学习方法

- 1、记笔记

题干、重点知识点、一些难记忆的概念

- 2、整理：

加注释、填知识点、收集报错信息

- 3、写总结

学会的知识点、常忘的知识点、归纳成册

大数据行业与技术趋势

www.huawei.com

HUAWEI

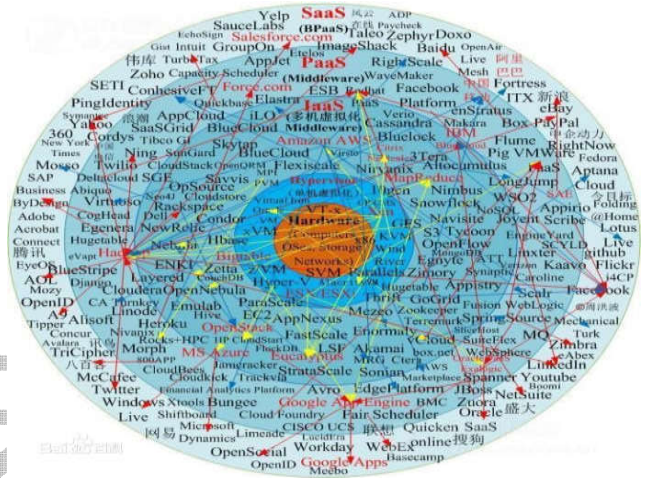
目录

大数据时代

什么是“大数据”？

大数据很抽象，表示数据规模的庞大。

- 本质上：大数据泛指巨量的数据集，因可从中挖掘出有价值的信息而受到重视。
- 技术上：大数据是一种对海量数据进行处理，从而抽取出能够创造价值的有意义的信息
- 商业上：大数据产业盛行疯长，大数据成为流行的商业模式。

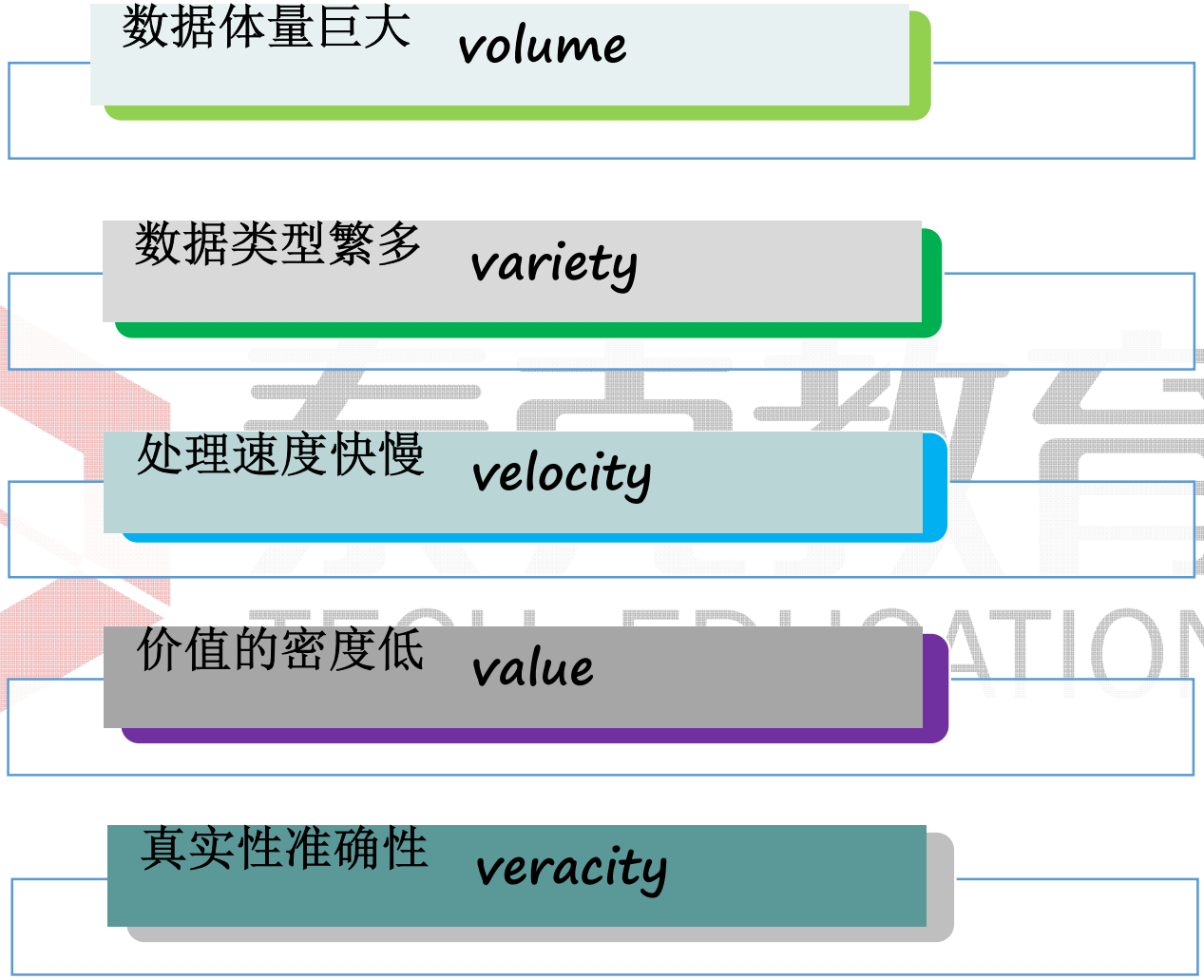


维基百科的定义：

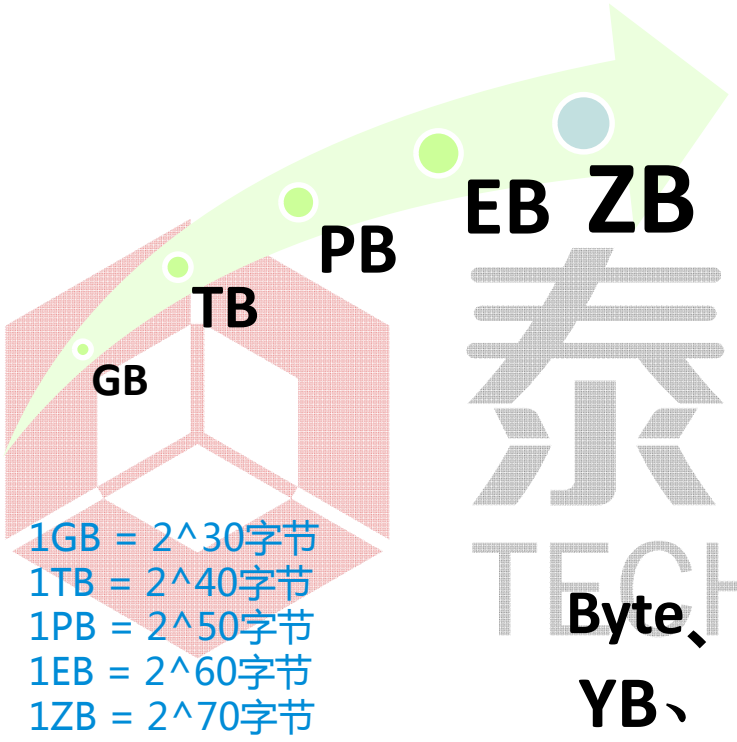
大数据是指利用常用软件工具捕获、管理和处理数据所耗时间超过可容忍时间的数据集。

——麦肯锡《大数据:下一个创新、竞争和生产力的前沿》

大数据特点



大数据时代的爆炸增长



1GB = 2^30字节
1TB = 2^40字节
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**2006** 年，个人用户才刚刚迈进**TB**时代，全球一共新产生了约**180EB**的数据；

在**2011** 年，这个数字达到了**1.8ZB**。

而有市场研究机构预测：
到**2020** 年，整个世界的数据总量将会增长**44** 倍，达到**35.2ZB (1ZB=10** 亿**TB) ！**

Byte、 KB、 MB
YB、 DB、 NB

1NB=2^100B=1267650600229401496703205360byte

大数据的来源

每天有超过2亿条消息

每天有超过3亿活跃用户

全球智能手机保有量28亿

每年售出数亿台支持GPS的设备

Facebook：每天产生**50TB**的日志数据，衍生分析数据超过**100TB**。

CERN：LHC对撞产生**1PB/s**的数据。

社交数据

机器数据

HUAWEI

所有生意都是数据生意

数据即生意

数据和客户同价

让数据流动

数据在变

数据即平台

流数据即商机

个性化大数据


物联网数据洪水


人工智能引领潮流

大数据时代已经到来

我国网民数量居世界之首，每天产生的数据量也位于世界前列。

淘宝网站	• 单日数据产生量超过**5万GB** • 存储量**4000万GB**
百度公司	• 目前数据总量**10亿GB** • 存储网页**1万亿页** • 每天大约要处理**60亿次**搜索请求
一个8Mbps的摄像头	• 一小时能产生3.6GB的数据 • 一个城市每月产生的数据达上千万GB
医院	• 一个病人的CT影像数据量达几十GB • 全国每年需保存的数据达上百亿GB

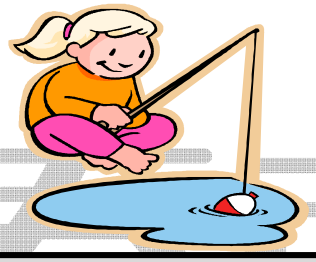
大数据时代已经到来

- 硬件成本的降低。
- 网络带宽的提升。
- 云计算的兴起。
- 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。

大数据时代与传统数据处理的差异

从数据库(database，DB)到大数据(big data，BD)

“池塘捕鱼”VS**“大海捕鱼”**，“鱼”是待处理的数据。



	数据库	大数据
数据规模	小（以MB为处理单位）	大（以GB、TB、PB为处理单位）
数据类型	单一（结构化为主）	繁多（结构化、半结构化、非结构化）
模式和数据的关系	先有模式后有数据 (先有池塘后有鱼)	先有数据后有模式，模式随数据增多不断演变
处理对象	数据（池塘中的鱼）	(“鱼”，通过某些“鱼”判断其他种类的“鱼”是否存在)
处理工具	One size fits all	No size fits all

目录

大数据时代引领未来

数据，已经渗透到每一个行业和业务领域，

洞见本质（业务）、预测趋势、指引未来是Big Data时代的核心

用未来牵引现在，用现在保证未来！

HUAWEI

大数据的浪潮

2012年3月29日奥巴马政府公布了”大数据研发计划”。该计划的目标是改进现有人们从海量和复杂的数据中获取知识的能力，从而加速美国在科学与工程领域发明的步伐，增强国家安全，转变现有的教学和学习方式。

谷歌搜索与流感预测

沃尔玛的啤酒与纸尿布

微博&投资

利用GPS数据了解交通状况

智能电表应用级家庭能源监测

Farecast与飞机票预测系统

大数据与乔布斯的癌症治疗

“魔毯”病人的监控

谷歌翻译系统

塔吉特预测少女怀孕

智慧城市&智能化交通

沃尔玛蛋挞与飓风用品的关系

大数据的应用领域

教育学　　情报学　　公共服务

企业管理　　市场营销　　总统选举

气候学

电子政务

天文学　　传媒业　　生活娱乐

金融学

生物医学　　商业智能　　图书馆学

For examples?

