The 8th China Cloud Computing Conference
第八届中国云计算大会
技术融合 应用创新

云计算基础架构与大数据应用实践

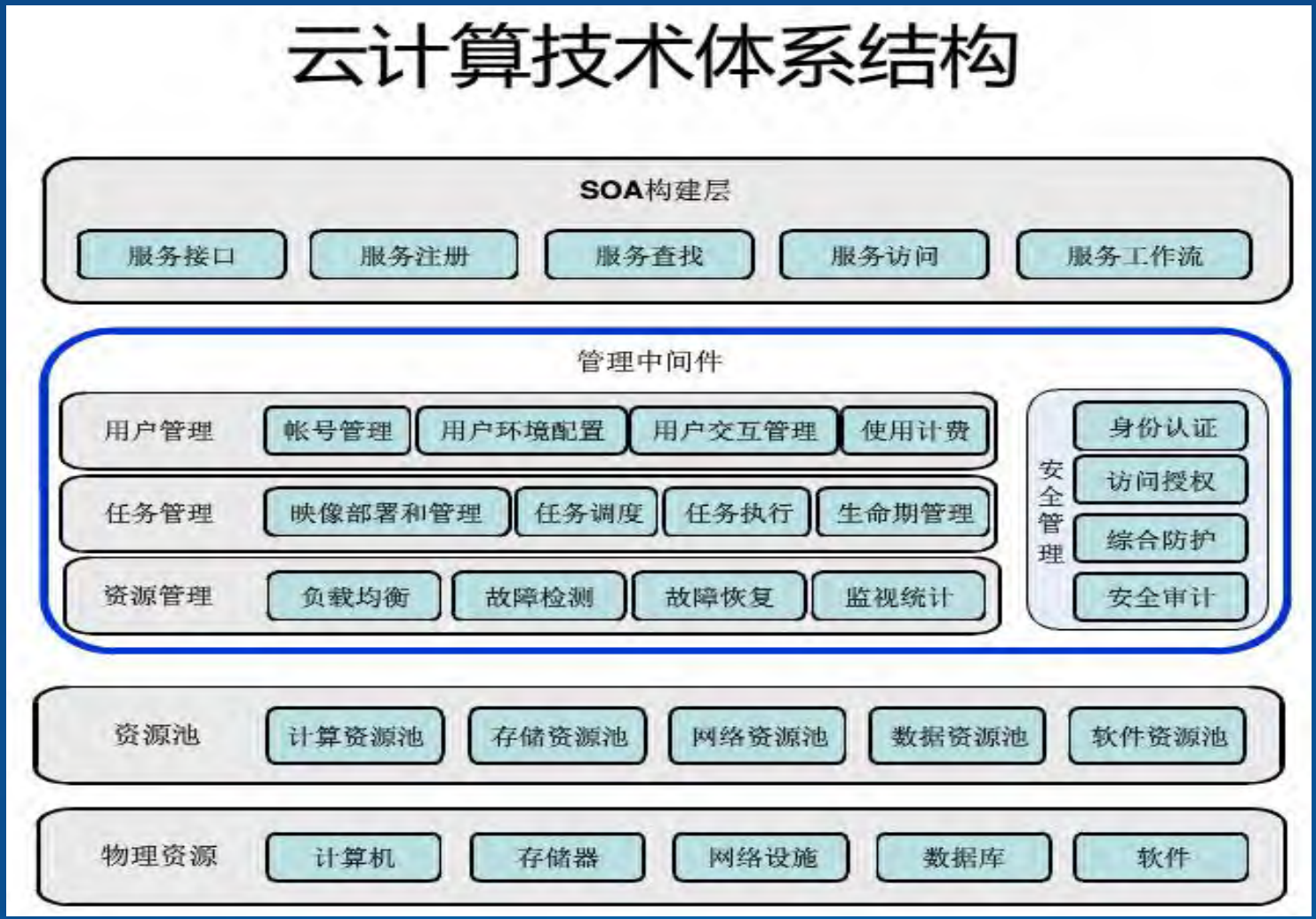
2016年5月19日 北京 张应福

主要内容

- 云计算基础架构及应用实践
- 大数据技术与业务应用实践

判断是不是云计算的三条标准

1. 用户使用的资源不在客户端而在网络中。
2. 服务能力具有优于分钟级的可伸缩性。
3. 五倍以上的性价比提升。


云计算技术体系结构
SOA构建层
服务接口
服务注册
服务查找
服务访问
服务工作流
管理中间件
用户管理
帐号管理
用户环境配置
用户交互管理
使用计费
任务管理
映像部署和管理
任务调度
任务执行
生命期管理
资源管理
负载均衡
故障检测
故障恢复
监视统计
安全管理
身份认证
访问授权
综合防护
安全审计
资源池
计算资源池
存储资源池
网络资源池
数据资源池
软件资源池
物理资源
计算机
存储器
网络设施
数据库
软件

云计算 = 平台 + 服务

云计算
- 云服务
  - 各种应用软件和服务
  - — 业务与应用软件
- “云计算”的灵魂
  - — 云平台软件
- 云平台
  - 连接数以万计服务器的以太交换机
  - 支撑海量信息处理的服务器和存储

云计算提供的服务类型

软件服务-SaaS，Software as a Service


办公应用


信息化应用


通讯应用


互联网应用

平台服务-PaaS, Platform as a Service


开发环境


运行环境
(Lib)


