大数据与工业大数据

刘海龙
西北工业大学
2018年图书馆交流

西北工业大学
NORTHWESTERN POLYTECHNICAL UNIVERSITY

提纲

- 大数据（Big Data）
  - *What\Where\Why\How*
- 大数据分析与管理技术
- 智能制造与工业大数据
- 结束语

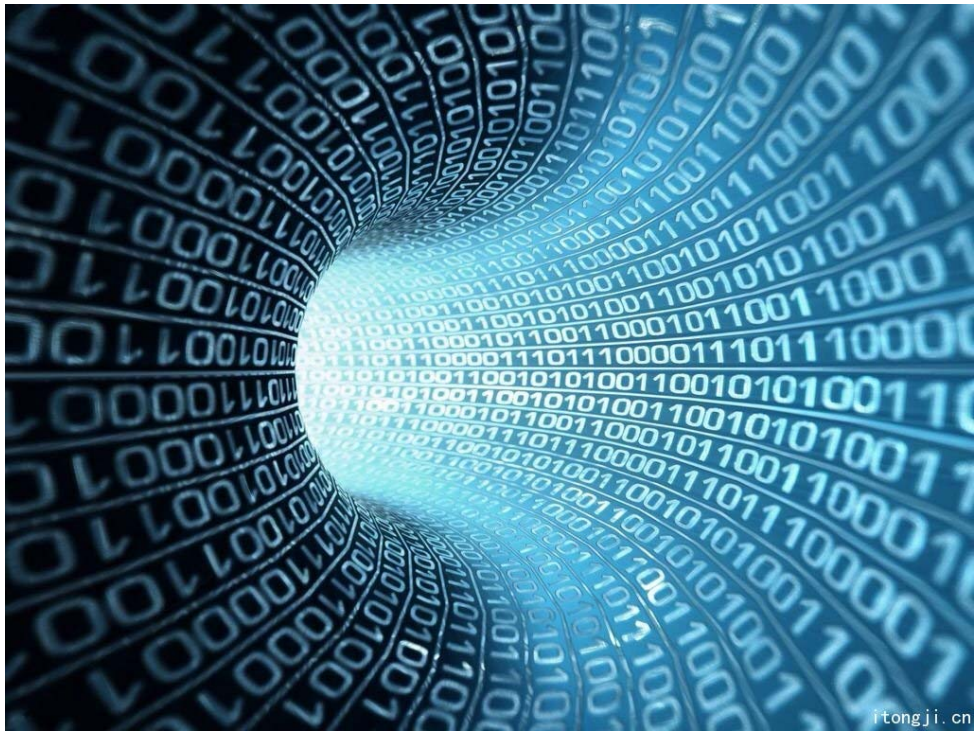
提纲

- **大数据（Big Data）**
  - ***What\Where\Why\How***
- 大数据分析与管理技术
  - 高性能高可用--并行数据库
  - 分布式并行分析引擎--MapReduce
  - 非关系型数据库--NoSQL数据库
  - 常驻内存速度为王--主存数据库
- 结束语

一. 认识大数据时代

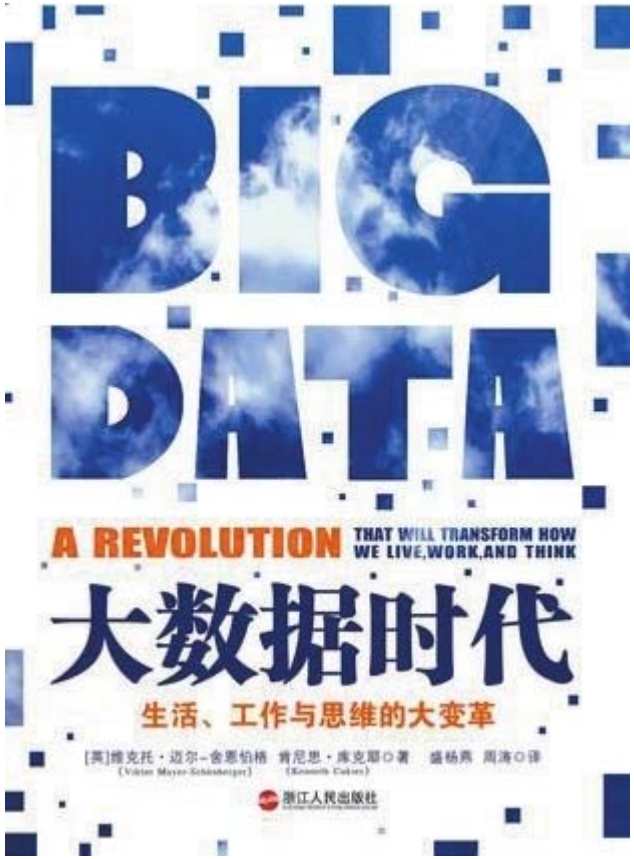
随着一系列标志性事件的发生和建立，人们越发感觉到大数据时代的力量。因此2013年被许多国外媒体和专家称为“大数据元年”。

当今『大社会』，
三分技术，七分数据，
得数据者得天下。



一. 认识大数据时代

拒绝大数据时代，可能会失去生命！



《大数据时代：生活、工作与思维的大变革》一书的作者维克托·迈尔·舍恩伯格，如是说，“如果你是一个个人，如果你拒绝的话，可能会失去生命，如果是一个国家的话，拒绝大数据时代的话，可能失去这个国家的未来，失去一代人的未来。”

这一句话恐怕不能算作耸人听闻，因为每当人们站在现在这个节点的时候，总会去眺望未来，但是未来往往在你不经意当中已经悄悄地来到你的身边。

一. 认识大数据时代

大数据时代到来的必然性：

- 硬件成本的降低
- 网络带宽的提升
- 云计算的兴起
- 网络技术的发展
- 智能终端的普及
- 电子商务、社交网络、电子地图等的全面应用
- 物联网

What--什么是大数据？

大数据的定义—不同的声音

对于运营商来说，这个“大数据”主要就是指大量的用户产生的行为数据。 *-- 某通讯运营商*

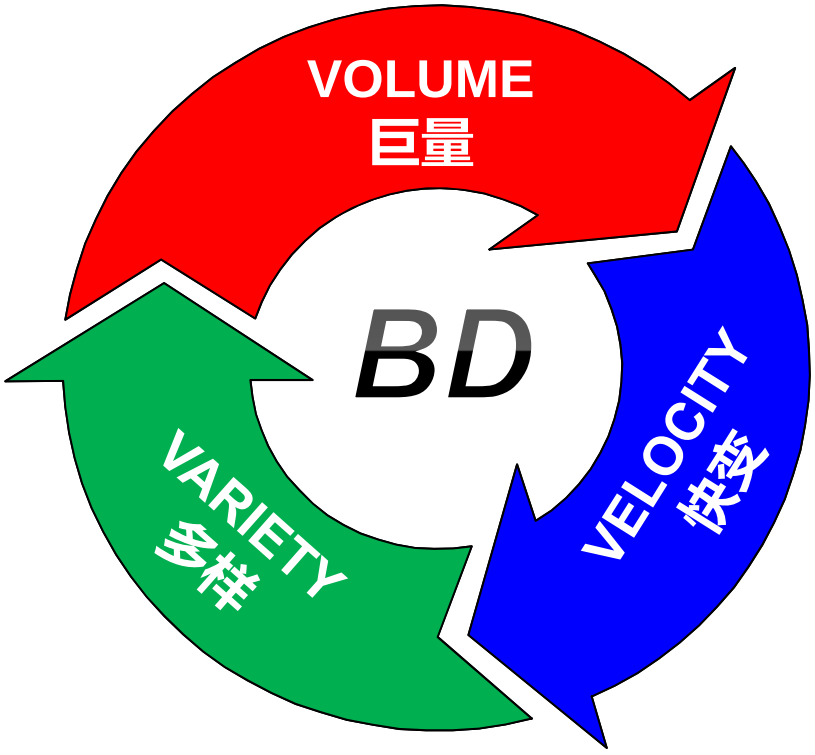
传统的基于集中式或者小规模分布式和并行系统无法满足大数据的计算需求，弹性的计算能力是大数据定义的重要维度。 *-- 某云服务提供商*

“Big data” refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze. *-- “Big data: The next frontier for innovation, competition*

“大数据是指无法在一定时间内用常规软件工具对其内容进行抓取、管理和处理的数据集。 *--维基百科*

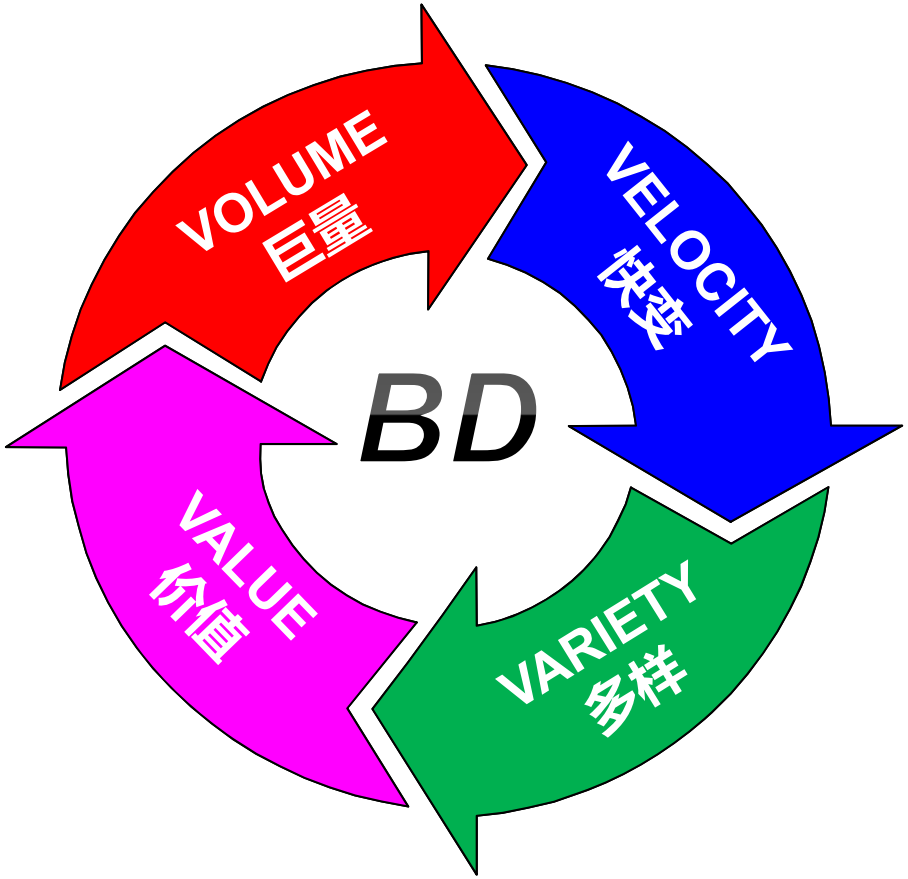
用 3 *V* 定义大数据

- “大数据”不仅仅是“大量数据”