挑选男友…

21.06%男性买文胸80前后更体贴女友

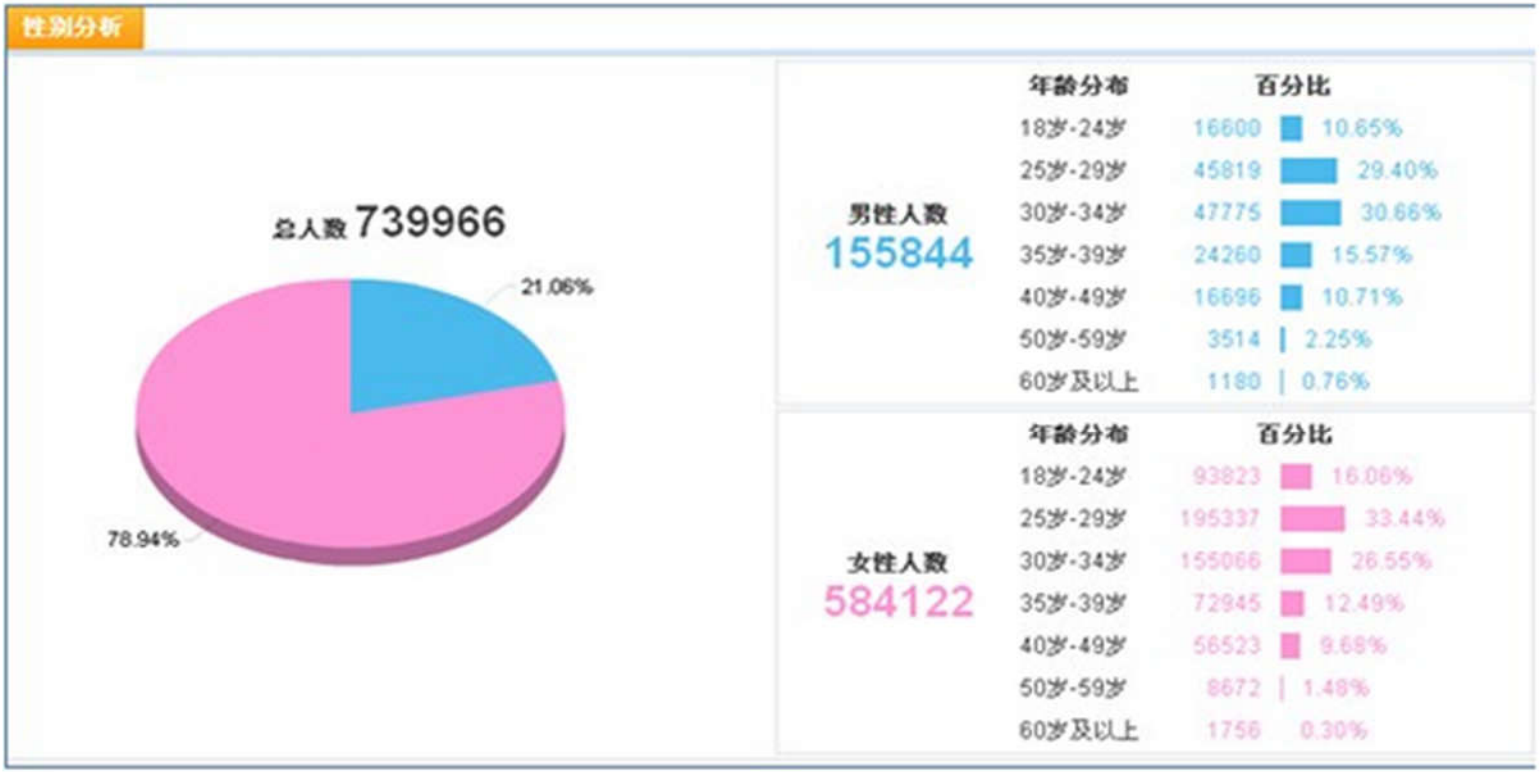


图5.文胸购买者性别年龄分析图片来源：淘宝数据平台

谁说文胸只能女人买？来自淘宝数据平台公布的数据显示，在6月14日至7月13日期间，共有739966人购买文胸。其中，有78.94%的购买者为女性，而另外21.06%的购买者却是男性。无论是男性还是女性，都以25岁至34岁为购买主力（如图5）。

找回自信

中国女性平均胸部B

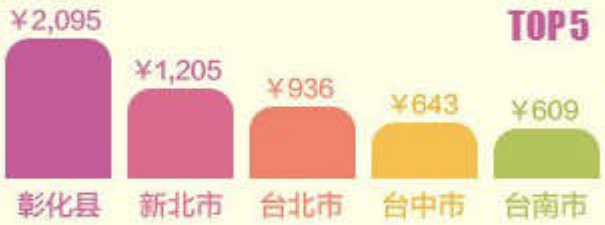
淘宝数据平台显示，杯，前9位中，B罩杯占比达41.45%，说明我国平均

其中，又以75B的销量ni的A罩杯，购买占比达25.26%，C罩杯则只有8.9

吃货集中营

食尚界

最爱吃零食的地方



零食人均购买金额最高的城市中，前五名居然都是台湾的！
台湾亲们真爱吃！

怪叔叔更爱买零食

买零食类男性买家中，年龄段最活跃的前三名分别是：
40-49岁、35-39岁、60岁以上。
怪叔叔打败了小萝莉。

最让国人上瘾的零食居然是槟榔

槟榔自古以来就是我国东南沿海各省居民迎宾敬客、款待亲
易让人上瘾，荣登最流行零食TOP1！令人意外的是：最爱购买

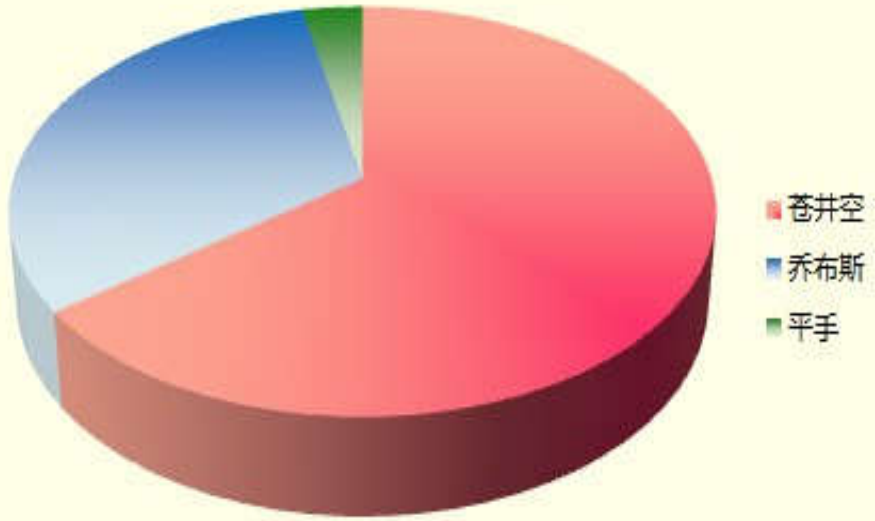
HUAWEI

当乔帮主遭遇苍老师

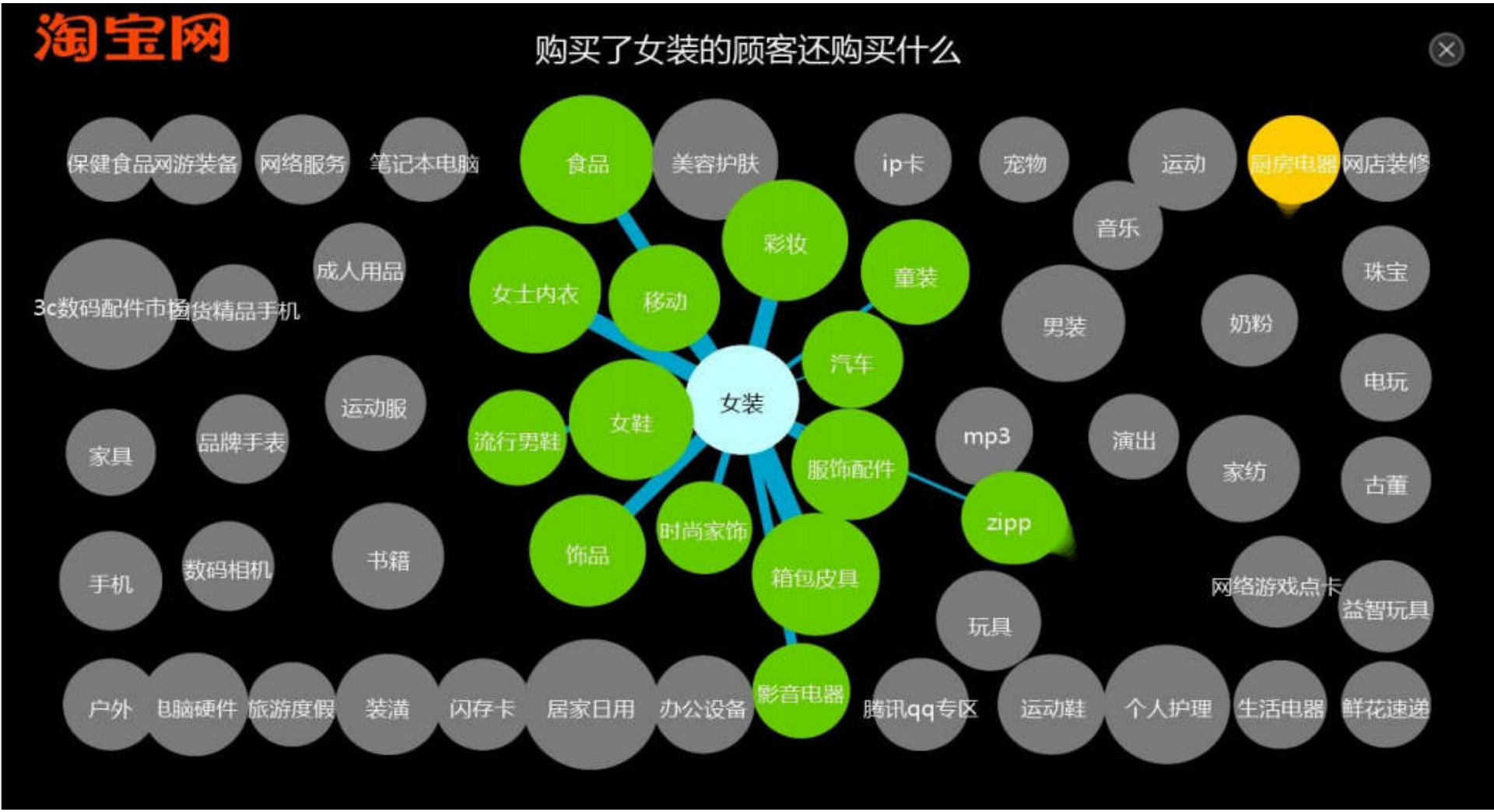
苍井空领先乔布斯 郭德纲入侵上海

淘宝网国内地区成交数据中，在260个地区购买苍井空相关商品的人数，多于该地区购买乔布斯相关商品的人数，在129个地区乔布斯后者领先前者，总体来说女神比教主更吸引眼球。