身份认证

基础设施服务- IaaS，Infrastructure as a Services

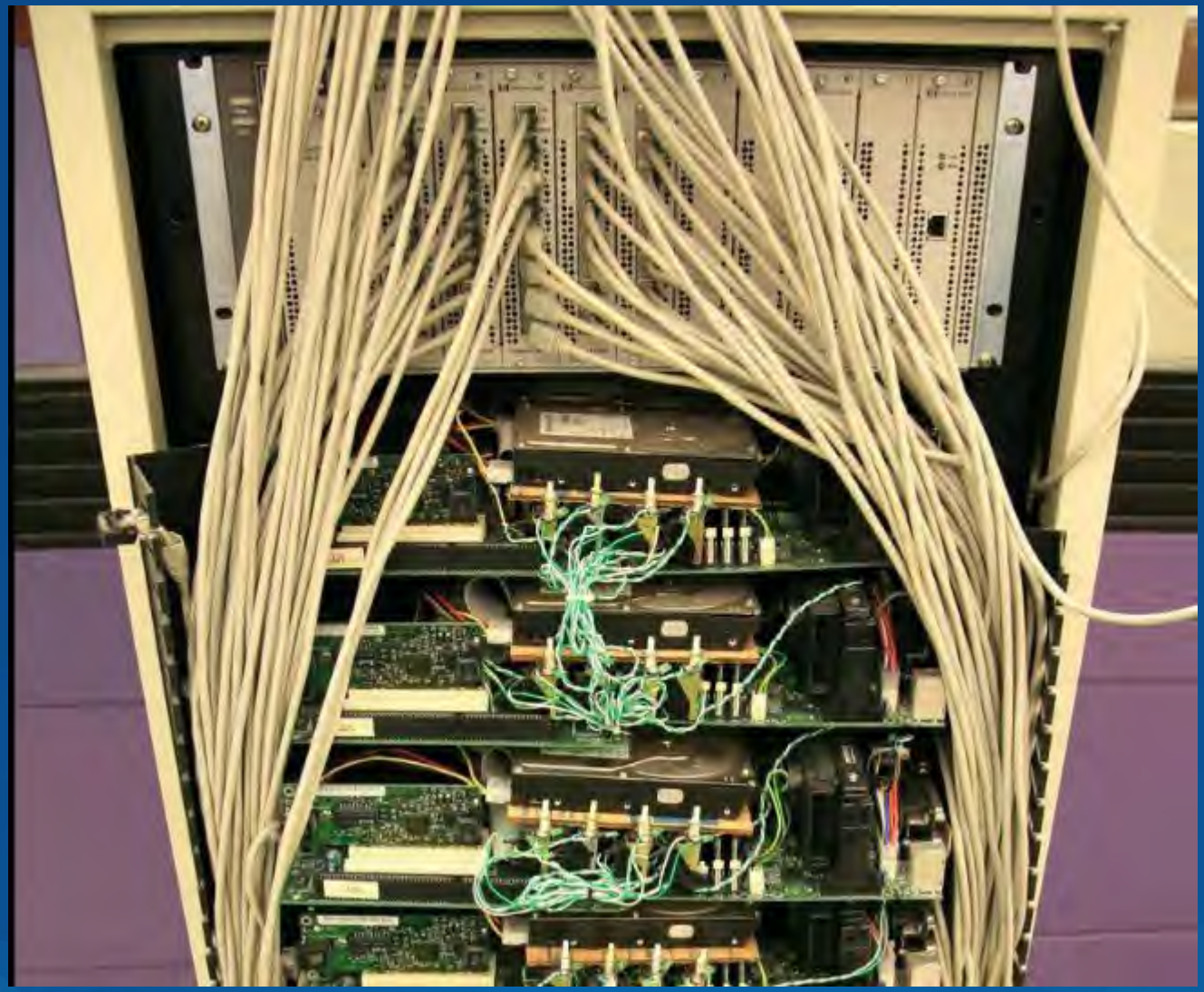

服务


存储服务


服务

云计算的核心理念

在一大堆烂机器上提供高性能可靠服务。



虚拟化过程与目标

共享和隔离

在单一物理服务器上可同时运行多个虚拟机，同时虚拟机之间相互隔离。提高资源利用率，降低能耗。

资源弹性

虚拟机可以根据其需求弹性增加或减少其分配的硬件资源。提高资源配置的灵活性

封装

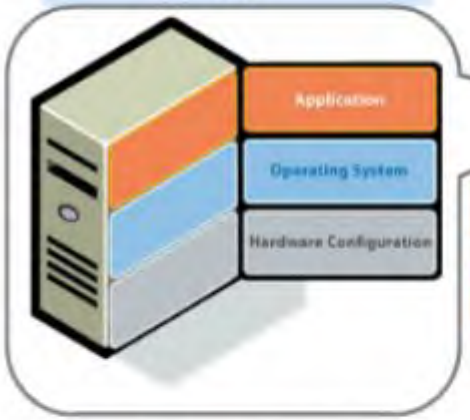
虚拟机将整个系统，包括硬件配置、操作系统以及应用等封装在文件里。用于系统快速部署、软件发布、系统备份