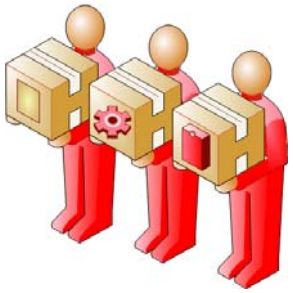
用 更多的*V* 定义大数据

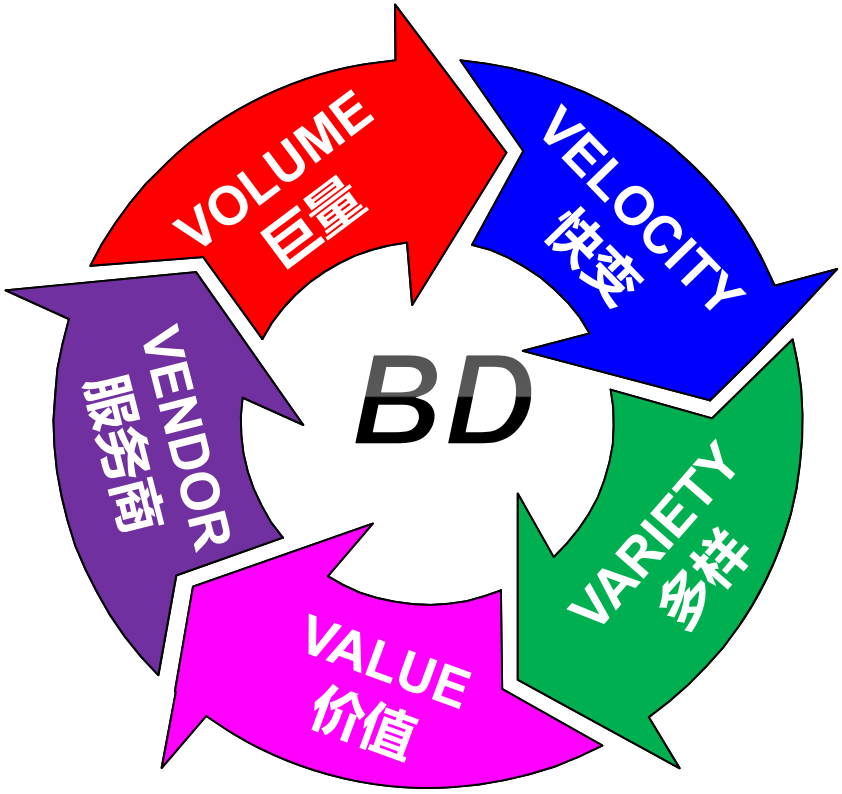
- 价值(Value)、服务商(Vendor)、向量(Vector). . .



用 更多的V 定义大数据

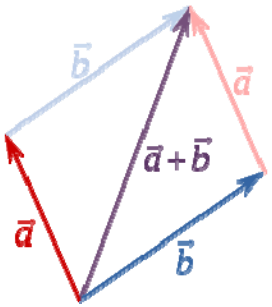
- 价值(Value)、服务商(Vendor)、向量(Vector). . .

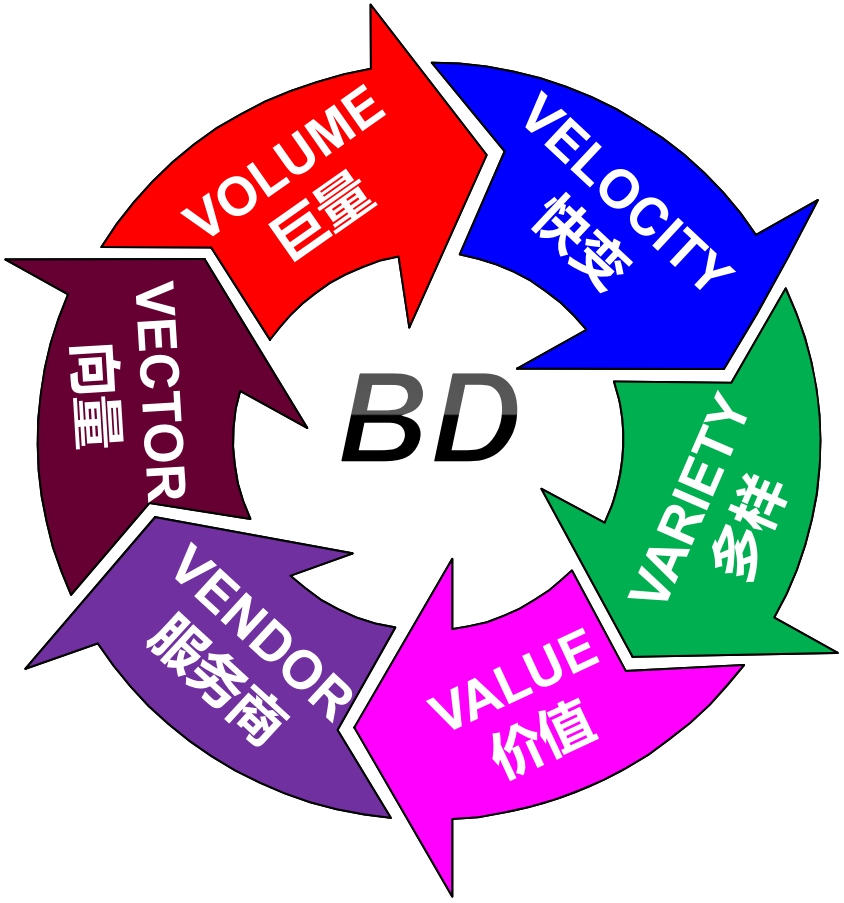




用 更多的V 定义大数据

- 价值(Value)、服务商(Vendor)、向量(Vector). . .





Where--大数据的来源 — 众流汇海

Number of Emails sent every second	2.9 Million
Data consumed by households each day	375 MB
Video upload to YouTube every minute	20 Hours
Data per data processed by Google	24 PB

Total minutes spent on Facebook each month	700 Billion
Tweets per day	50 Million
Data sent and received by mobile internet users	1.3 EB
Products ordered on Amazon per second	72.9 items

