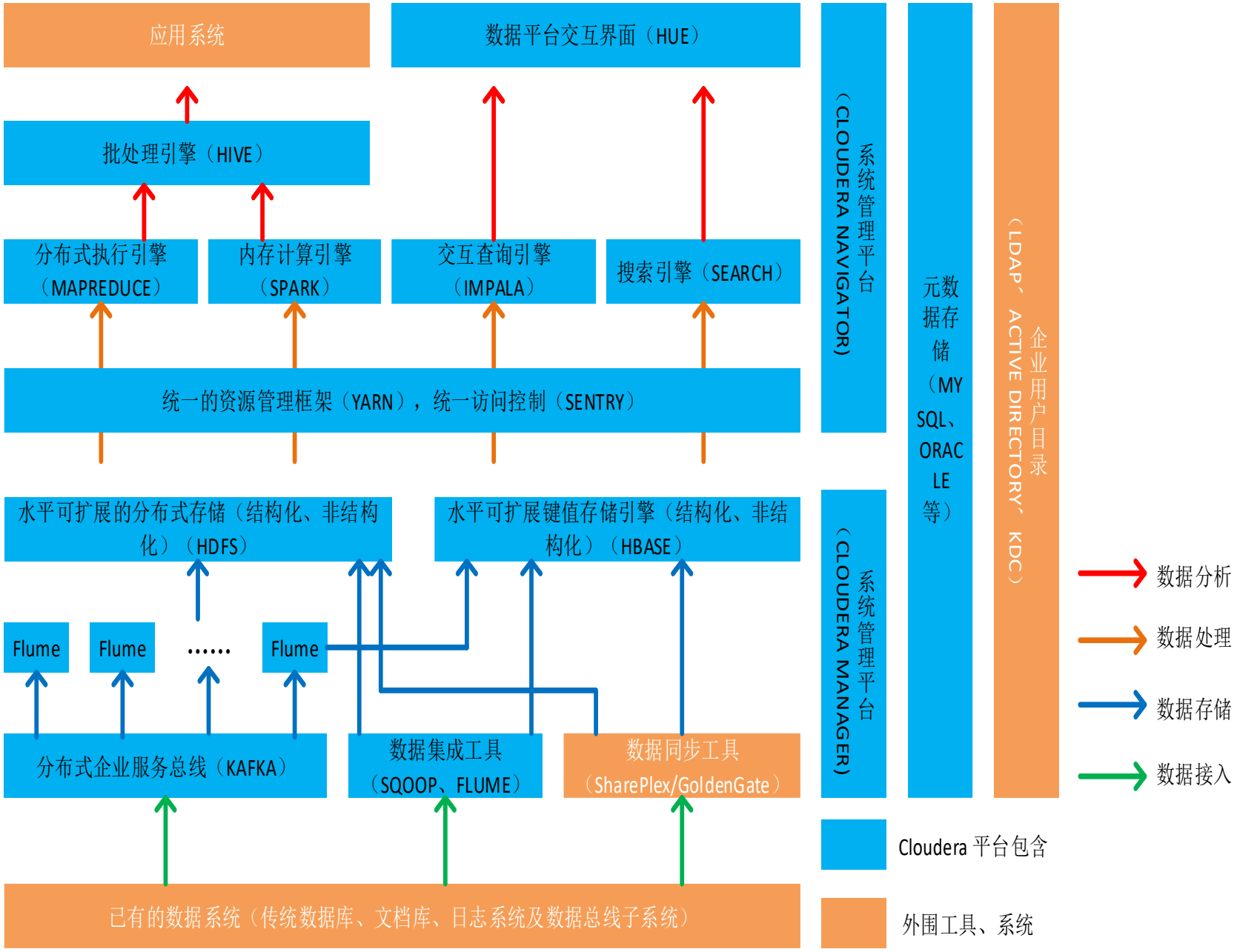


一般有婴儿的家庭，母亲在家中照看孩子，父亲前去超市购物，在购买尿布的同时往往顺便为自己购买啤酒。因此，**“尿布”和“啤酒”常出现在同一购物货架区内**



当飓风到临前，最畅销的商品除了矿泉水、手电筒、电池等必须品和啤酒之外， 还有**草莓夹心饼干-销量是平常的7倍!**

CLOUDERA大数据技术架构


应用系统
数据平台交互界面（HUE）
批处理引擎（HIVE）
分布式执行引擎（MAPREDUCE）
内存计算引擎（SPARK）
交互查询引擎（IMPALA）
搜索引擎（SEARCH）
统一的资源管理框架（YARN），统一访问控制（SENTRY）
水平可扩展的分布式存储（结构化、非结构化）（HDFS）
水平可扩展键值存储引擎（结构化、非结构化）（HBASE）
Flume
Flume
……
Flume
分布式企业服务总线（KAFKA）
数据集成工具（SQOOP、FLUME）
数据同步工具（SharePlex/GoldenGate）
已有的数据系统（传统数据库、文档库、日志系统及数据总线子系统）
系统管理平台（CLOUDERA NAVIGATOR）
系统管理平台（CLOUDERA MANAGER）