迁移

可以在不同服务器上不加修改直接迁移正在运行的虚拟机，增强系统的可靠性和可扩展性。

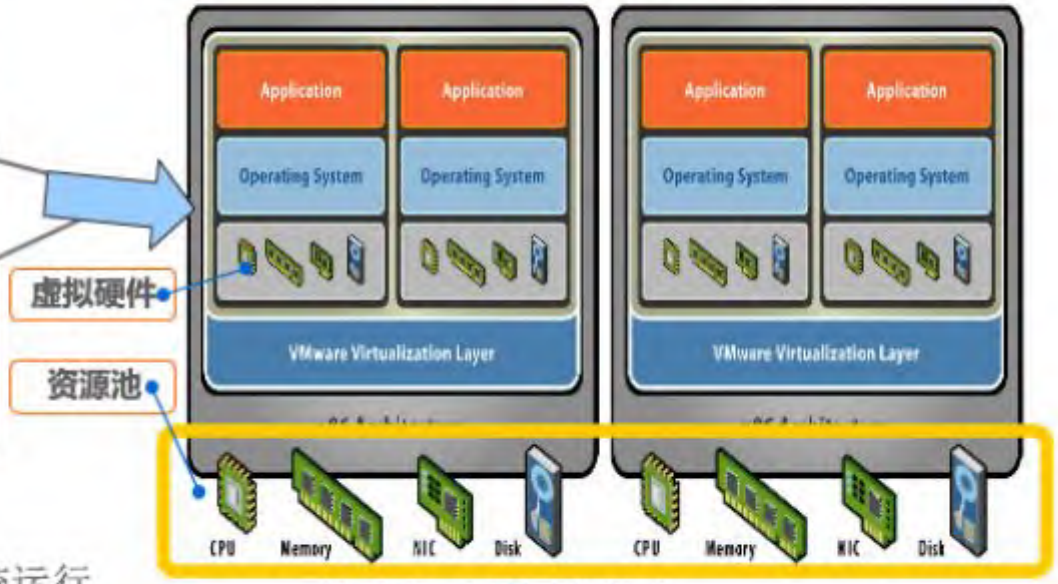
虚拟化技术

传统计算系统计算模式



- 软件必须与硬件相结合
- 每台机器上只有一个系统运行
- 操作系统只能管理本地服务器

虚拟化计算系统计算模式



- 软件相对于硬件独立
- 每台机器上可以有多个系统运行
- 虚拟化管理软件能管理多台服务器

中国电信云计算

中国电信“8+2+x”云资源布局

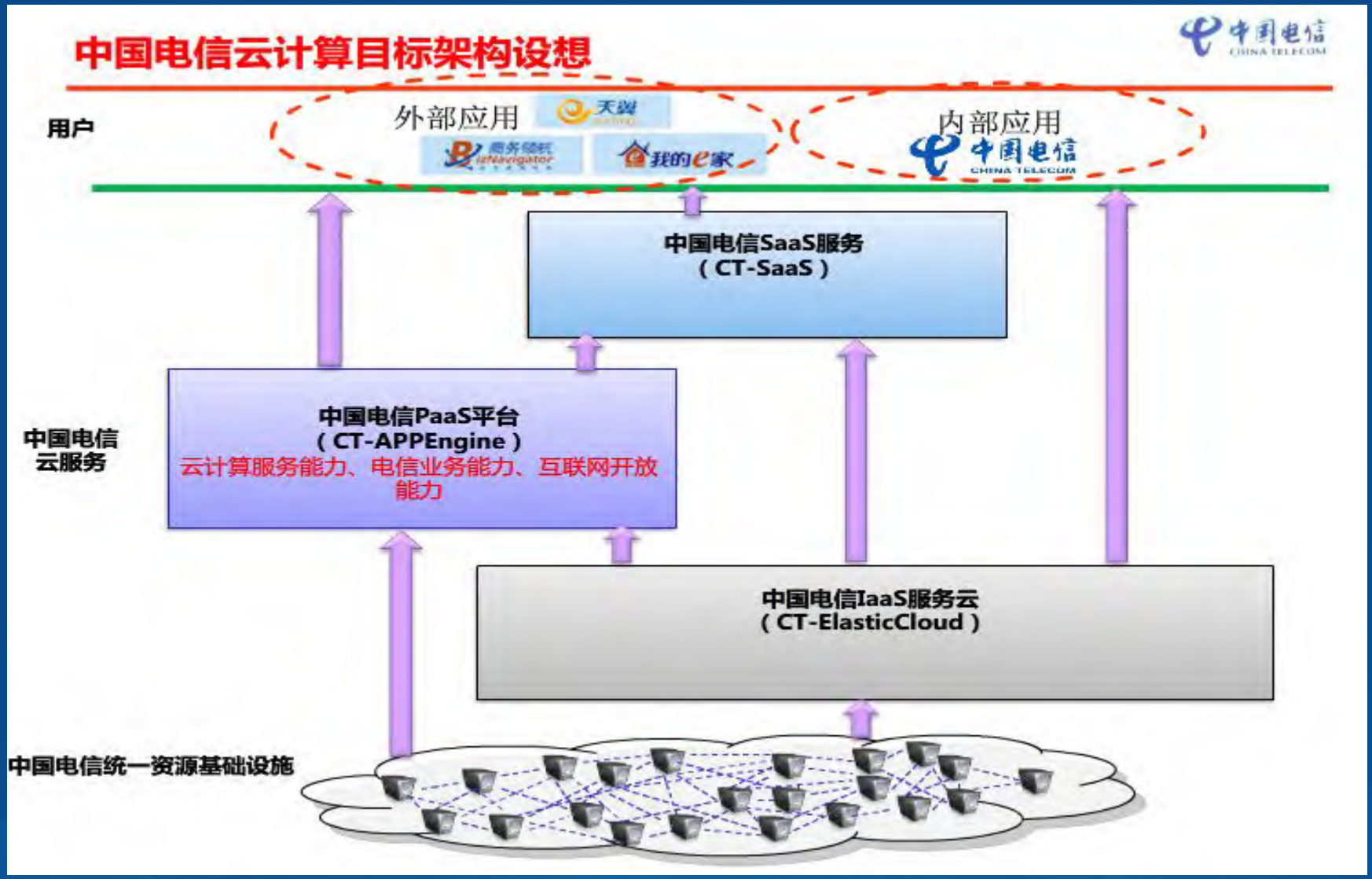
8：由先前的4个资源池拓展到8个（区域性质）
2：加上2个云基地，内蒙、贵州
X：根据客户需求按需布点