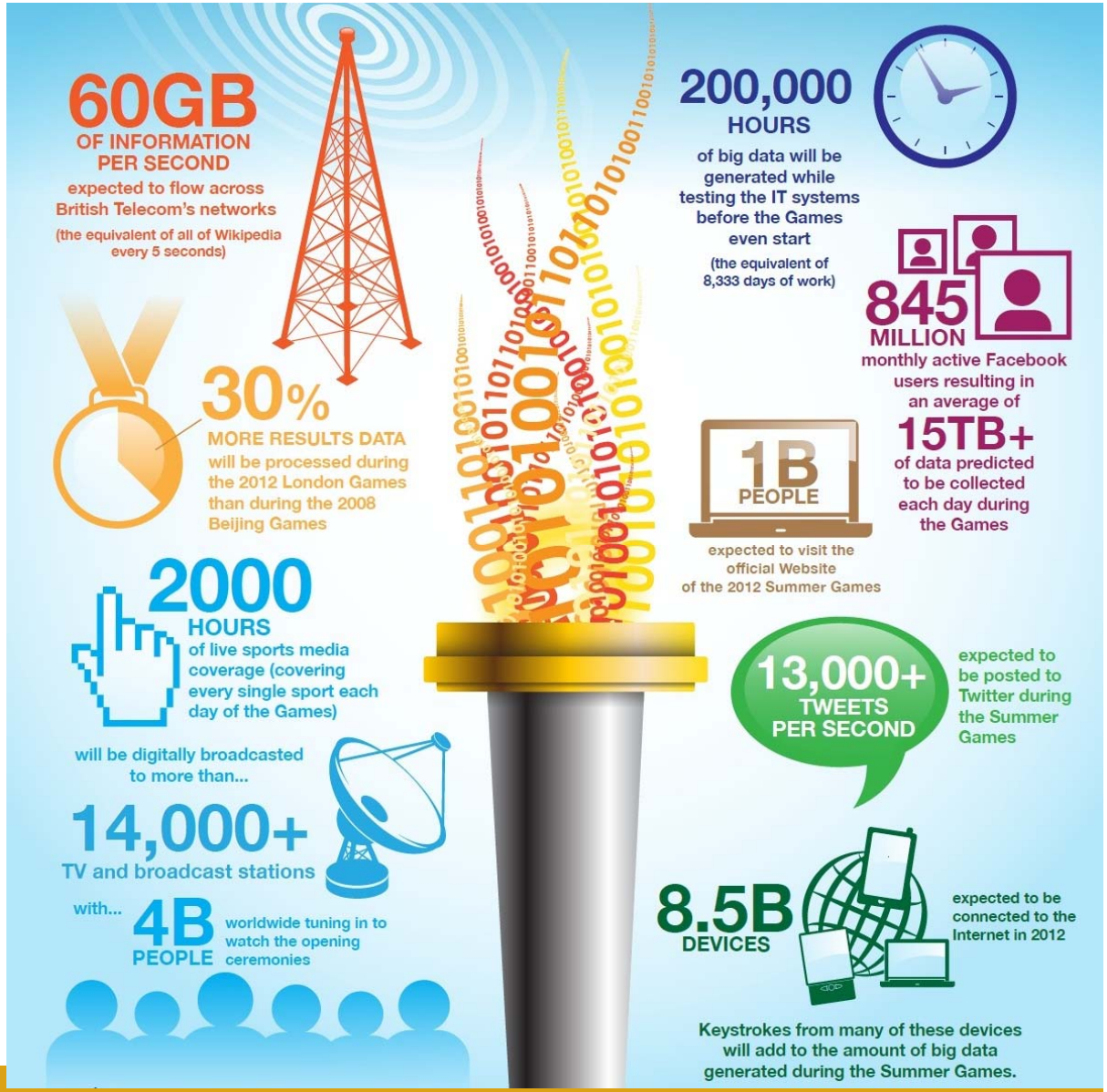
2012伦敦奥运产生的数据 一瞥

2012 London Summer Games

BIG DATA BY THE NUMBERS

The Age of Big Data Is Dawning

For two weeks this summer, when the world comes together for the 2012 Summer Games, an unprecedented spike in the sheer volume of big data is expected to be generated on a global scale.



西北工业大学
NORTHWESTERN POLYTECHNICAL UNIVERSITY

Why--大数据衍生大变革

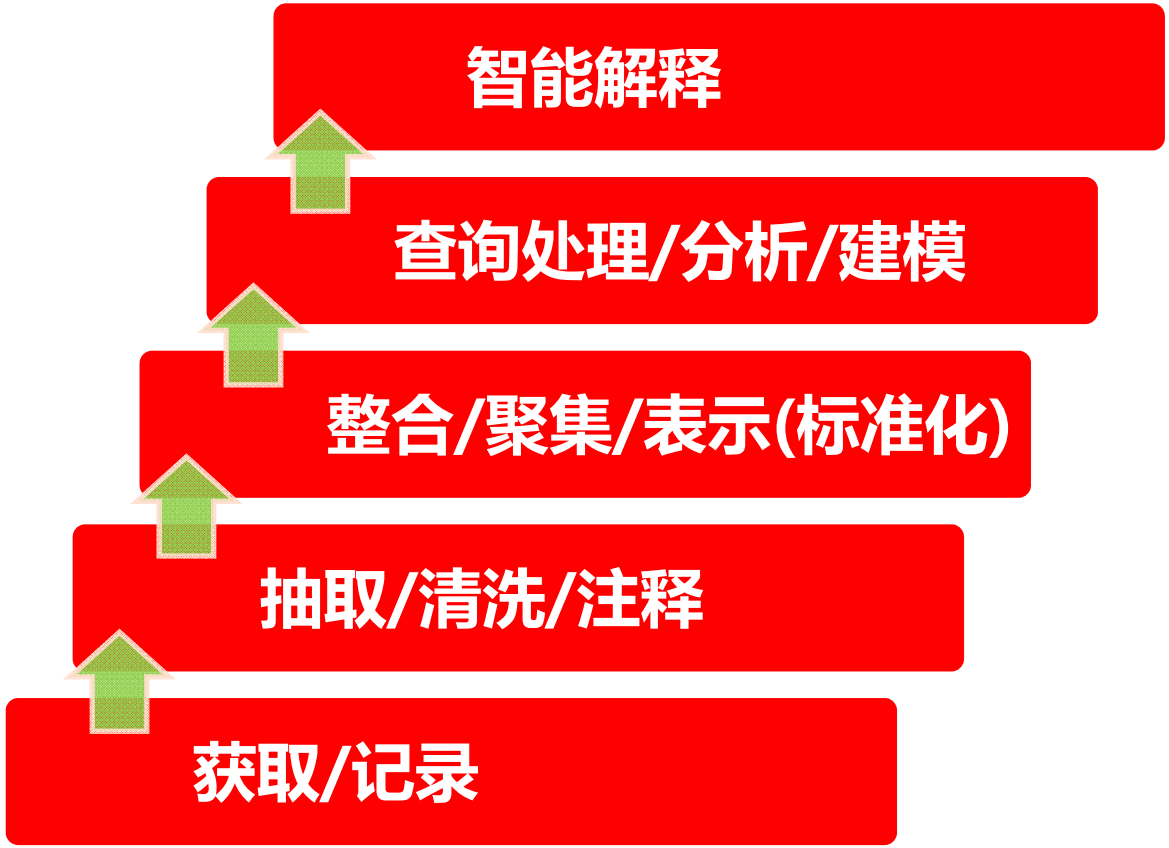
- ***智慧城市***：大数据时代的到来给智慧城市的建设带来了新兴的增长机会。
- ***产业创新***：从IT业，制造业，金融业，甚至是体育产业和旅游产业中都能发现大数据所带来的变革。
- ***科研创新***：在天文观测、气象监测、生物基因、物理仿真等数据密集型科学研究中都将遭遇大数据的挑战。
- 2012年奥巴马政府公布了2亿美元的“大数据研发计划”。

Why--大数据产生大价值

美国 卫生保健	制造业	全球个人 位置数据	欧洲公共 部门管理	美国零售业
年增加产业价值	*减少开发和 组装成本*	*增加服务 提供者收益*	*增加产业价值*	*增加净利润率*
$300 B	**–50%**	**$100 B**	**€250 B**	**60+%**

*Source: * McKinsey Global Institute: Big Data – The next frontier for innovation, competition and productivity (May 2011)*

How--大数据处理的主要阶段



From: H.V. Jagadish et al. "Challenges and Opportunities with Big Data"

大数据处理的主要挑战 – 3V

- 异构
  - 异构平台支持（硬件环境、软件系统）
  - 异构数据支持（复杂内联）
- 数据规模
  - 分布式存储架构与并行数据处理技术（高可扩展）
  - 压缩技术
  - 对算法设计的挑战（可扩展的高效并行算法）
- 时效性
  - 处理“流”或接近实时的分析与处理的框架与平台
  - （多维）索引技术

大数据处理的主要挑战 – *数据质量*

- 数据的”多面性“
  - 大数据涉及到处理过程复杂，每一个阶段都会引入数据质量问题。
    - 尤其是在数据垮组织流动时
  - 繁多的数据类型及应用
- 高度的复杂性以及上下文无关性
  - 处理对象及处理过程都非常复杂
  - 数据质量问题呈现出不同的模式
- 缺乏”利器“
  - 需要利用一系列的工具来控制数据质量
  - 需要设置一系列的规则来利用数据

Data Quality and Data Cleaning: An Overview. Theodore Johnson. johnsont@research.att.com. AT&T Labs – Research.

大数据处理的主要挑战 – *数据隐私*

- 隐私数据加密技术
  - 内容保护（如位置信息、疾病史）
  - 身份保护
  - 关系保护
  - 基于可检索、可计算的加密技术
- 基于加密隐私数据的挖掘技术
- 敏感查询的保护
  - 高频查询
  - 查询结果
- 非技术层面（管理、政策、法律...）

大数据处理的主要挑战 – *其他*

- 可视化
- 交互式协作
  - 众包（crowdsourcing）
    - 参与感知（Participatory Sensing）
- 低能耗（绿色计算）
  - 高效数据中心
- …

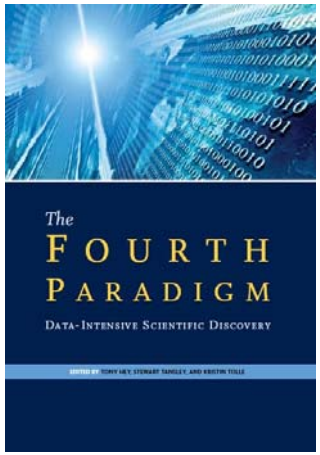
提纲

- 大数据（Big Data）
  - *What\Where\Why\How*
- **大数据分析与管理技术**
  - 高性能高可用--并行数据库
  - 分布式并行分析引擎--MapReduce
  - 非关系型数据库--NoSQL数据库
  - 常驻内存速度为王--主存数据库

从“大数据”到“智能数据”

- **数据驱动的决策（Data-driven decision-making）**：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，用数据创造价值。

- **科学发现的第四范式**
  - **实证科学(Empirical Science)**
    - *描述自然现象*
  - **理论建模(Theoretical Modeling)**
    - *利用模型和泛化*
  - **计算模拟(Computational Simulations)**
    - *模模拟复杂的现象*
  - **数据密集计算(Data-intensive Computing)**
    - *综合利用理论、实验和模拟应对规模数据*



大数据管理的基本原则

- 分区(Partition)与键值(key-value)存储
  - *违背第一范式*
- 容忍不一致性
  - *违背ACID 原则*
- 用副本(Replica) 支持容错处理
  - *不单是热冷备或服务器镜像*
- 高扩展和高性能
  - *用scale-out替代scale-up*

舞动大数据 — 大数据分析技术

- 大数据分析技术是多种技术的组合：
  - *分布式并行计算技术*
  - *NoSQL数据库技术*
  - *并行数据库技术*
  - *主存数据库技术*
  - 。。。