元数据存储（MYSQL、ORACLE等）
企业用户目录（LDAP、ACTIVE DIRECTORY、KDC）
数据分析
数据处理
数据存储
数据接入
Cloudera 平台包含
外围工具、系统

软件体系架构

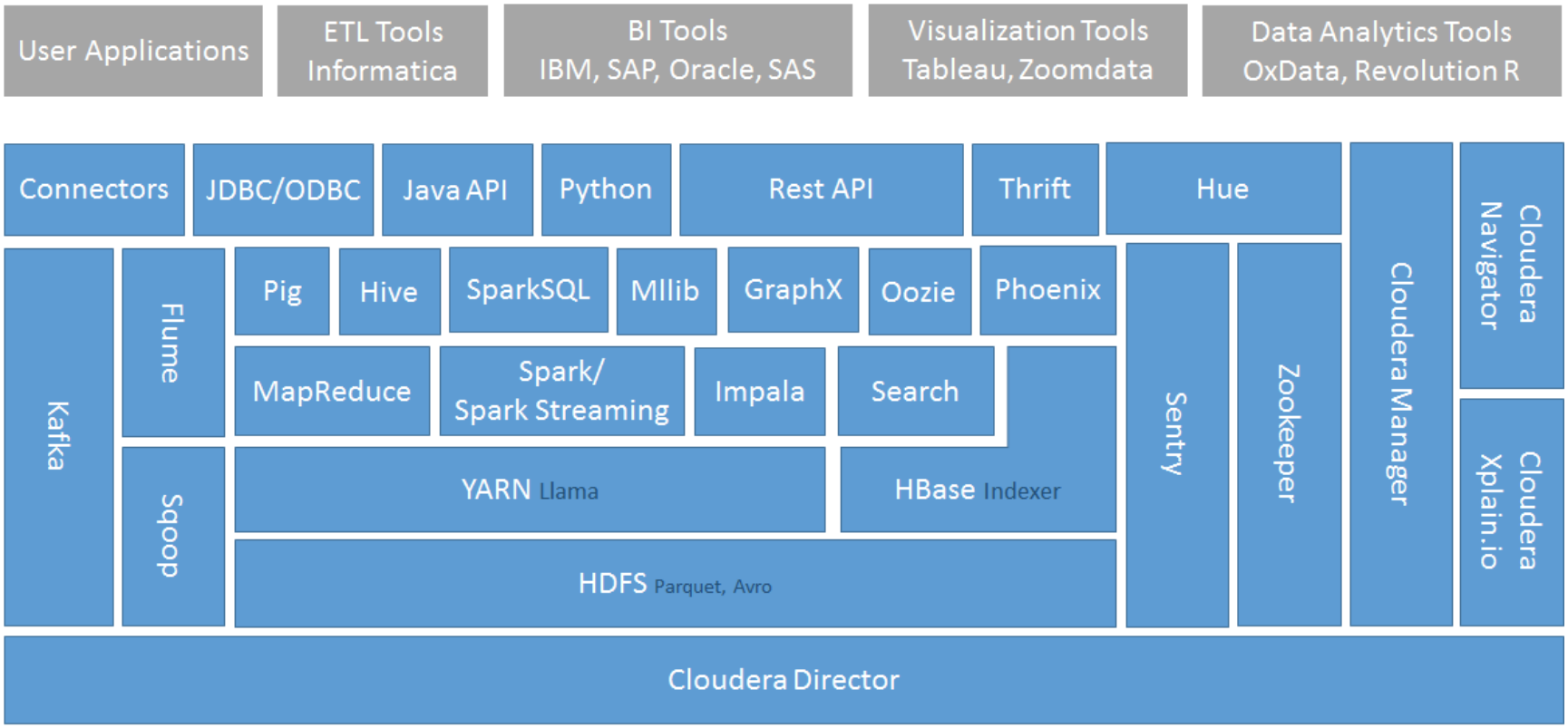
- Cloudera平台包含的分布式企业服务总线Kafka，数据库导入导出工具Sqoop以及实时日志采集组件Flume，再辅以一些商业的数据库增量同步工具，比如GoldenGate，SharePlex，使得大数据平台和传统数据平台能进行无缝对接。
- 基于HDFS和HBase构建的存储系统可以实现高度水平可扩展性，同时可以承载不同类型的数据，结构化，半结构化以及非结构化的数据，将之前割裂的数据系统进行融合存储，极大方便后期的关联数据分析。
- 在存储层之上，平台同样提供了多种高度可扩展的分布式处理组件，包括MapReduce，分布式内存计算引擎Spark，交互式SQL引擎Impala以及搜索引擎Cloudera Search，为融合数据平台上的分析提供工具支撑
- 为方便基于平台的开发，并和其他系统的集成，平台提供易于开发的接口，如SQL，JAVA，RestAPI，Python等等；同时也有Hue这样的工具提供了界面化操作。