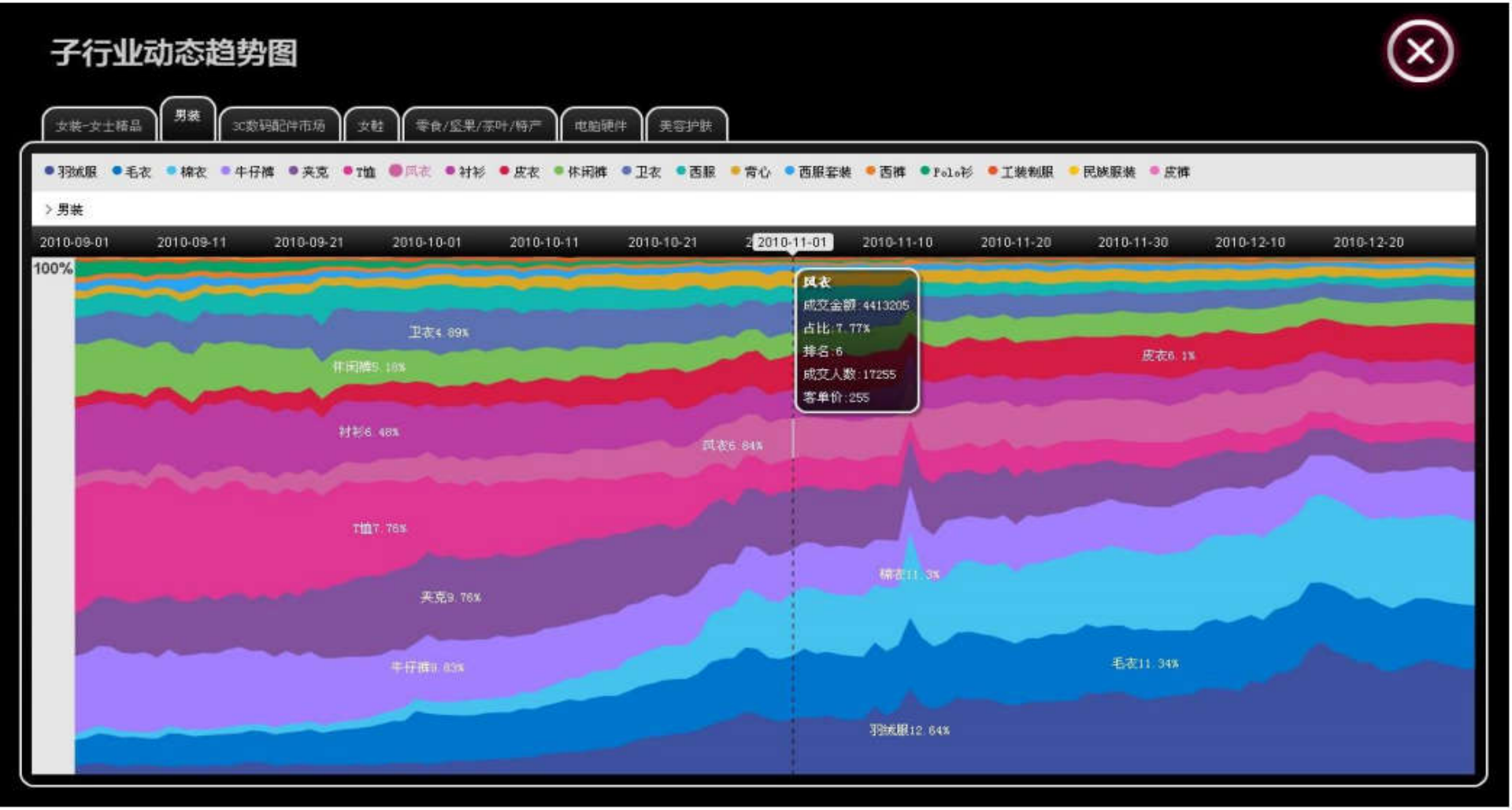
苍井空与乔布斯相关商品交易地区大PK



数据化运营...



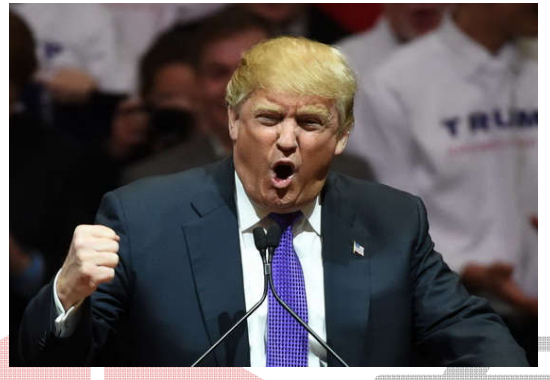
行业分析…



HUAWEI

大数据的应用领域

大数据的应用 – 政治



The cave
(数据分析中心)

大数据心理学分析帮助特朗普赢得美国大选

- 特朗普聘用CA公司对美国选民进行性格和需求分析，掌握了2.2亿美国人的个性。
- 利用选民在Facebook上的点赞行为，分析出他的性格特征已及政治取向，将选民分为三类，共和党支持者，民主党支持者，和摇摆者，重点拉拢摇摆不定的选民。
- 特朗普以前从没发过电子邮件，甚至在参加总统选举后才第一次购买智能手机，并迷上了发推特，而且他发出的每一条推特都是数据驱动的。针对不同的选民，都会有不同的微调版本。
- 针对非洲裔美国人，他们可以看到希拉里把黑人称为捕食者的视频，从而远离希拉里的投票箱，这些黑暗的帖子都是只有特定用户可见。