你需要哪种技术？ – *了解需求*

- 读密集（Read-intensive） vs. 写密集 （Write-intensive）
- 易变数据 vs. 静态数据
- 即时一致性 vs. 最终一致性
- 低访问延迟 vs. 高访问延迟
- 可预测访问模式 vs. 不可预测的访问模式
- 负载(workload)类型:*OLTP\OLAP*

创建更多的自己的checklist！

One Size Does Not Fit All

“ Attempting to force one technology or tool to satisfy a particular need for which another tool is more effective and efficient is like attempting to drive a screw into a wall with a hammer when a screwdriver is at hand: the screw may eventually enter the wall but at what cost? ”

- 多样的大数据
- 多变的应用需求
- 多维的技术方案

需求为导，顺势而动！

Source: E.F. Codd et al. “Providing OLAP to User-Analysts: An IT Mandate”

提纲

- 大数据（Big Data）
  - *What\Where\Why\How*
- 大数据分析与管理技术
- **智能制造与工业大数据**
- 结束语

1、工业大数据

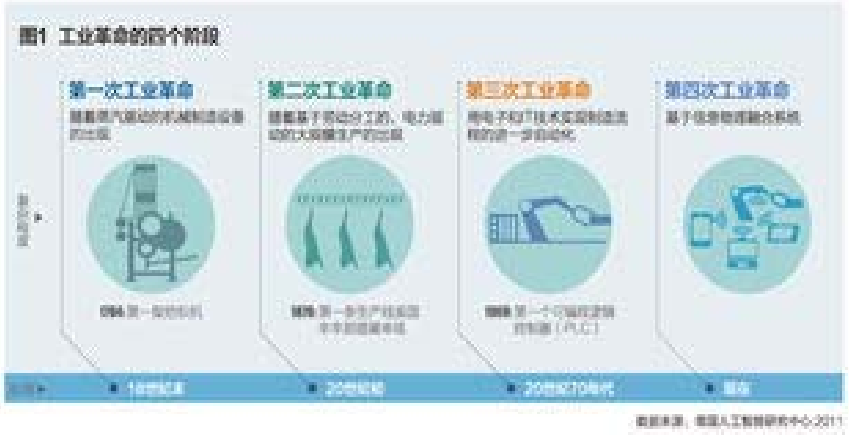
- 简单来讲，工业大数据就是在工业领域信息化相关应用中所产生的海量数据，注意这里的“相关应用”意味着不仅包括企业内和产业链，还包括客户用户和互联网上的数据。

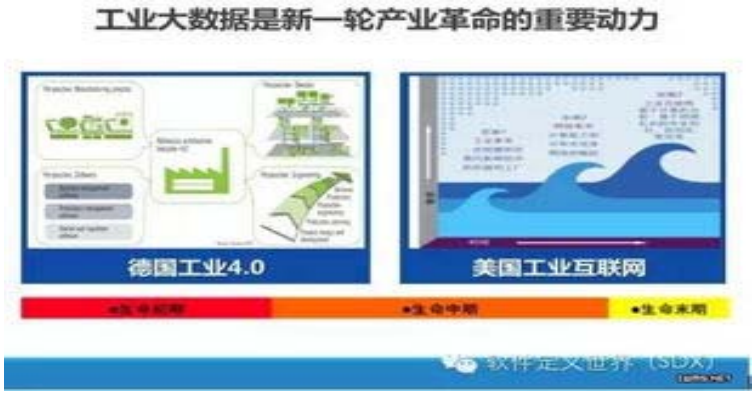
清华大学 王建民教授

- 2012年，GE公司率先明确了“工业大数据”的概念。同年麦肯锡的报告中给出了一个有趣的事实：那就是在虚拟经济占主导地位的美国，其工业界蕴含的数据总量反而是最大的。

工业大数据发展背景（1）

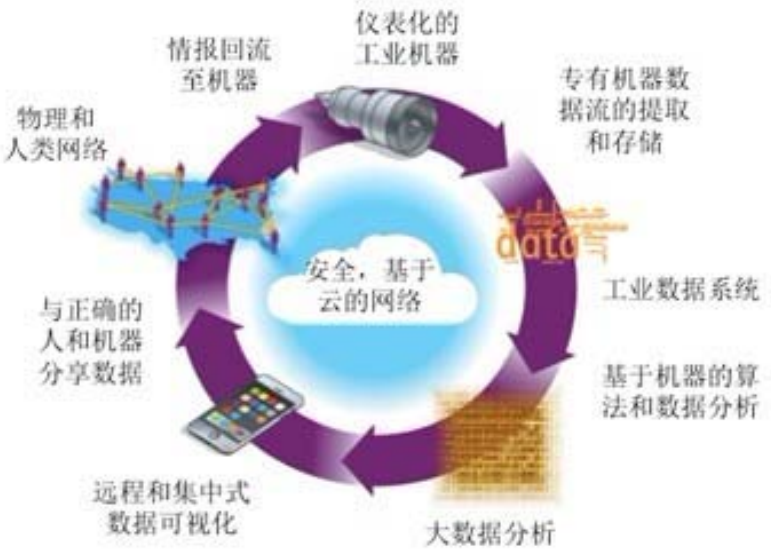
- 工业发展进入新阶段
  - 经历自动化、进入网络化、智能化发展新阶段;
  - 美国提出智能制造；德国提出工业4.0;
  - 我国提出“互联网+”和“中国制造2025计划”





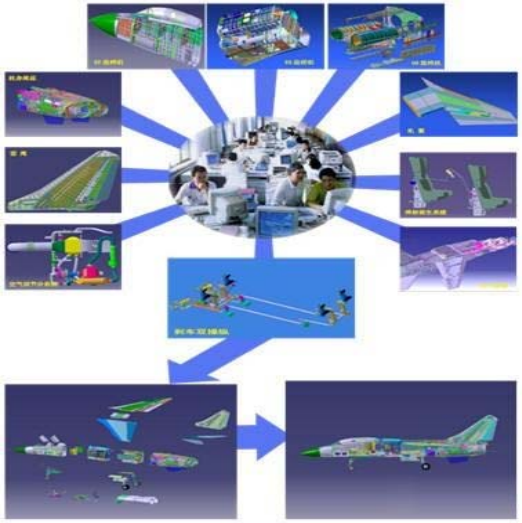
工业大数据发展背景（2）

- “制造业数字化网络化智能化”是新工业革命的核心技术
  - 制造业创新的三个层次：产品创新、制造技术创新、产业模式创新；
  - 数字化网络化智能化是制造业创新的重要途径与共性使能技术；
  - 以数字化网络化智能化为主线，必然形成工业大数据，需要大数据技术、平台与应用支撑；



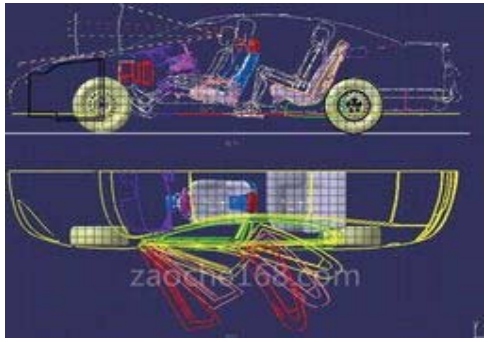
工业大数据发展背景（3）

- 机器数据正在高速增长
  - Wikibon发布《工业互联网与大数据分析：机遇与挑战》;未来10年，工业数据增速将是其它大数据领域的两倍。
  - 大数据咨询公司Think Big在“正在塑造大数据和企业未来的十大趋势”中，将“机器数据和物联网将占据中心舞台”列为首位，并指出“从RFID标签和工业仪器，到喷气发动机和消费电器，整个世界正在生产着越来越庞大的数据量;

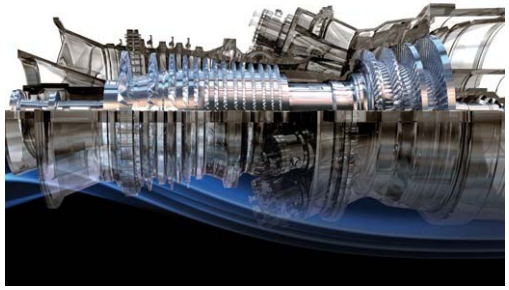


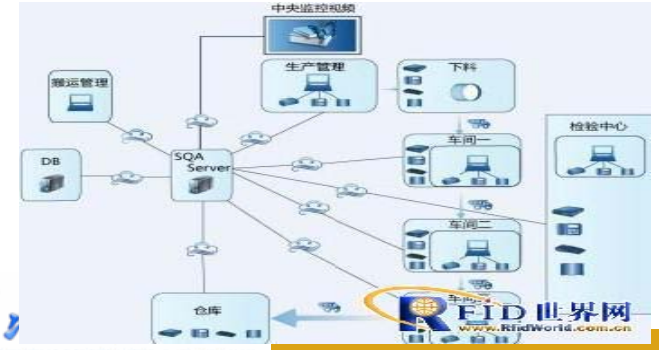
工业大数据特点分析（1）

- **工业大数据来源**
  - **数字化设计**　如 飞机全数字化设计；波音公司利用CATIA软件产生波音777的300万个零部件的尺寸和形状数据；我国飞豹飞机首次实现全数字化三维设计；
  - **智能化制造**　以智能工业机器人为典型代表的智能制造装备已经开始在多个领域得到应用；我国今年的工业机器人超过日本；