平台的管理和安全

- 软件平台提供了业界最流行的管理工具Cloudera Manager，平台的部署、配置、监控及诊断等都可以自动化完成
- 丰富的安全组件给平台提供细粒度的安全控制
  - 多层级的身份认证（Cloudera Manager, Kerberos, AD, Hue）
    - 管理平台，运维人员，客户端，BI工具
  - 统一的授权访问控制（Apache Sentry）
    - 在平台上提供统一的访问安全控制策略
  - 数据保护（HDFS At-Rest Encryption, Navigator Encrypt, Navigator KeyTrustee）
    - On-the-wire和at-rest数据保护，并内置有Key Management方案
  - 全面的审计（Cloudera Navigator）
    - 不管以什么方式进行访问集群，都会得到审计

平台数据治理

- Cloudera Navigator为平台的数据运维人员提供了全面的、自动化的数据治理方案
  - 数据发现和探索
    - 快速检索相关数据，加速数据发现流程
    - 自动发现元数据并允许用户自定义可定制化标签与注释，便于数据追踪与归类
  - 数据溯源
    - 帮助用户直观理解数据集的上下游血脉关系，验证数据源头与数据演变过程
    - 可以导出数据溯源信息到其他的溯源信息管理系统中
  - 生命周期管理
    - 定义并自动化复杂的数据生命周期管理工作，包括分类，保留及加解密策略 – 一切都基于Navigator丰富的元数据管理能力


User Applications
ETL Tools
Informatica
BI Tools
IBM, SAP, Oracle, SAS
Visualization Tools
Tableau, Zoomdata
Data Analytics Tools
OxData, Revolution R
Connectors
JDBC/ODBC
Java API
Python
Rest API
Thrift
Hue
Cloudera Navigator
Kafka
Flume
Pig
Hive
SparkSQL
Mllib
GraphX
Oozie
Phoenix
Sentry
Zookeeper
Cloudera Manager
Cloudera Xplain.io
MapReduce
Spark/
Spark Streaming
Impala
Search
Sqoop
YARN Llama
HBase Indexer
HDFS Parquet, Avro
Cloudera Director
Legacy Systems (RDBMS/DW, Data Mart, Archive, ESB, Log Data), Sensors, Internet


向其他系统提供数据服务
面向集团动车组信息服务（集团核心数据中心）
面向企业动车组信息服务（决策、应急指挥）
面向用户动车组信息服务（CRM）
面向协作企业信息服务（SRM）
基于大数据技术的高速动车组全生命周期数据集成化管理平台
数据集成管理
BW数据管理
销售/市场分析
项目管理
生产计划
客户关系管理
检修分析
质量分析
财务分析
资源计划
人力资源
行业指标
知识管理
故障数据挖掘
关联关系分析
专家知识
信息融合
BOM管理
BOM转换组件
BOM可视化组件
BOM视图映射
BOM版本控制器
数据引擎
部件基础库
设计资源库
制造资源库
维修资源库
故障知识库
企业服务总线（ESB）
消息组件
服务组件
管理组件
安全组件
PDM系统
ERP、MES系统
MRO系统
动车组设计、制造、运维全生命周期
设计创新实践
设计持续改进
设计变更管理
—
工艺创新实践
工艺持续改进
工程变更管理
—
预测性维修技术实践
检修修程修制优化
精细化计划修管理
精细化故障修管理
质量追溯管理
履历变更管理
维修上下游协同
远程故障诊断支持
增 值 业 务