大数据的应用 – 金融

数据挖掘重要性

传统客户

固定时间地点获取服务
被动接受数据
寻求更多资源
信任市场信息
被动接受传播

随时随地获取服务
分析、创造数据
寻找有意义的体验
审视细节
互动参与内容、产品和体验的创建

新客户

运营客户

全渠道

关注场景

互动

分析

效率

推荐

营销客户

个性化的灵活服务

服务客户

传统金融

标准化和产业化提供服务
关注过程和步骤
被动接受信息且信息来源单一
通过客户经理联系客户
固定渠道单一交互

新金融机构

HUAWEI

大数据的应用 – 金融案例

Walmart 沃尔玛

东西岸时差4个小时

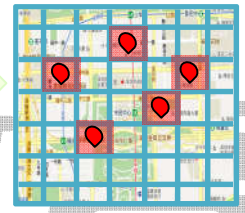
沃尔玛

同一天，东海岸的销售分析结果指导西海岸的商品如何摆放。

HUAWEI

大数据的应用--商业

沃尔玛



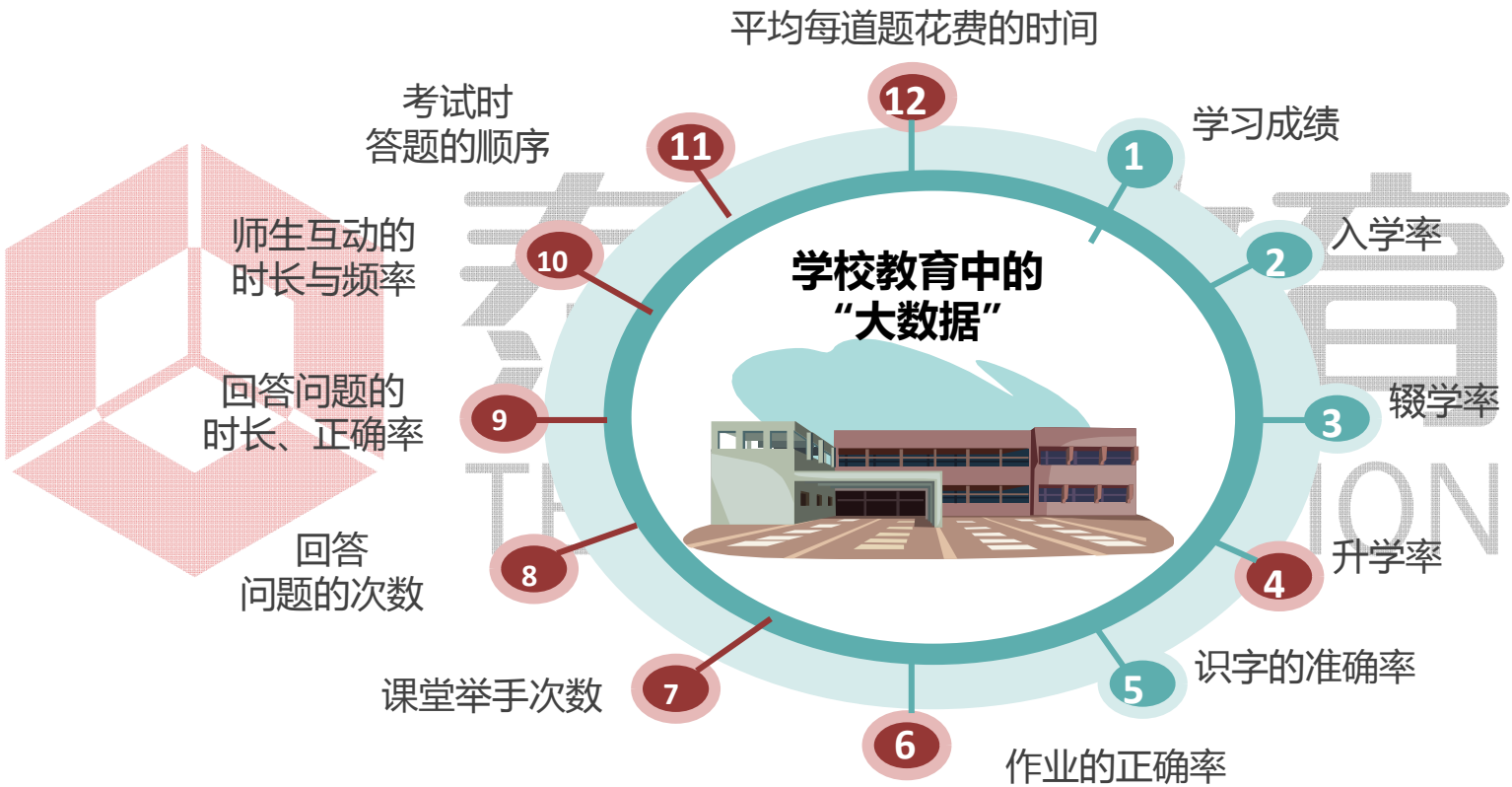
类型选择 → 开店自定义 → 自助辅助查看 → 开店候选区域

开店目标区域

开店目标人群

大数据的应用 – 教育

现在，大数据分析已经被应用到美国的公共教育中，成为教学改革的重要力量。

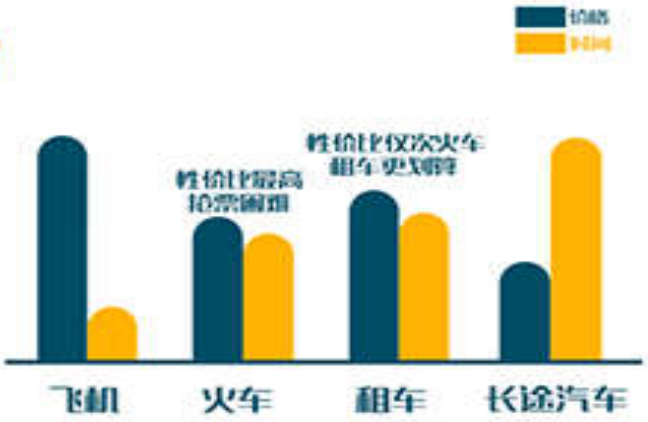


大数据的应用 – 出行

500公里内大多选择铁路出行，但...

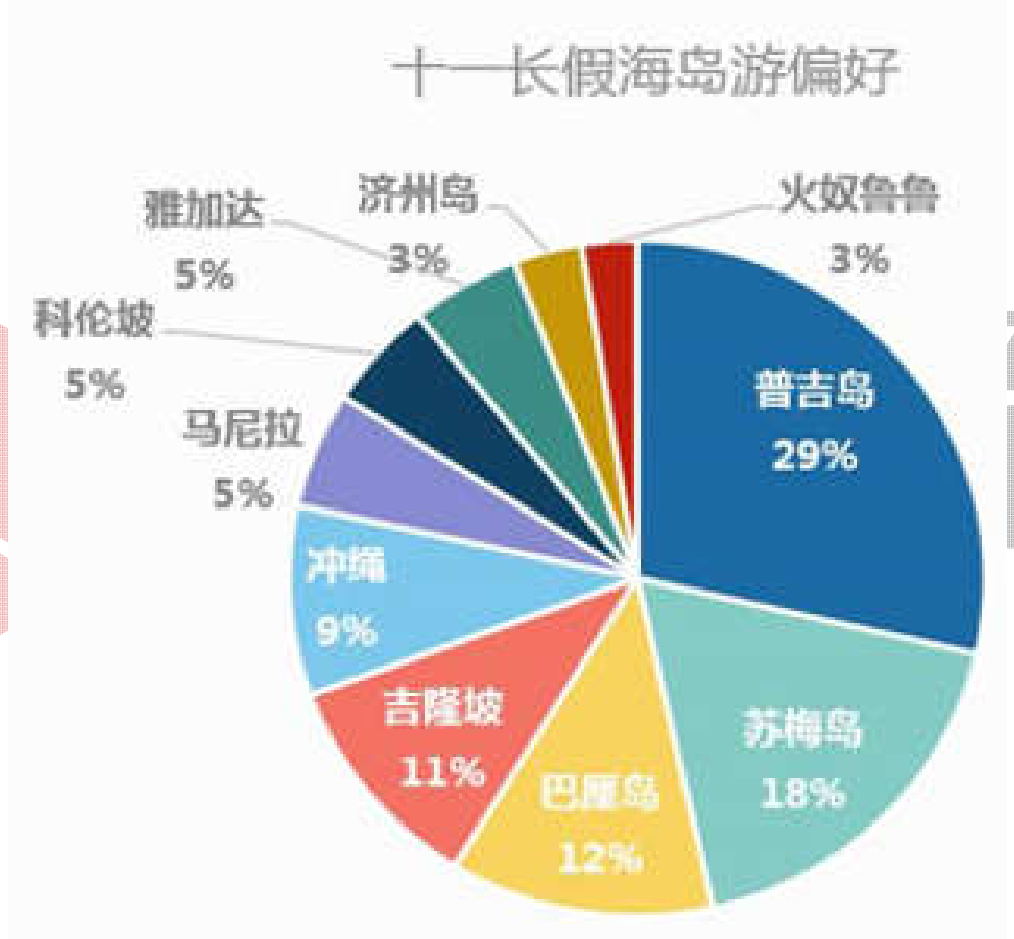
500公里以内出行方式 示例：
北京—太原

2018年春运出行方式

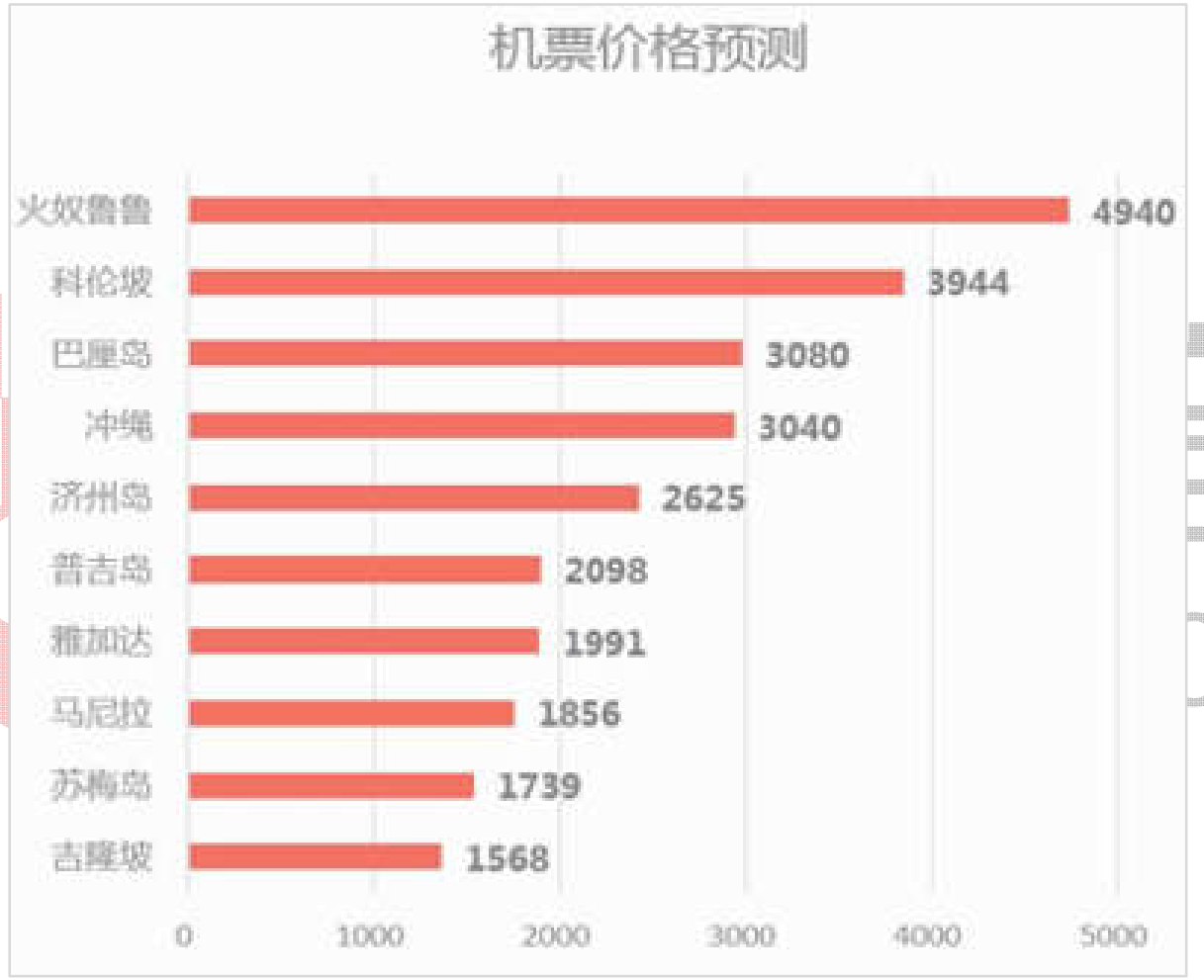


500公里6小时车程范围内，铁路出行虽然性价比最高，但是抢到票的机率看脸看神看运气：租车性价比仅次于火车，据调查在抢不到火车票的情况下，超过70%的人选择租车回家。