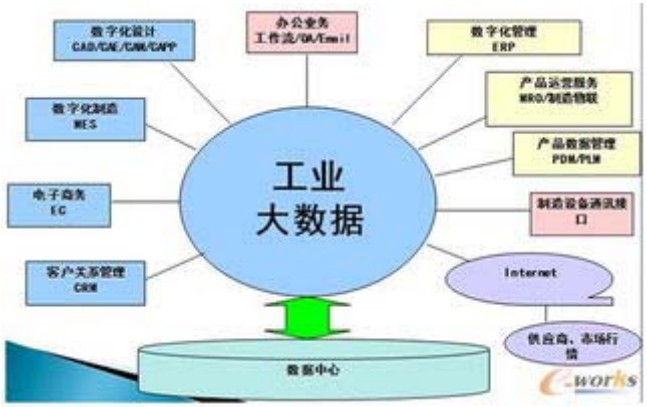
- **网络化监测** 大型工业装备运行状态网络化远程动态监测；例如，波音737发动机在飞行中每30分钟产生10 TB数据；陕鼓实现数百台旋转机械远程在线监测及故障诊断，
- **物联化管理** 工业生产过程开始大量使用RFID实现零件与产品管理；

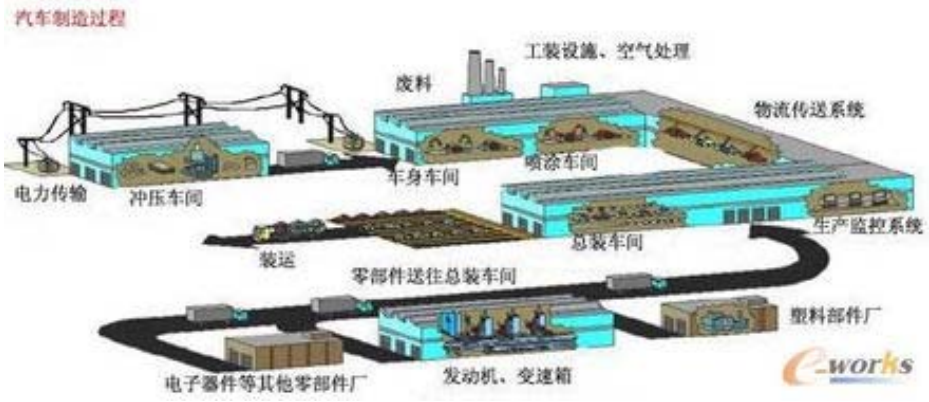




工业大数据特点分析（2）

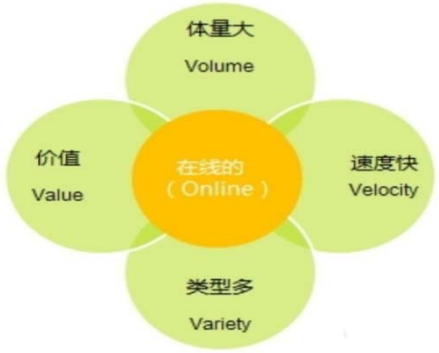
- 工业大数据分类
  - **产品数据**　设计、建模、工艺、加工、测试、维护、产品结构、零部件配置、变更记录等数据。
  - **生产数据**　组织结构、业务管理、装备状态、质量控制、生产过程、采购库存、目标计划等数据。
  - **价值链数据** 市场营销、电子商务、客户、供应商、合作着等数据。
  - 外部数据　行业、市场、竞争对手等数据。

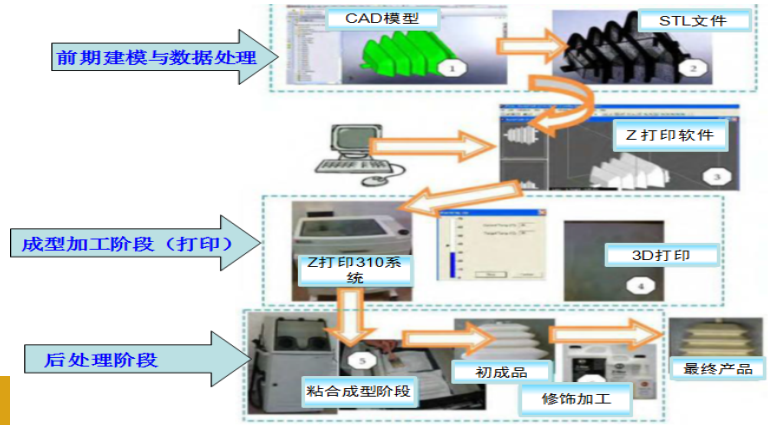




工业大数据特点分析（3）

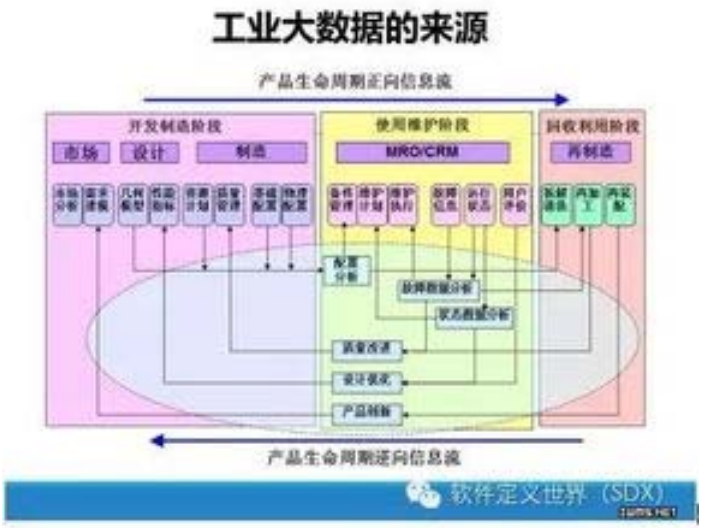
- 工业大数据同样具有“4V”特点
  - **大体量**（Volume）；一类家电智能工业生产过程一天产生20TB以上数据；3D打印一个中规模部件产生几十GB数据；
  - **多样性**（Variety）；三维图形数据，监控视频数据，文字档案数据；RFID数值数据；。。。。。
  - **快速性**（Velocity）；发动机运行监测数据；工业机器人运动数据
  - **价值性**（Valve）；工业大数据的价值是显性的；直接的；

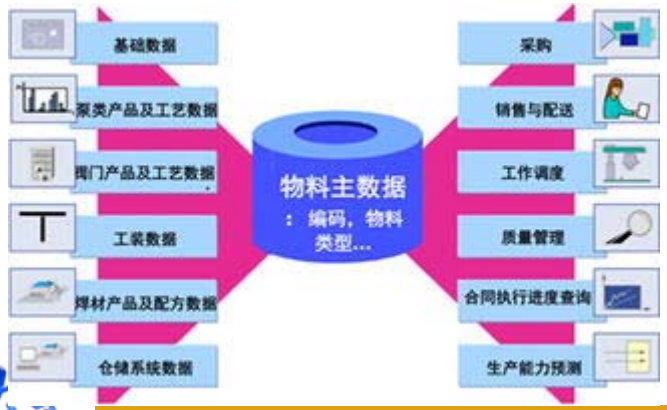




工业大数据特点分析（4）

- 工业大数据具有自身特点
  - 多源性获取，数据分散，**非结构化数据比例大**；
  - 数据蕴含信息复杂，**关联性强**；
  - 持续采集，具有鲜明的**动态时空特性**；
  - 采集、存贮、**处理实时性要求高**；
  - **与具体工业领域密切相关**；