大数据不同观点：个人信息滥用

值得注意的是，随着数据化的深入和应用，对于大数据应用安全和准确性的担忧声音也在日益加剧。

英国剑桥大学心理测量学中心研究员米哈乌·科辛斯基称，根据收集到的用户数据对用户行为和喜好进行推测，既可以为善，也可以作恶，尤其被广告商不当利用之时，就可以把顾客玩弄于股掌之中。

•

数据分析的客观性问题

- 麻省理工学院公民媒体中心访问学者凯特·克劳福认为，由于数据科学家根据统计方式或者预设的程序，能从海量无序的数据中提炼出相关信息，却不能确保这些数据一定是客观的，因为都是人为设定的程序，存在偏差。克劳福以桑迪飓风期间Twitter上的推文为例，称讨论桑迪话题最多的是未受影响的曼哈顿地区，而真正受影响的却很少发出推文。

  克劳福称，如何解决大数据科学中潜在的认识偏差问题，短期内数据科学家应该学习社会科学家的一些方法论，知道数据来自哪里，采用何种方法来收集和分析等。从长期来看，应该知道如何通过小规模的数据的研究来获得大数据的方法论。

数字孪生（digital twin）

数字孪生

是以数字化方式创建物理实体的虚拟模型， 这个虚拟物可以完美地模拟其物理产品的物理属性和动态性能，通过虚实交互反馈、 数据融合分析、决策迭代优化等手段，为物理实体增加或扩展新的能力。

数字孪生体

机械或系统的精准、虚拟副本----数字孪生体。

正给工业带来革命性变化

数字孪生本质与特征

本质

信息世界中对物理系统内在规律和外在属性最直观与最全面的描述，是物理系统全生命周期所有数据和模型集成与共融载体。

数字孪生五维模型

- ✓ **物理实体：** 客观存在，由各功能子系统间的协作完成特定任务。部署在物理实体上的各种传感器实时监测环境数据和运行状态。
- ✓ **虚拟模型：** 是物理实体忠实的数字化镜像，集成与融合了几何、物理、行为及规则 4 层模型。
- ✓ **服务系统：** 集成了监控、评估、控制、优化等各类信息系统，基于物理实体和虚拟模型提供智能运行、精准管控与可靠运维服务，决策及优化。
- ✓ **孪生数据：** 包括物理实体、虚拟模型、服务系统的相关数据，领域知识，并随着实时数据的产生不断更新与优化。是核心驱动。
- ✓ **连接：** 将以上 4 个部分进行两两连接，通过实时交互以保证各部分间的一致性与迭代优化降低装备的整个寿命周期成本

数字双胞胎优化船舶设计，维护和性能

旨在建立一个行业标准的开放源码数字仿真平台

--**令人兴奋的技术，以提高航运的安全性和效率**

2017年7月Rolls-Royce Marine, DNV GL and SINTEF

在挪威的海事中心NMK签署备忘录

2018年3月正式启动

功能：

设计阶段的优化、需求设计和类型批准

建设阶段的集成、测试和验收、接口管理和认证

操作阶段的变更管理、故障诊断预测、培训和分类

该项目已经完成，并在云上运行。

机器人+数字孪生：哈工大联合实验室

数字孪生挑战

数据困难： 数据类型多、数据缺失、多类型融合、数据拥有权

模型挑战： 没有通用的标准、规范和方法，每个系统、部件都要从零建模

协同协作： 多学科紧密合作的专家团队，知识的共享

（目前存在工业界以商业机密原因与学术界联系不紧密）

本章思考题

1、简要描述云计算概念，并给出云架构。

2、简要描述大数据技术的技术特征，并给出大数据技术架构。

3、简要描述数字孪生的概念和技术特征，并给出面向机器人应用的技术架构。