目标：构建差异化的服务能力

中国电信投120亿在呼和浩特建云计算基地

- 2011年12月6日，中国电信宣布启动 哈尔滨数据中心项目。
- 该项目总投资120亿元，一期总投资超过50亿元，建40万台服务器

中国电信投120亿在呼和浩特建云计算基地


中国电信云计算目标架构设想
中国电信
CHINA TELECOM
用户
外部应用
天翼
我的e家
内部应用
中国电信
CHINA TELECOM
中国电信SaaS服务
（CT-SaaS）
中国电信
云服务
中国电信PaaS平台
（CT-APPEngine）
云计算服务能力、电信业务能力、互联网开放能力
中国电信IaaS服务云
（CT-ElasticCloud）
中国电信统一资源基础设施

中国电信发展云计算应遵循的原则

- **合作发展、争取主导运营**
  - 借助主流云计算服务提供企业，积极参与云计算标准化制定，发挥电信运营商优势，**合作共赢，共同推进产业链发展**
  - 技术合作为主，积极争取**主导云计算服务运营**
- **聚焦平台建设，增强差异化**
  - 打造IaaS，PaaS和SaaS等不同类型的服务平台，服务于不同目标客户
  - 与已有业务能力深度定制集成，增强平台服务能力和服务管理能力，提升企业信息化、移动互联网的融合业务与应用竞争力，提升业务与应用的创新能力
- **内外并举、稳步发展**
  - 内部应用于资源优化、平台整合；**外部优化现有业务运营体系**、提供新业务服务
  - 根据市场成熟度和技术成熟度，有步骤地推出不同模式的云服务

主要内容

- 云计算基础架构及应用实践
- 大数据技术与业务应用实践

大数据现象

- 信息爆炸：2006年，全球产生161 EB 的数据，印成书是地球到太阳距离的10倍；2007年，全球产生280 EB 数据，全世界平均每人45G；而人类历史5000年的文字记载只有5EB

全球数据大爆炸，大数据时代来临

- 随着移动互联网、云计算、物联网技术和业务的发展，数据呈爆炸性增长
- 麦肯锡全球研究机构认为**大数据是创新、竞争和生产力的下一个前沿领域**，数据将会给社会带来更大的价值。

- **全球进入ZB时代**
  - –2013年全球数据量达到4.4ZB
  - –到2020年，将达到44 ZB
- **“数据太多，知识太少”**
  - – 传统数据分析方式无法进行辨析和处理，只有“大数据应用”才能从**数据汇聚**到**知识生成**