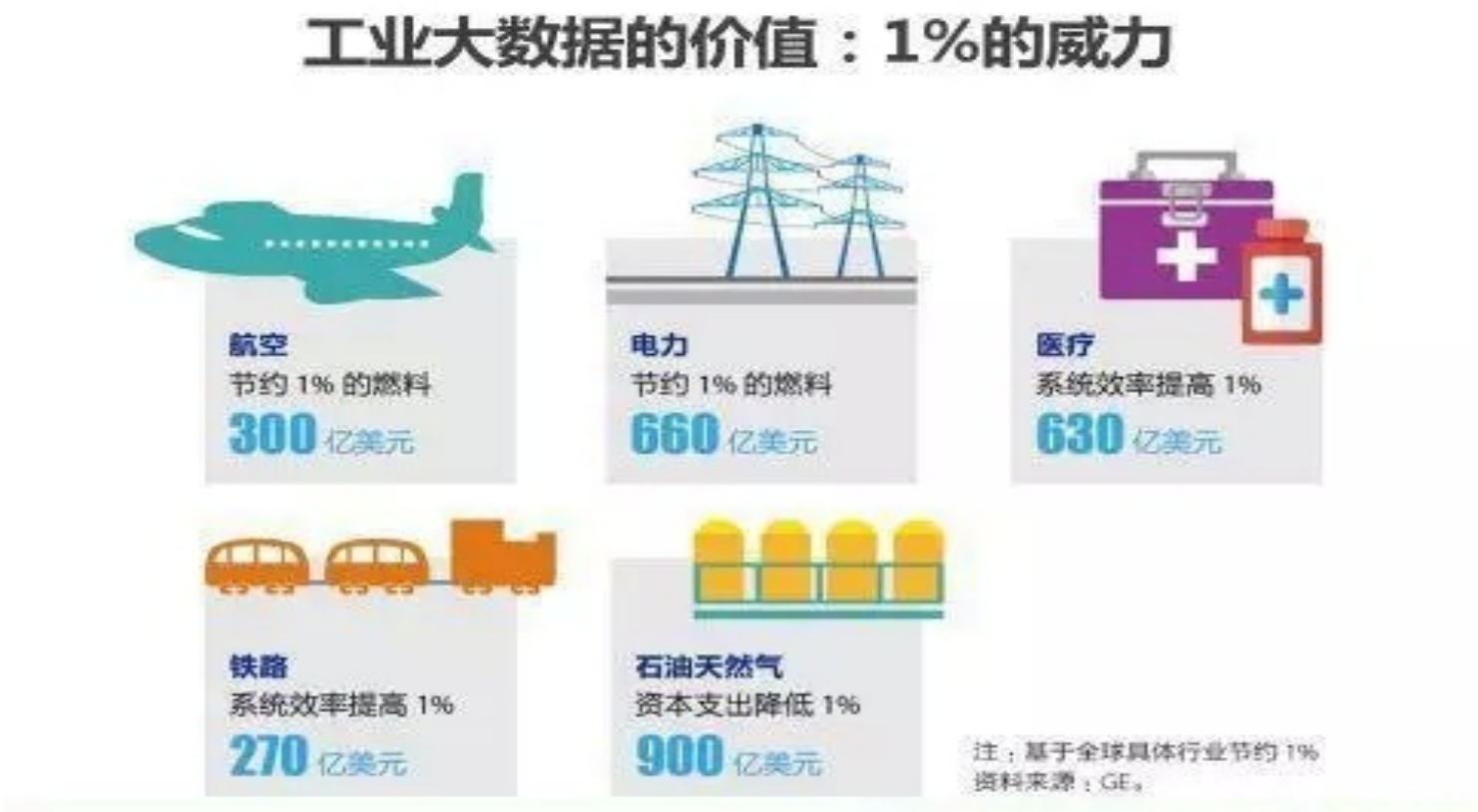
工业大数据特点分析（4）

- 数据价值密度
  - 20%的SQL小数据具有80%的价值密度
    - 例如：产品图纸、试验分析、加工工艺等
  - 80%的工业大数据密度只有20%，需要分析挖掘
    - 例如：工况数据、图片数据、文本数据等
- 数据的依赖关系
  - 20%的SQL小数据将引爆80%工业大数据价值
    - 不举小数据之“纲”，难张大数据之“目”

工业大数据领域2/8法则

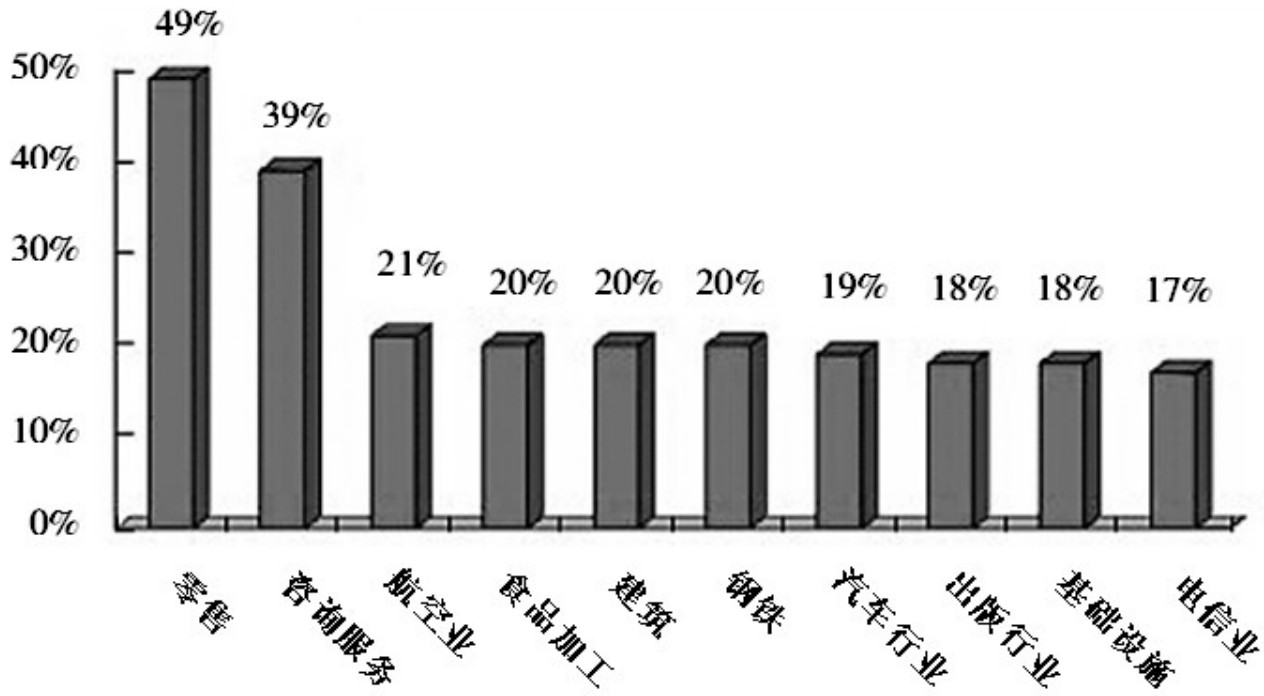
工业大数据潜在价值

同时GE公司的报告还揭示了工业大数据所蕴含



工业大数据潜在价值

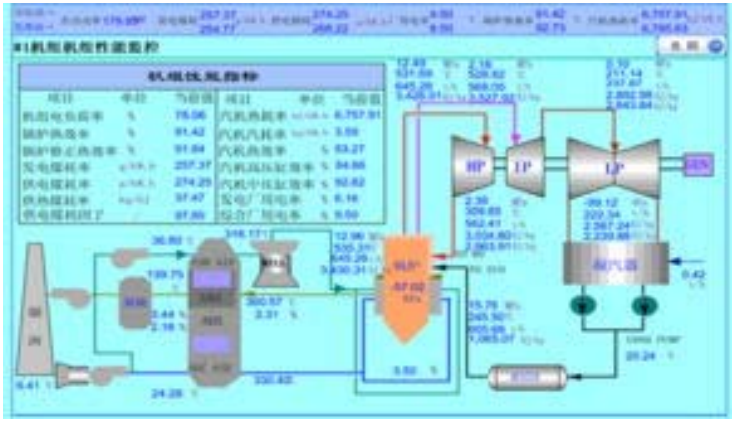
数据使用率提升 10%对行业人均产出的平均提升幅度



数据来源:《Measuring the Business Impacts of Effective Data》

工业大数据潜在价值

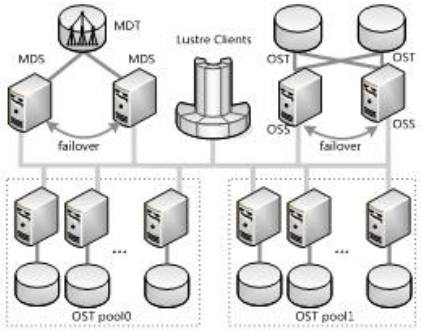
- 工业大数据颠覆传统制造过程的十条途径
  - 监控生产过程
  - 加快业务整合
  - 提高企业制造绩效
  - 改进生产流程
  - 预测供应商绩效
  - 监测生产设备状况
  - 合理计划生产
  - 细化质量管理环节
  - 追踪产能与财务状况
  - 监测产品运维状态



《如何利用大数据改进制造业》麦肯锡咨询公司

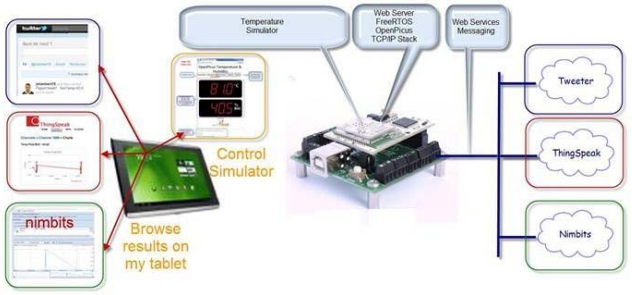
工业大数据处理关键技术（1）

- 工业大数据获取与预处理
  - 数据来源众多；
  - 数据采集设备种类多、接口复杂；
  - 必须支持数量众多的硬件连接驱动；
  - 支持万级以上大规模数据点快速采集；
  - 超效的数据压缩；
  - 分布式实时服务器数据存贮；
  - 支持数万事件精确时间标签分辨率；
  - …



工业大数据处理关键技术（2）

- 工业大数据管理
  - 工业大数据具有鲜明的多样化特点
    - 历史数据；
    - 当前生产流程的实时数据
    - 设计数据；
    - 工业设备运行状态的监测数据
  - 工业大数据存贮成本
    - 合理的数据存贮模式构建
    - 高比率压缩和高效存储技术的实现