大数据的应用 – 旅游



大数据的应用 – 旅游



大数据的应用 – 政府公共安全

公共安全场景 - 自动预警与联动

人流异常增多预警

可满足城市人口流向监测分析。

自动预警系统：
富华大厦右侧异常人群超警戒。

监管部门即时定位初期问题。

发送确认

因群体性斗殴，而引发群众聚集围观。

通报上级部门

城市/社区监控系统确认

事务处理职务部门

HUAWEI

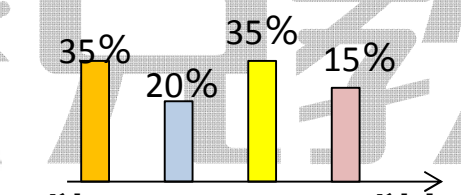
大数据的应用 – 交通规划

交通规划场景 - 多维度交通人群分析

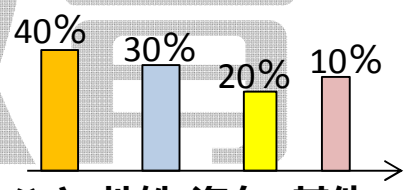
历史人流量超阈值区域

- 工体北门：人流超500/H
- 三里屯：人流超800/H
- 北京工体：人流1500/H

按人群分析



按出行方式比例分析



结合人群的交通预测建议

路网规划

公交线路规划

HUAWEI

大数据的应用 – 体育

目录

各国都将大数据作为国家战略

美国

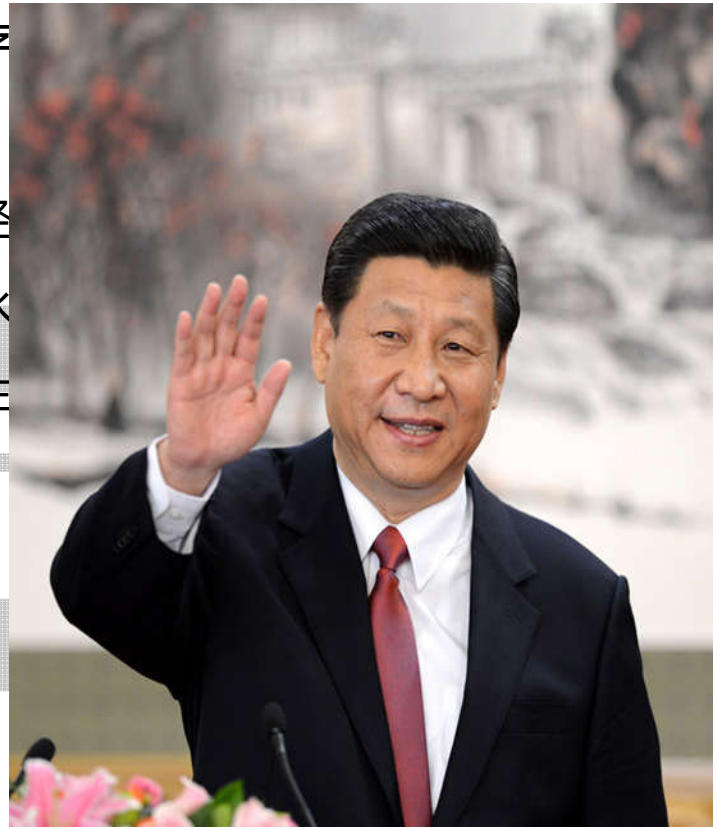
G8

英国

- 八国集团发布了《G8开放数据宪章》，提出要加快推动数据开放和利用。
- 欧盟力推《数据价值链战略计划》，用大数据改造传统治理模式，降低公共部门成本，并促进经济增长和就业增长。
- 安倍内阁正式公布新IT战略《创建最尖端IT国家宣言》，以开放大数据为核心的IT国家战略。
- 英国政府发布《英国数据能力发展战略规划》，旨在利用数据产生商业价值、提振经济增长，承诺开放交通、天气、医疗方面的核心数据库。

中国实施国家大数据战略

- 实施国家大数据战略加快建设数字
  - 要推动大数据技术产业创新发展；
  - 要构建以数据为关键要素的数字经
  - 要运用大数据提升国家治理现代水
  - 要运用大数据促进保障和改善民生
  - 要切实保障国家数据安全。



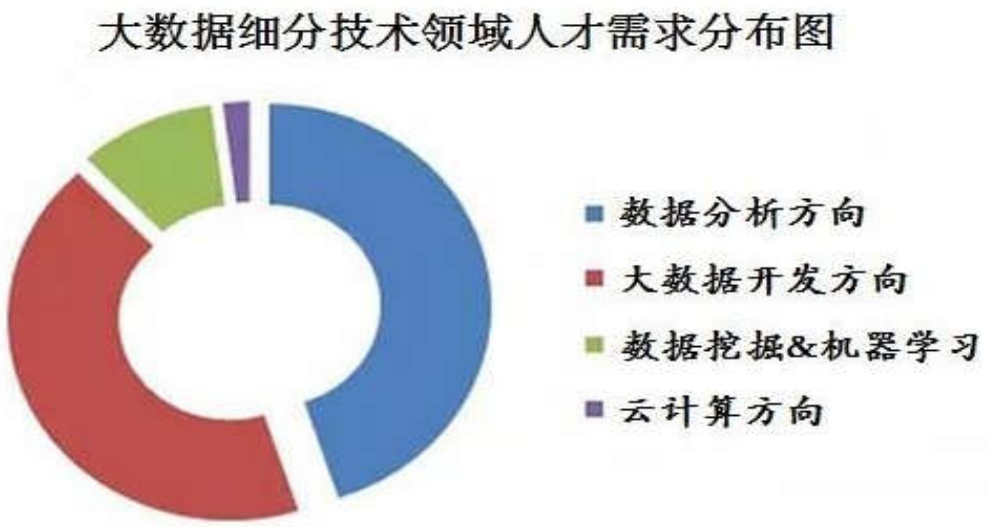
大数据的市场规模

工信部正式印发的《大数据产业发展规划（2016-2020年）》估计的。

到2020年，技术先进、应用繁荣、保障有力的大数据产业体系基本形成。大数据相关产品和服务业务收入突破1万亿元，年均复合增长率保持30%左右，加快建设数据强国，为实现制造强国和网络强国提供强大的产业支撑。

大数据的人才需求

全球顶尖管理咨询公司麦肯锡(McKinsey)出具的一份详细分析报告显示：预计到 2018年，大数据或者数据工作者的岗位需求将激增，其中大数据科学家的缺口在14万到19万之间，对于懂得如何利用大数据做决策的分析师和经理的岗位缺口则将达到150万。根据中国数据分析行业网的数据显示，目前全国的大数据人才只有46万，未来3-5年内大数据人才的缺口将高达150万多，大数据行业将面临全球性的人才荒。



目录

大数据发展所需要的人才类型

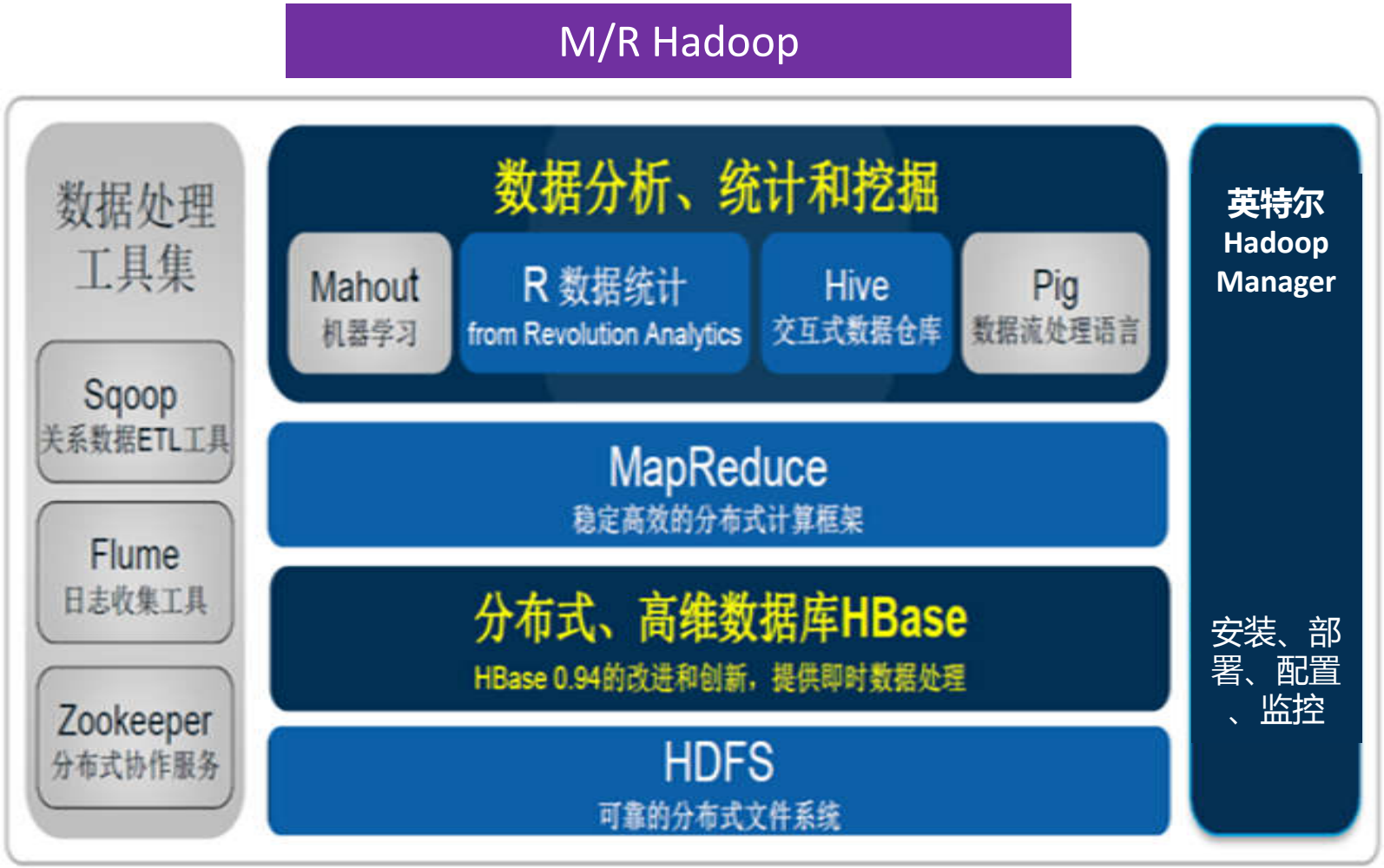
- 大数据系统研发工程师。
- 大数据应用开发工程师。
- 大数据分析师。
- 数据可视化工程师。
- 数据安全研发工程师。
- 数据科学研究人才。

大数据时代的机遇

机遇 - 大数据蓝海成为企业竞争的新焦点

大数据所能带来的巨大商业价值，被认为将引领一场足以与20世纪计算机革命匹敌的巨大变革。大数据正在对每个领域都造成影响，包括商业、经济等领域。大数据正在促生新的蓝海，催生新的经济增长点，正在成为企业竞争的新焦点。