*注：
1 ZB = 1024 EB
1 EB = 1024 PB
1 PB = 1024 TB
1 TB = 1024 GB

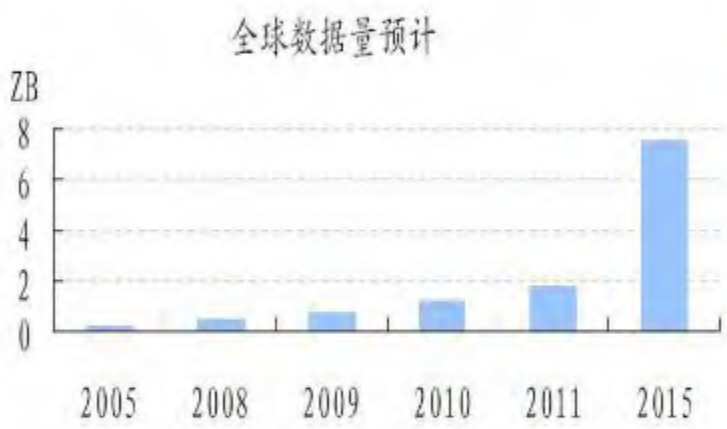
来源： IDC Digital Universe Study

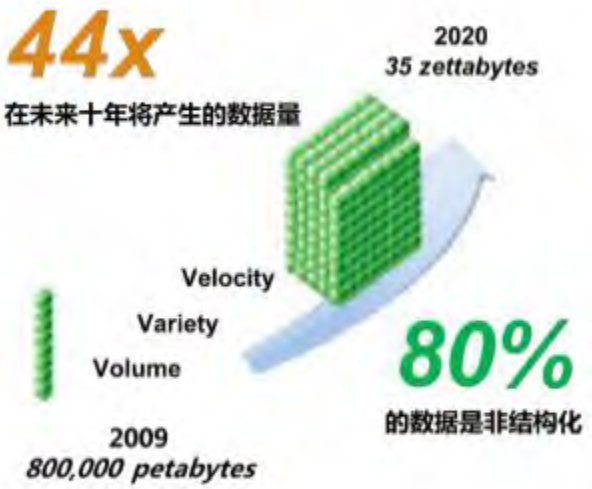
- Facebook用户每天上传**3亿**张照片，超过**500TB**的数据增长量，**100PB**单集群存储容量
- Google索引的在线数据2002年是5EB，到2009年增长到**280EB**
- 淘宝网注册用户达到**3.7亿**，在线商品数达到**9亿**，**100PB**海量数据存储