数据特性

- 多模态的非结构化工程数据
- 强关联的图数据
- 高通量的时序数据

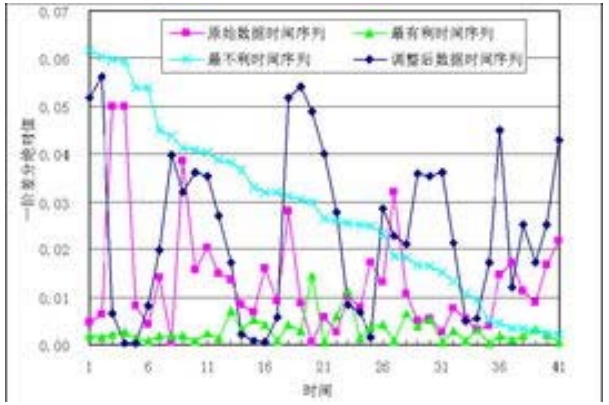
应用特性

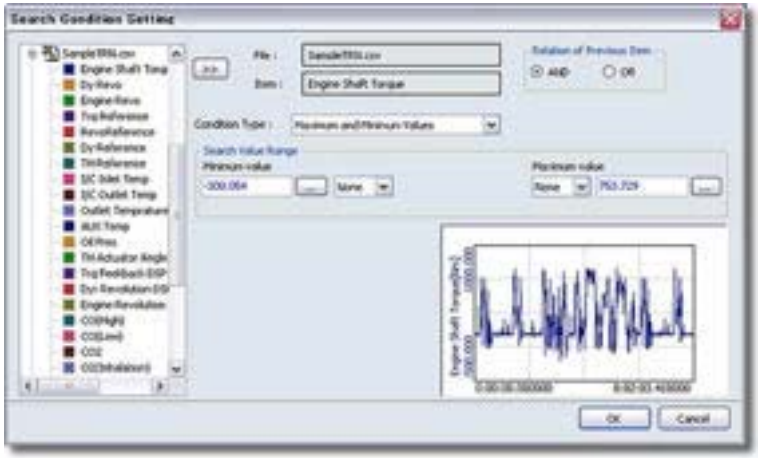
- 跨界关联
- 产业链关联
- 信息物理关联

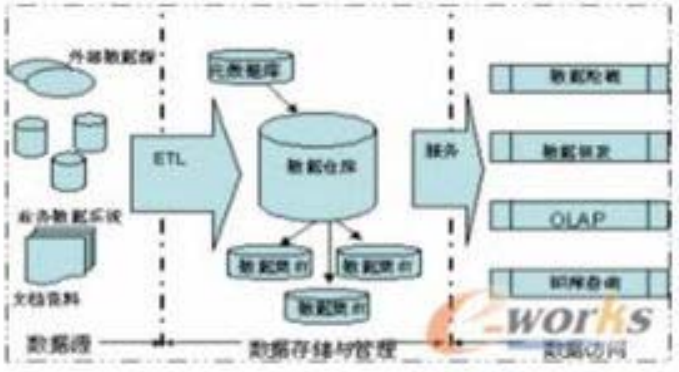
工业大数据处理关键技术（2）

- 大规模时间序列数据管理
  - 过程实时数据的高性能“专用”引擎设计；
  - 历史数据库与过程数据源的无缝连接集成；
  - 适应不同类型数据的统一访问化接口；
  - 基于时间/事件的高效数据检索；

清华大学自主开发了国产时序数据库 IoTDB，已开源，http://tsfile.org/index



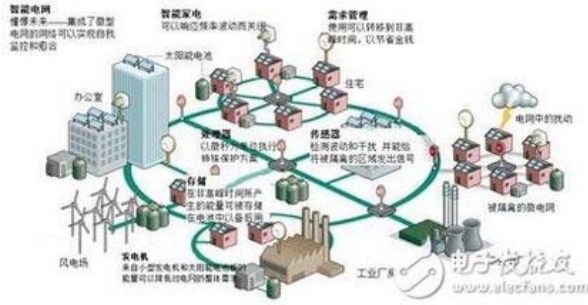




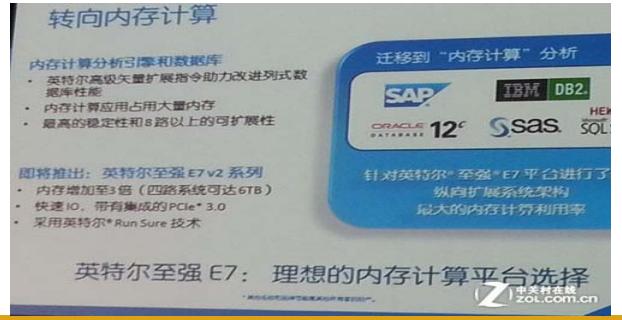
工业大数据处理关键技术（3）

工业大数据应用集成

- 工业大数据采集网络集成
  - 多种传感器；如西门子H-Class燃气轮机配备1500个传感器，用于以秒为间隔精确测定诸如温度、压力、气体成分和发电功率等等关键工作参数值。
  - 多种通讯协议；工业以太网；工业无线网；。。。。
  - 多种数据格式；
- 多计算平台整合
  - 分布式计算平台
  - 内存计算平台
  - 流计算平台



▲图2 基于云计算的数据挖掘平台架

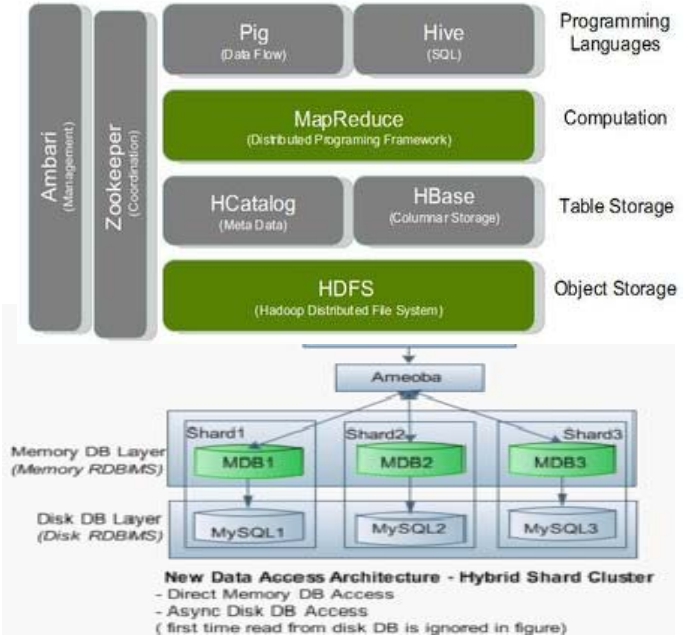


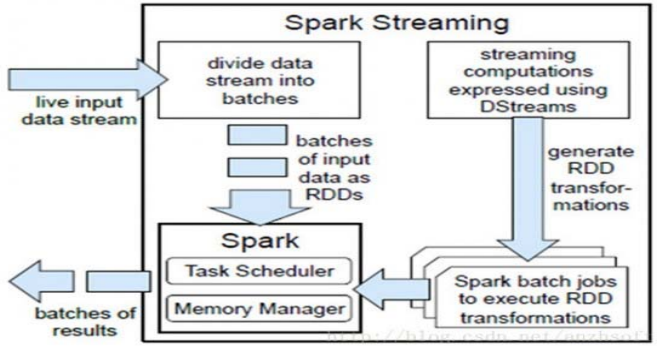
工业大数据处理关键技术（4）

- 工业大数据分析
  - 工业大数据分析是工业大数据计算的重点，是能否体现工业大数据价值的关键所在
    - 适应各类工业大数据分析的通用方法研究与开发
    - 面向具体工业领域数据分析的专用方法研究与开发
  - 关联分析是工业大数据分析需要解决的首要问题
  - 快速分析成为工业大数据分析应着重实现的分析方法
  - 面向具体优化目标的工业大数据应用分析
  - 工业大数据分析与应用需要考虑不同的企业信息化发展阶段的适应性。

工业大数据处理关键技术（5）

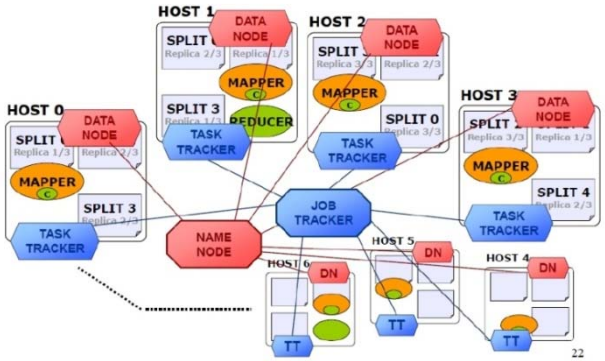
- 工业大数据分析模型
  - 分布式并行计算模型是基础
    - Hadoop/MapReduce
      - 主要面向静态数据的批处理
      - 实时性不够；
  - 流式计算模型是核心
    - S4、Storm、Spark、…
      - 利用内存，减少I/O
      - 提供面向动态数据处理的原语
      - 表达能力不足，无法较好支持复杂算法
      - 容错机制、负载均衡等还需完善
  - 增量计算模型为提升
    - Nectar、Nova、Percolator





工业大数据处理关键技术（5）

- 提升大数据计算方法
  - 数据流实时分析;
  - 可扩展统计分析;
  - 异质数据混合计算;
  - 基于机器学习的智能分析;
  - 基于领域知识的分析;



- 创新大数据计算应用模式
  - 大数据计算应用的关键是服务企业核心业务；如红领，海尔
  - 注重与企业主流业务的紧密对接与融合;
  - 成功的大数据分析重在为企业揭示风险并识别新发展机会;

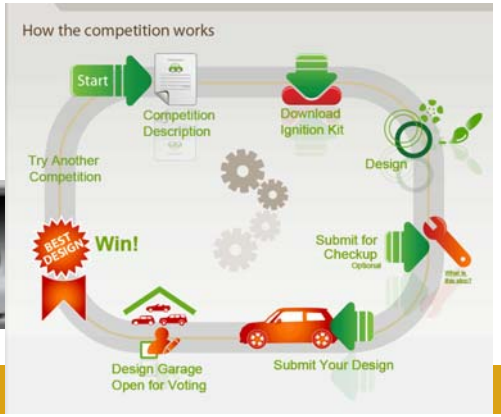
波音787的协同制造服务平台

- 采用了基于网络协同、制造服务外包的模式，组织全球40多个国家和地区协同研发
- 使之与波音777相比，研发周期缩短了30%，成本减少了50%，并在上市前即获得了1400多亿美元的订单，取得显著的效益。