Hadoop业界参考实践



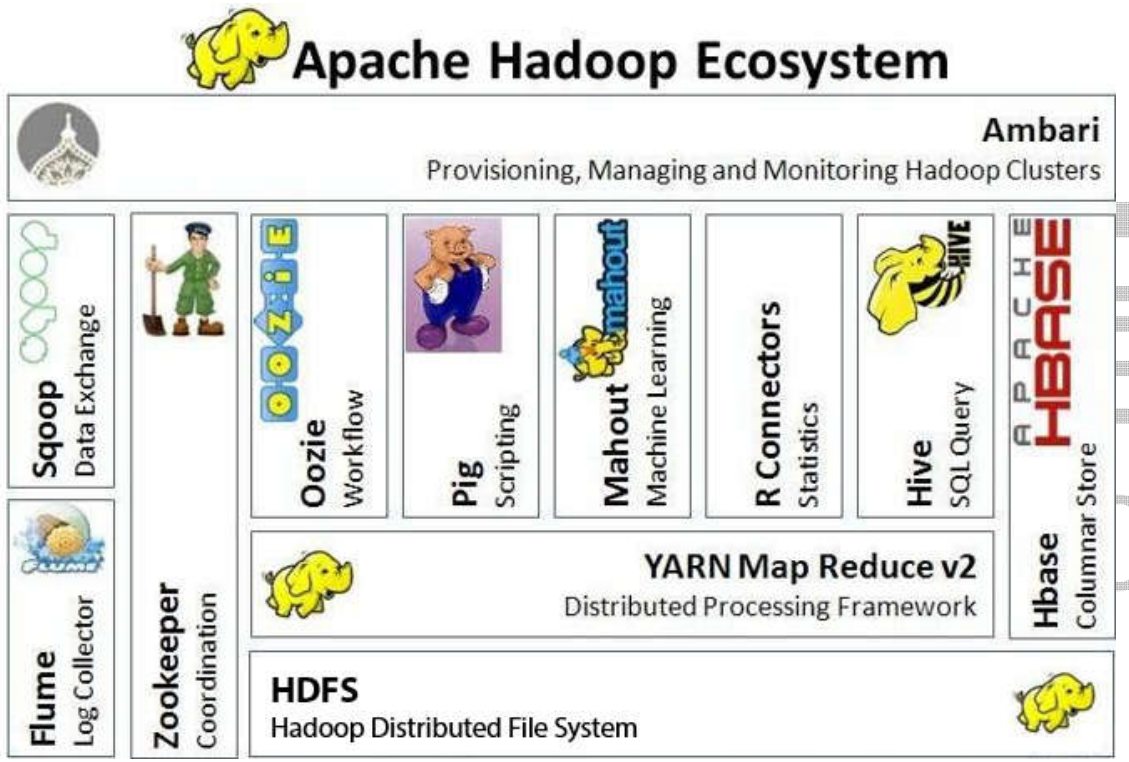
流计算业界参考实践

IBM InfoSphere Streams是IBM BigData战略的核心部件之一，支持结构化、非结构化数据的高速处理，Processing in Motion，每秒支持百万事件吞吐，支持高可扩展，支持SPL语言。

Hstreaming ,对Hadoop M/R框架进行流化改造，可以和现有主流Hadoop基础设施兼容。在不/微小改动的前提下，以流式M/R的模式处理数据，曾被Gartner评为最Cool ESP厂商，现在已经可以支撑文本和视频处理，采用PIG语言，提供Hadoop的高可扩展性能力，数百万件每秒吞吐和毫秒级时延。

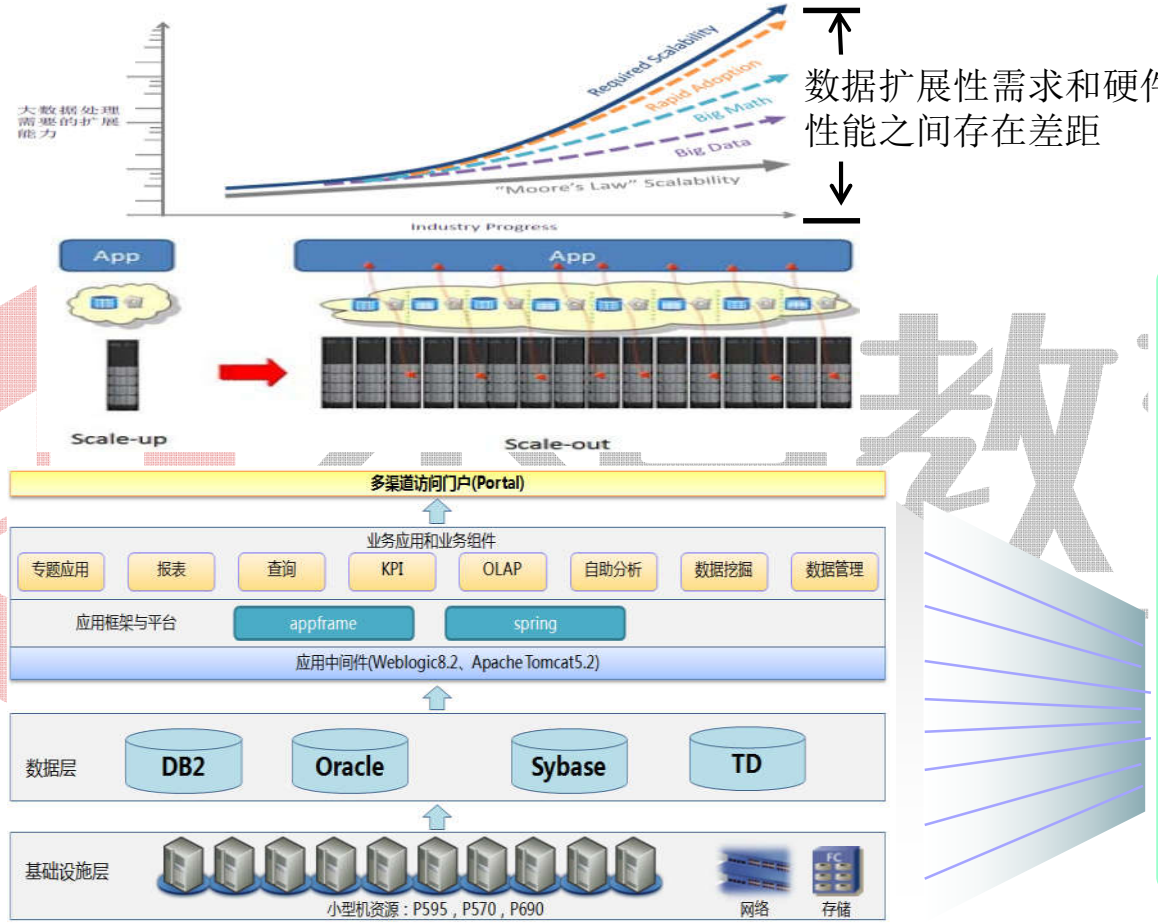
从批量处理到实时分析

Hadoop成为大数据批量处理的基础，但无法提供实时分析



对高价值高度聚合的信息和知识实时智能化是主要商业诉求

面对挑战，传统数据处理遭遇天花板



- 海量数据的高存储成本
- 数据批量处理性能不足
- 流式数据处理缺失
- 有限的扩展能力
- 单一数据源
- ……

传统框架：小型机＋磁阵＋商用数据仓库

大数据时代的机遇和挑战

当你有了锤子，好像什么问题都看上去像钉子！

今天，大数据似乎成了“灵丹妙药”，“包治百病”，无所不能。但千万别把“大数据”用做解决世界上所有问题的全能办法，人类的思想、个人的文化和行为模式、不同国家及社会的存在发展都非常复杂、曲折和独特，显然不能全部由计算机来“数字自己说话”。**无论到何时，其实都还是人在思考和“说话”。**

大数据不能做什么？

- **不能替代管理的决策力**

人群聚集：群架or演唱会？

- **不能替代有效的商业模式**

如何盈利，线上or线下？

- **不能无目的地发现知识**

华尔街的小目标

- **不能替代专家的作用**

AlphaGo的起始

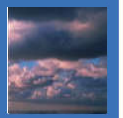
- **不能一次建模终生受益**

反馈、学习、更新

企业级大数据平台应用场景

经营类	管理类	监管类	专业类
经营分析 电信信令 金融细账 金融票据 电力调度 智能电网	绩效管理 报表分析 历史分析 社保分析 纳税分析 决策支持和预测	公安网监 国安技侦 舆情监控 银监会稽查 食品溯源 环保监测	音视频 地震勘探 气象云图 卫星遥感 雷达数据 物联网
电信、金融	金融	政府	政府
结构化	结构化+半结构化	结构化+半结构化	非结构化



电信、金融、政府等行业数据分析的诉求强烈，互联网已开始应用新技术处理价值密度低的大数据。

企业所面临的挑战（1）

- 挑战一：

业务部门无清晰的大数据需求

很多企业业务部门不了解大数据，也不了解大数据的应用场景和价值，因此难以提出大数据的准确需求。由于业务部门需求不清晰，大数据部门又是非盈利部门，企业决策层担心投入产出比不高，在搭建大数据部门时犹豫不决，甚至由于暂时没有应用场景，删除了很多有价值的历史数据。

企业所面临的挑战（2）

- 挑战二：

企业内部数据孤岛严重

企业启动大数据最重要的挑战就是数据的碎片化。在大型企业中，不同类型的数据常常散落在不同部门，使得同一企业内部数据无法共享，无法发挥大数据的价值。

企业所面临的挑战（3）

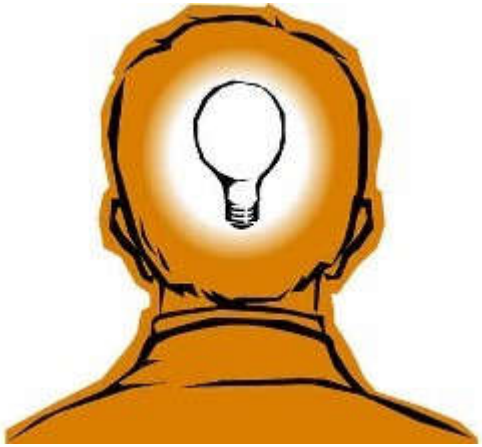
- 挑战三：

数据可用性低，质量差

很多大中型企业每天会产生大量的数据，但很多企业在大数据的预处理阶段很不重视，导致数据处理很不规范。大数据预处理阶段需要抽取数据把数据转化为方便处理的数据类型，对数据进行清洗和去噪，以提取有效的数据等操作。Sybase的数据表明，高质量的数据可用性提高10%，企业效益提高10%以上。

企业所面临的挑战（4）

- 挑战四：**数据相关管理技术和架构**
  - 传统的数据库部署不能处理百TB及以上级别的数据。
  - 传统的数据库没有考虑数据的多样性，尤其对结构化数据，半结构化数据和非结构化数据的兼容。
  - 传统的数据库对数据处理时间要求不高，而大数据需要实时处理数据。
  - 海量数据运维需要保证数据稳定，支持高并发的同时减少服务器负载。



企业所面临的挑战（5）

- 挑战五：

数据安全

网络化生活使得犯罪分子更容易获得关于人的信息，也有了更多不易被追踪和防范的犯罪手段。

如何保证用户的信息安全成为大数据时代非常重要的课题。此外，大数据的不断增加，对数据存储的物理安全性要求会越来越高，从而对数据的多副本与容灾机制也提出更高的要求。

企业所面临的挑战（6）

- 挑战六：

大数据人才缺乏

大数据建设的每一个换件都需要依靠专业人员完成，因此必须培养和造就一支掌握大数据，懂管理，有大数据应用经验的大数据建设专业队伍。全球每年将新增数十万个大数据相关的工作岗位，未来将会出现100万以上的人才缺口。因此高校和企业共同努力去培养和挖掘人才。