新摩尔定律

- 全球数据总量每18个月翻番。
- 大数据已经成为一种自然资源
- 大数据不被利用就是成本

未来增长不可限量





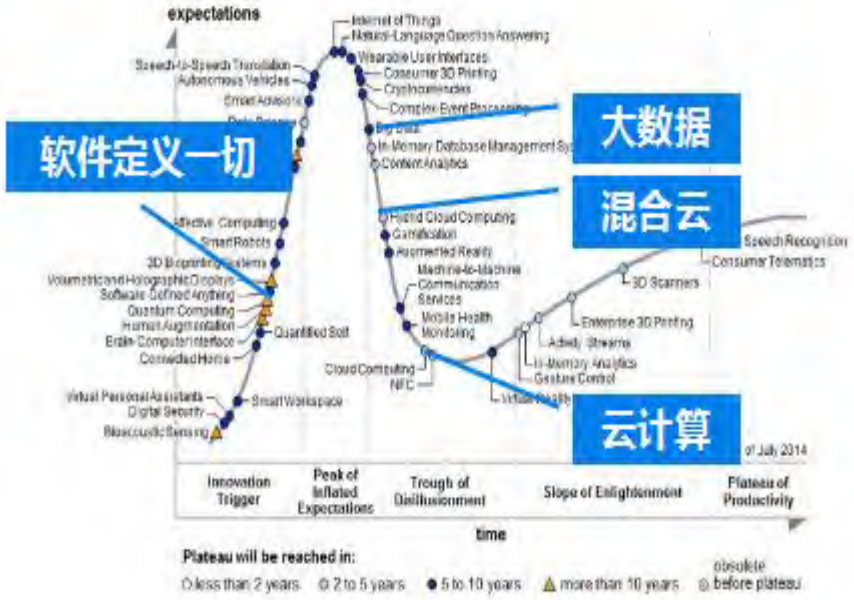
来源：IDC数字宇宙研究报告，2011.11

据IDC预测，未来10年全球数据量将以40+%的速度增长，2020年全球数据量将达到35ZB（35,000,000PB），为2009年（0.8ZB）的44倍

总体判断

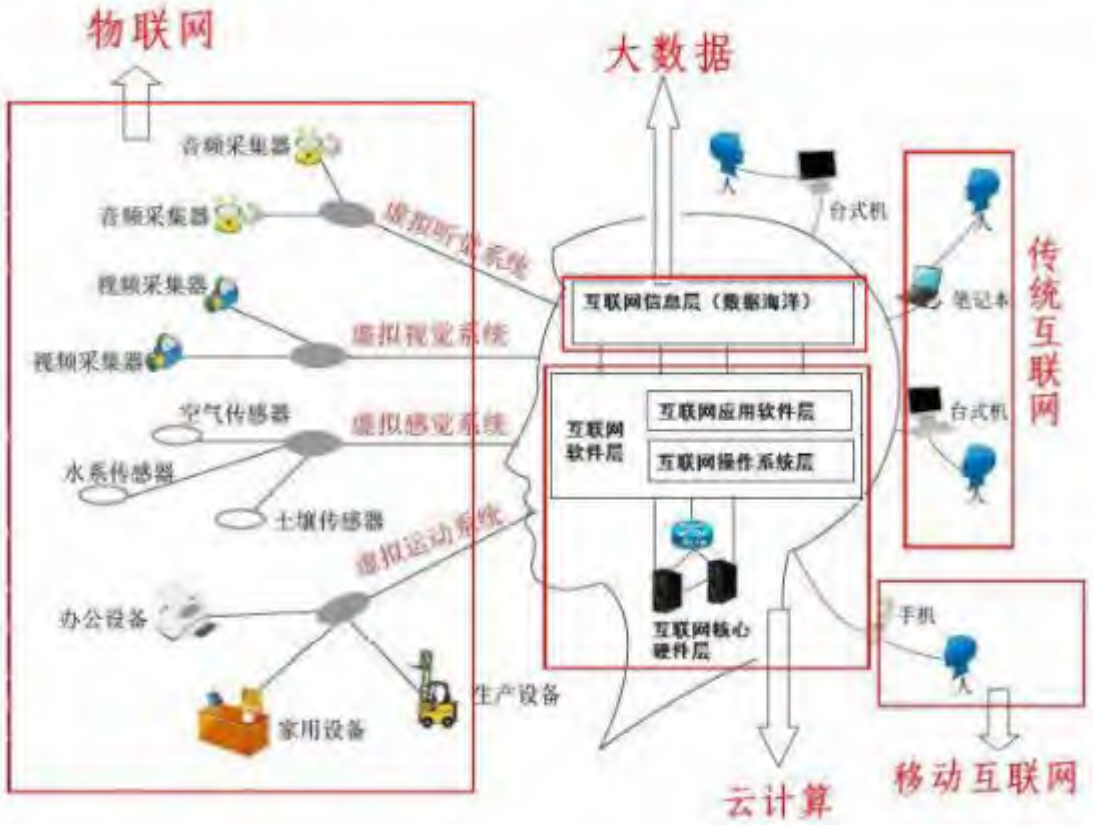
2015是大数据由概念走向具体、由技术走向应用的元年，处于产业大爆发的前夜

- 基于大数据的RTB广告和互联网金融发展迅速
- 互联网公司日益成为大数据公司，大数据营销、搜索、推荐平台相继推出
- 教育、医疗、电商等行业的大数据极具潜力



大数据和物联网、云计算的结合，将产生更大的价值

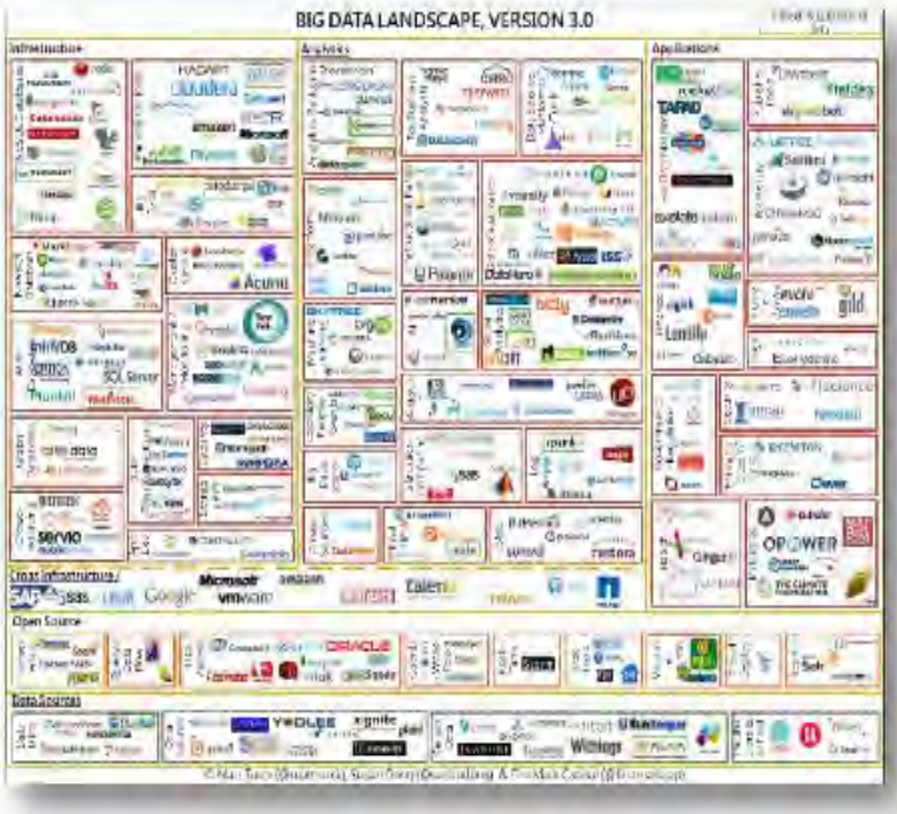
- 物联网：提供感知层——互联网的感觉和运动神经系统。
- 云计算：提供计算层——互联网的核心硬件层和核心软件层的集合。
- 大数据：提供智慧层——互联网的信息（数据海洋）