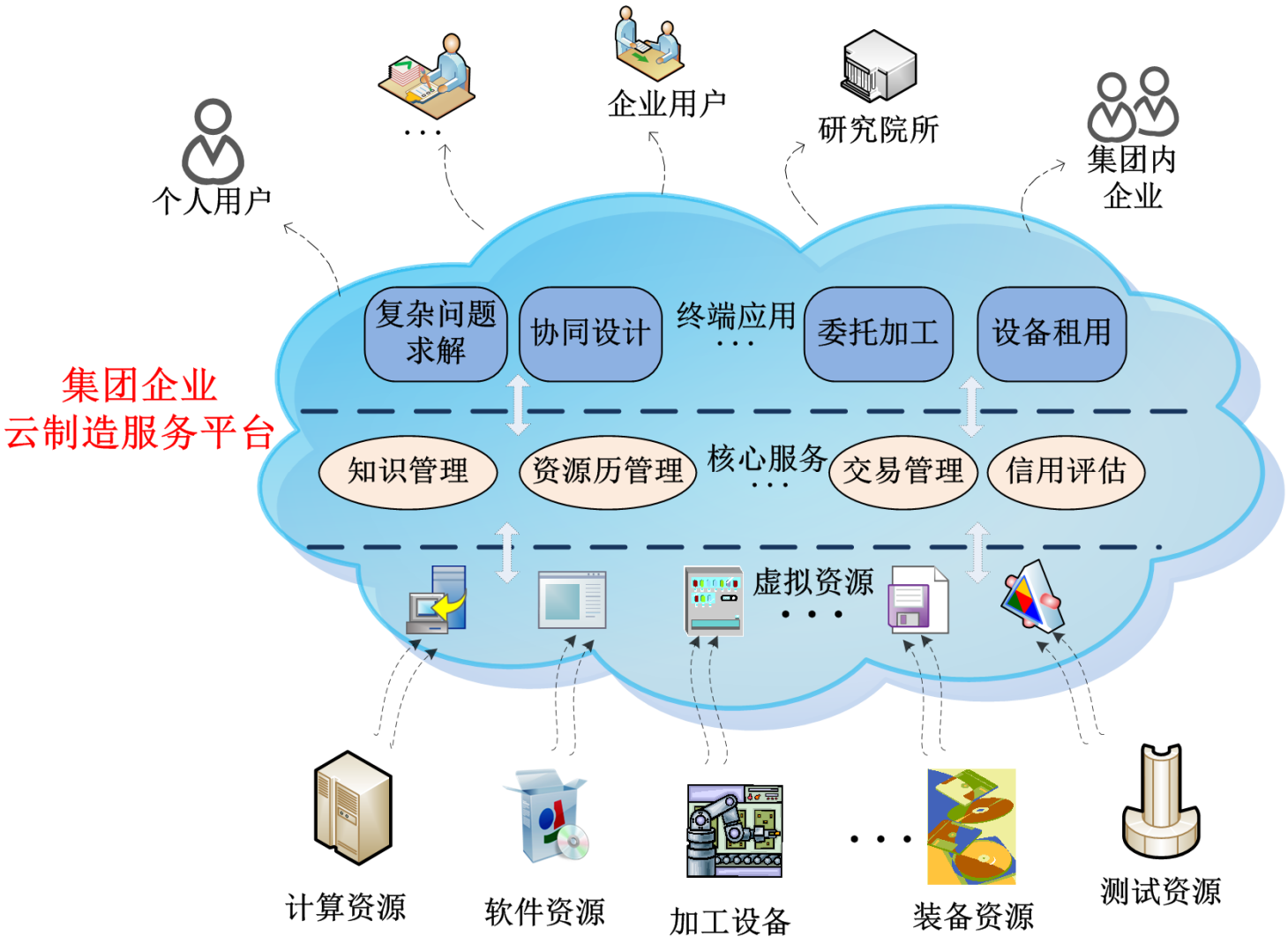
众包制造的典型应用 Local-Motors

- 将越野赛车的全部个性化设计与制造过程众包给社区
- 从图纸到面市只需18个月，仅相当于底特律对车门做规格调整的时间
- 美国一家只有干洗店大小的工厂，拉动整个供应链

LOCAL MOTORS



集团企业的云制造服务平台

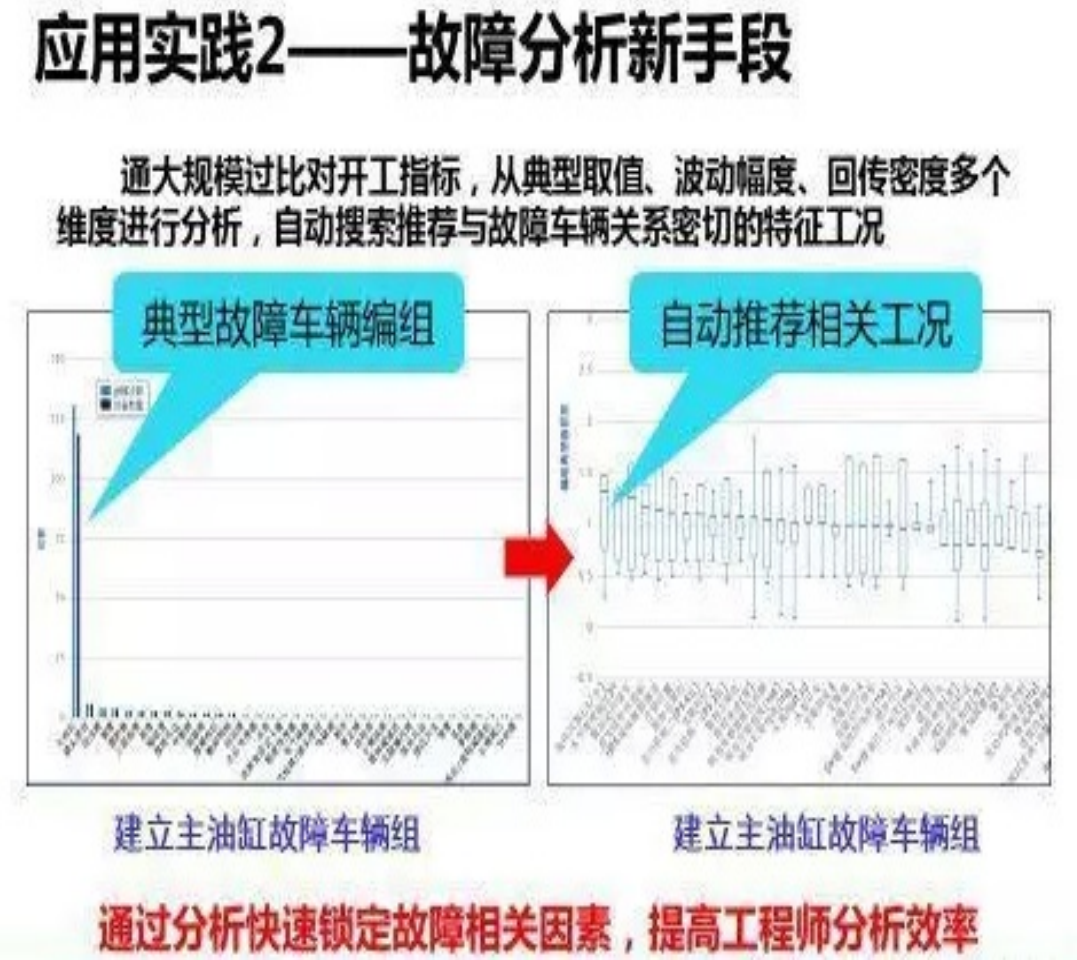


应用实践

- 再举个例子，我们知道液压系统是工程机械的核心系统之一，导致故障的原因有很多，例如：密封套腐蚀，内壁刮花，密封环损坏，阀块受损，等等。有了工况大数据就可以寻找深层次原因。

应用实践

- 有了工况大数据，我们通大规模过比对开工指标，从典型取值、波动幅度、回传密度多个维度进行分析，自动搜索推荐与故障车辆关系密切的特征工况，发现这些故障车辆的每分钟换向次数在变化幅度上高度相关。



应用实践

- 再通过引入互联网上的行政区划数据和历年高铁建设数据（企业外部数据），可以得出这样一个结论，这些典型故障均发生在2012年~2013年期间在建重大工程“杭深高铁”沿线，这为我们寻找更深层次的原因提供了重要线索。

应用实践

- 最后，我们可以通过大规模工况数据透视宏观装备应用情况，可以根据这些信息，进行易损配件需求的预测，优化调配我们的服务资源，甚至我们可以推测各地宏观经济情况。

应用实践3——宏观决策分析

实时展现各类主机的区域分布及设备保有量情况及时掌握在外设备开工情况，助力高层决策。

提纲

- 大数据（Big Data）
  - *What\Where\Why\How*
- 大数据分析与管理技术
- 智能制造与工业大数据
- **结束语**

- 国家重点研发计划“面向高端制造领域的大数据管理系统”
- 清华大学、北京大学、中国人民大学、哈尔滨工业大学、***西北工业大学***、复旦大学、武汉大学、中国运载火箭技术研究院、中国商飞、金风科技、国家气象中心

大数据存储与管理
工业和信息化部重点实验室
MIIT Key Laboratory of Big Data Storage and Management

在研项目中：主持或参与**国家重点研发计划**课题4项，主持**国家自然科学基金重点项目**2项，参与国家自然科学基金重点项目2项，主持国家自然科学基金近10余项。科研经费充裕！

西北工业大学
NORTHWESTERN POLYTECHNICAL UNIVERSITY

欢迎大家报考实验室的研究生！或者合作研究！

参考文献及致谢

本课件核心内容来源于下列文献：

- 李战怀，西北工业大学. 大数据时代
- 周兴社, 西北工业大学. 工业大数据特点、价值及其计算.
- 王建民 ，清华大学. 中国工业大数据的实践与思考.
- ……

（若有疏漏，请谅解！）

特此致谢！

Thank You

liuhailong@nwpu.edu.cn