企业所面临的挑战（7）

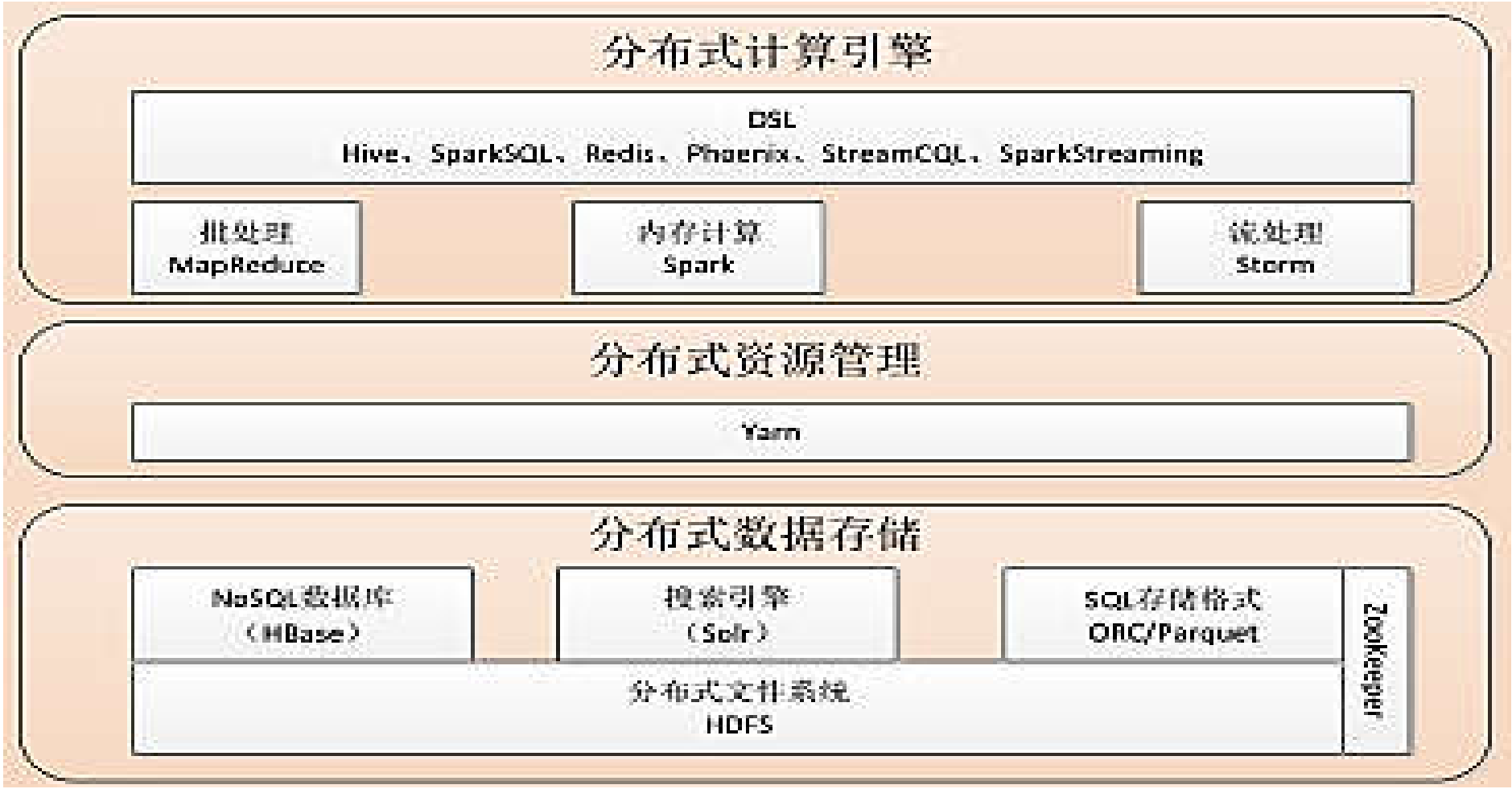
- 挑战七：

数据开放与隐私的权衡

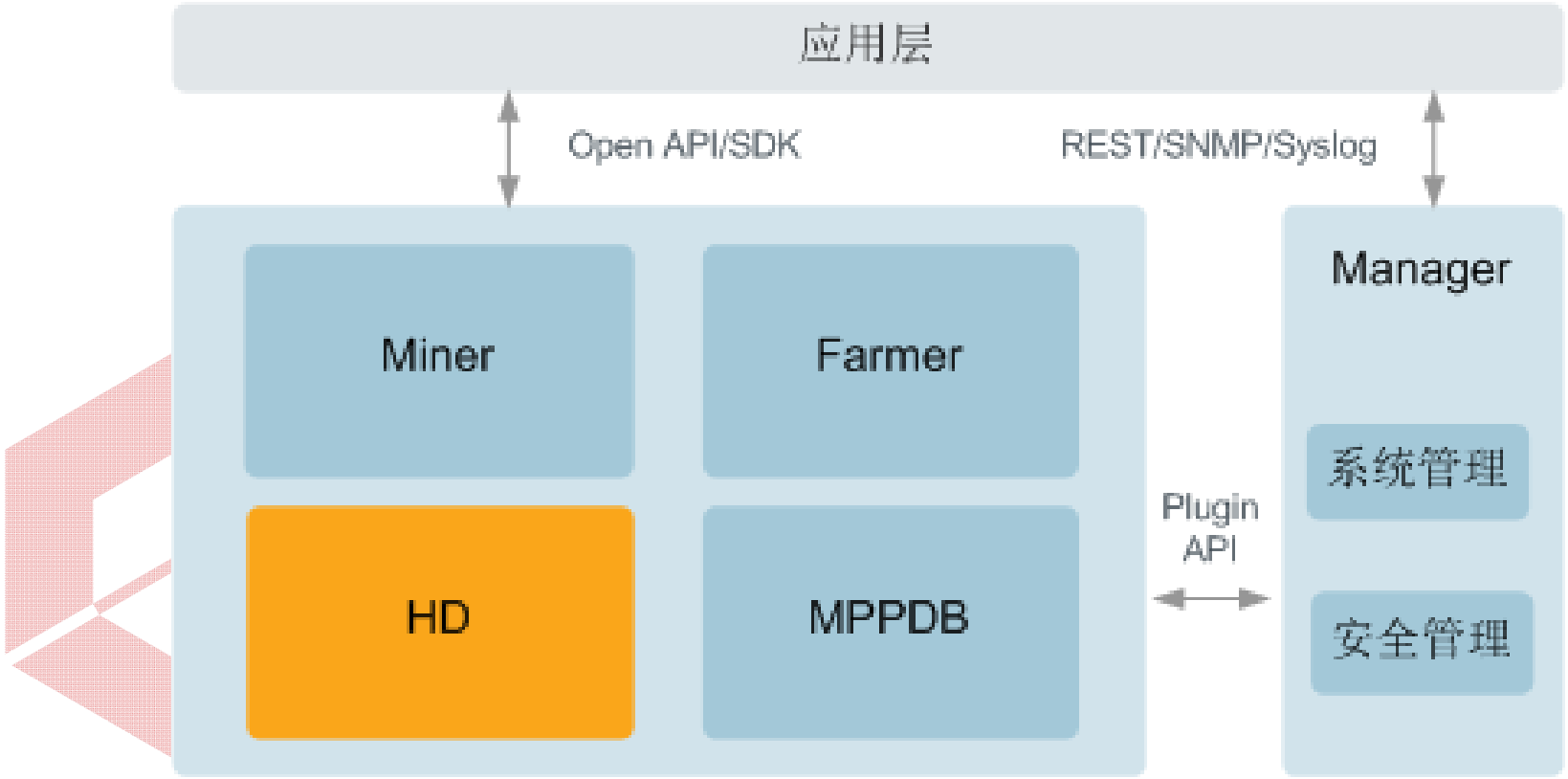
在大数据应用日益重要的今天，数据资源的开放共享已经成为在数据大战中保持优势的关键。但是数据的开放不可避免的会侵害一些用户的隐私。如何在推动数据全面开放，应用和共享的同时有效地保护公民，企业隐私，逐步加强隐私立法，将是大数据时代的一个重大挑战。

目录

通用大数据平台框架

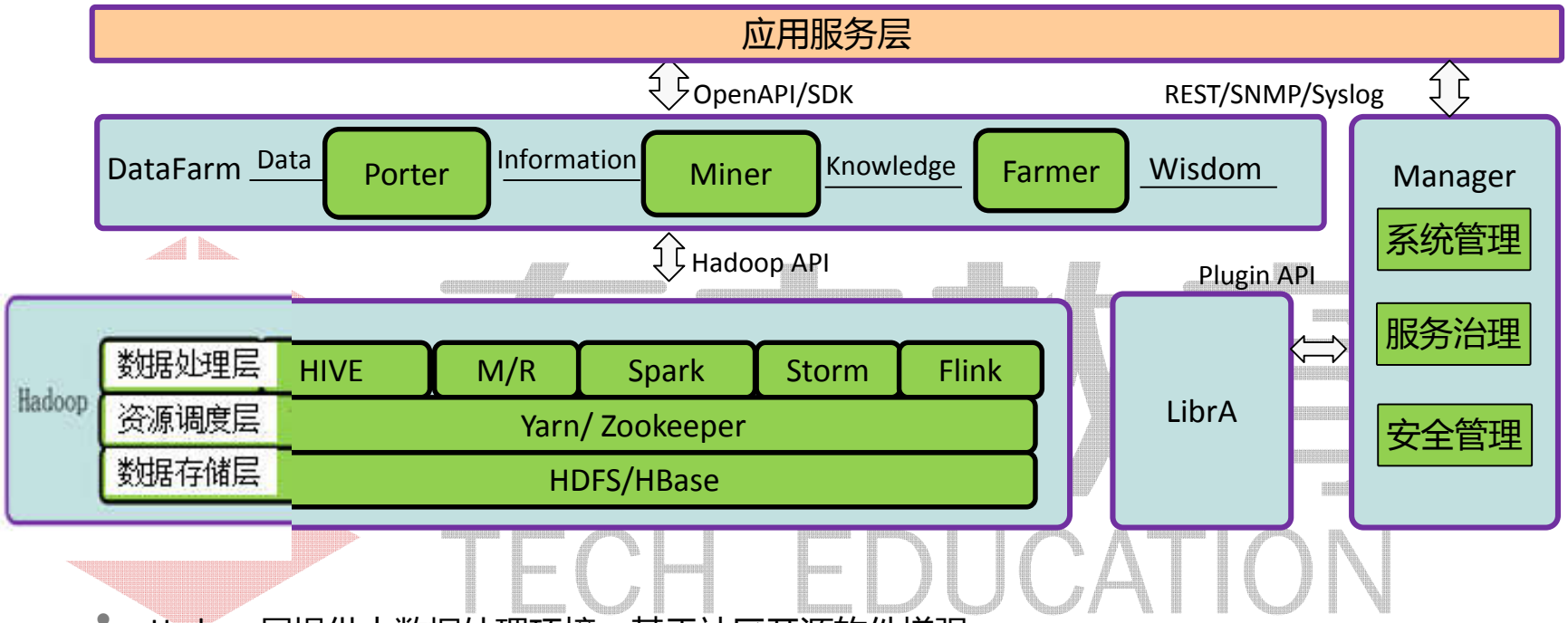


华为大数据平台架构框架



FusionInsight 解决方案由 5 个子产品 FusionInsight HD、FusionInsight MPPDB、FusionInsight Miner、FusionInsight Farmer 和 FusionInsight Manager 构成。