- 物联网、传统互联网、移动互联网向大数据层汇聚数据和接受数据。



大数据发展趋势

1. 大数据产业链初步形成
2. 数据开放与产业整合是大数据发展的关键
3. 数据开放与隐私安全成为大数据发展的主要障碍
4. Hadoop 2.0逐步取代1.0成为主流的大数据平台
5. 云计算和大数据技术开始融合
6. 推荐系统、可视化、知识图谱在大数据中广泛采用
7. 大数据行业应用蓬勃发展，初显成效

- 包含数据提供者、技术提供者、服务提供者、数据使用者的完整产业链已经形成，为行业用户提供数据、基础设施、分析技术和数据服务
- 大型互联网公司、电信运营商和政府是主要数据来源
- 传统IT巨头和新型创业公司瓜分大数据基础设施市场
- 创业公司占据大数据行业应用的主要位置
- 数据开放与隐私安全成为大数据发展的主要障碍



大数据产业视图

产业链上游 产业链下游

数据	基础设施	分析工具	行业应用
1. 大型互联网公司和电信运营商是主要数据源 Google verizon 2. 政府数据公开将成趋势 DATA.GOV	1. 计算平台多基于Hadoop、Spark等开源技术 cloudera Hortonworks 2. NoSQL、MPP、NewSQL等新型数据库涌现 mongoDB SkySQL	1. 