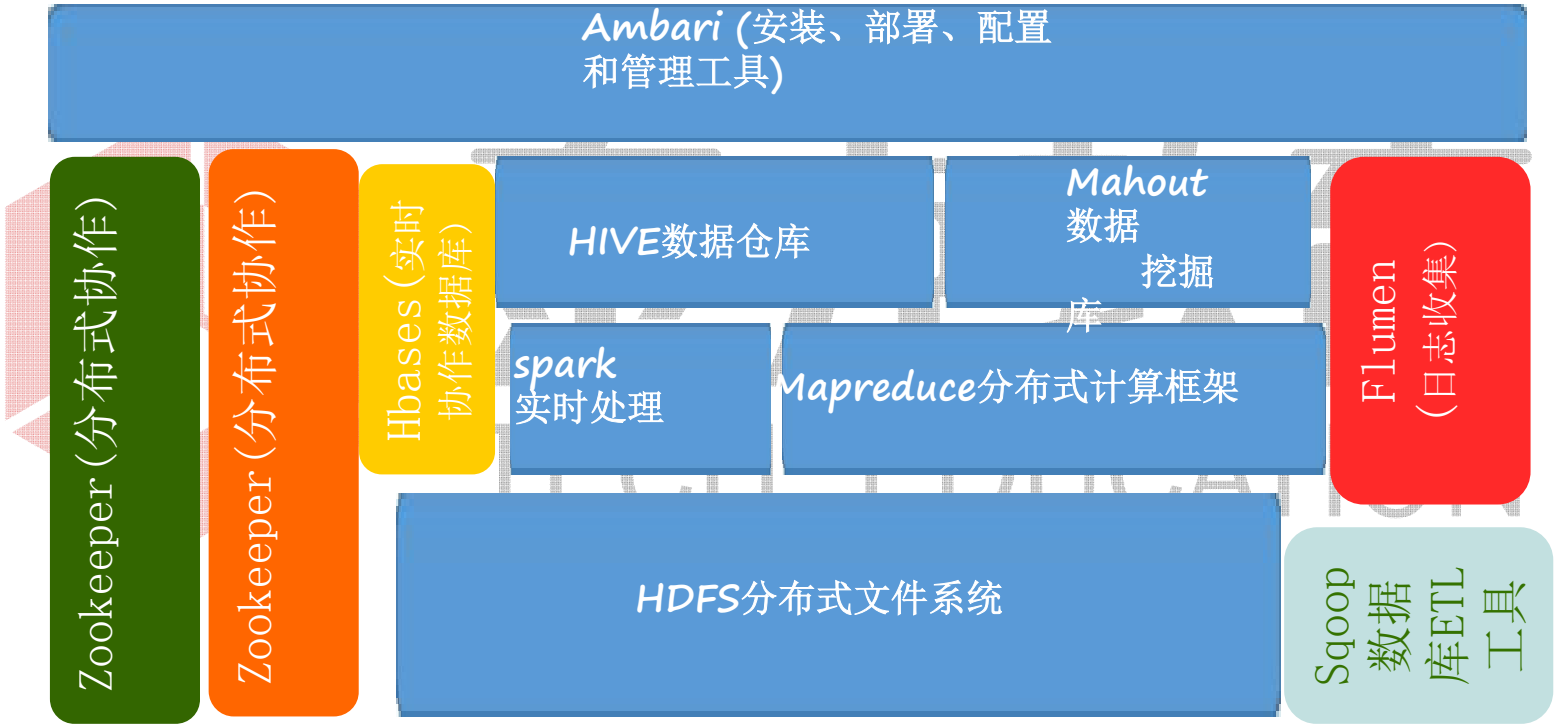
华为大数据平台架构框架



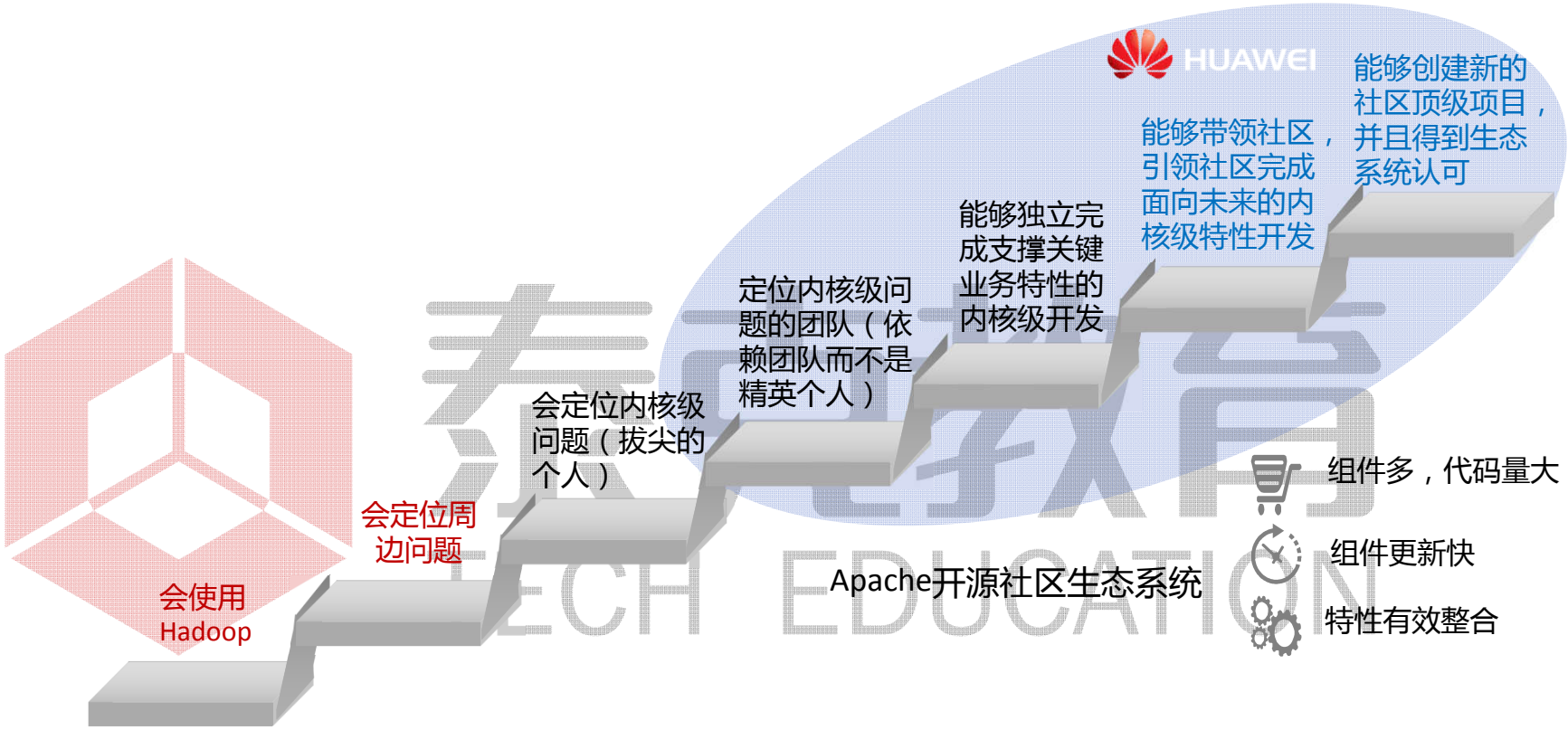
- Hadoop层提供大数据处理环境，基于社区开源软件增强。
- DataFarm层提供支撑端到端数据洞察，构建数据到信息到知识到智慧的数据供应链， 其中数据集成服务Porter，数据挖掘服务Miner和数据服务框架Farmer。
- Manager是一个分布式系统管理框架，管理员可以从单一接入点操控分布式集群，包括系统管理、数据安全管理和数据治理。

普通公司大数据平台框架

Hadoop 是一个能够对大量数据进行分布式处理的软件框架。具有可靠、高效、可伸缩的特点。
Hadoop 的核心是 HDFS 和 Mapreduce，hadoop2.0 还包括 YARN。



华为大数据团队核心能力



强大的Hadoop内核团队支持的开发与产品交付能力，电信级运营支撑能力。

金融与运营商大数据平台合作伙伴

Industrial and Commercial Bank of China
工商银行

China Merchants Bank
招商银行

Pacific Insurance
太平洋保险

China Mobile
中国移动

China Unicom
上海联通

Top 3
China Telco中国运营商企业

50%
China Financial Industry Top 10 Customers
中国金融行业Top10企业

本章总结

- 本章介绍了我们当前已经处在大数据的时代，并介绍了大数据已经成功应用在生活中的各行各业。然后抛出当前大数据给我们带来的基于和挑战，最后介绍在此机遇下，华为大数据解决方案。
- 简单介绍了我们这星期要学的一些工具组件

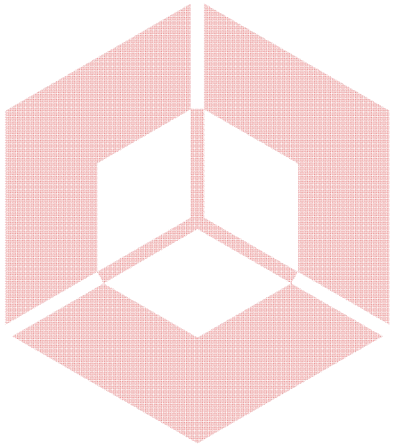
  HDFS　Map Reduce　streaming

  Spark　miner　Hbase　Hive

  FusionInsight　HD

思考题

1. 大数据从什么地方来？这些数据有哪些特点？
2. 大数据可以应用在哪些社会领域？
3. 华为大数据解决方案叫什么？



谢谢

www.huawei.com

泰克教育

TECH EDUCATION