数据统计工具依旧热门 sas SPSS 2. 可视化领域前景广阔 visual ly 3. 社交分析等垂直领域分析工具涌现 simple reach	1. 广告、金融、市场是当前主战场 doubleclick 广点通 tanx 融360 RelateIQ 2. 医疗、教育、人工智能是未来重要方向

数据开放与产业整合是大数据发展的关键

- **大数据开放是业界趋势**
  - 广告联盟、广告交换等形式数据开放已经形成产业，国内成立地区性、行业性数据产业联盟；
  - 国内外有一些公司开始汇集各种数据源，为客户提供数据产品服务
  - 政府数据开发逐渐成为现实

各种数据资源在有效组合和关联后会产生巨大价值，各数据持有方进行数据开放是前提；大数据经过产业整合，各个角色找到定位，有助于整个产业的蓬勃发展

大数据遗失的“V”特征

Volume
（体量巨大）

数据量将增长几百倍
巨量数据存储技术

大部分是非结构化数据
非结构化数据处理技术

Variety
（类型繁多）

大数据

Velocity
(实时处理)

通常要求在几秒响应
实时数据处理技术

数据价值密度低
新型数据挖掘技术

Value
（价值密度低）

大数据的最重要的“V”往往被忽视——最关键的“V”是 Value！

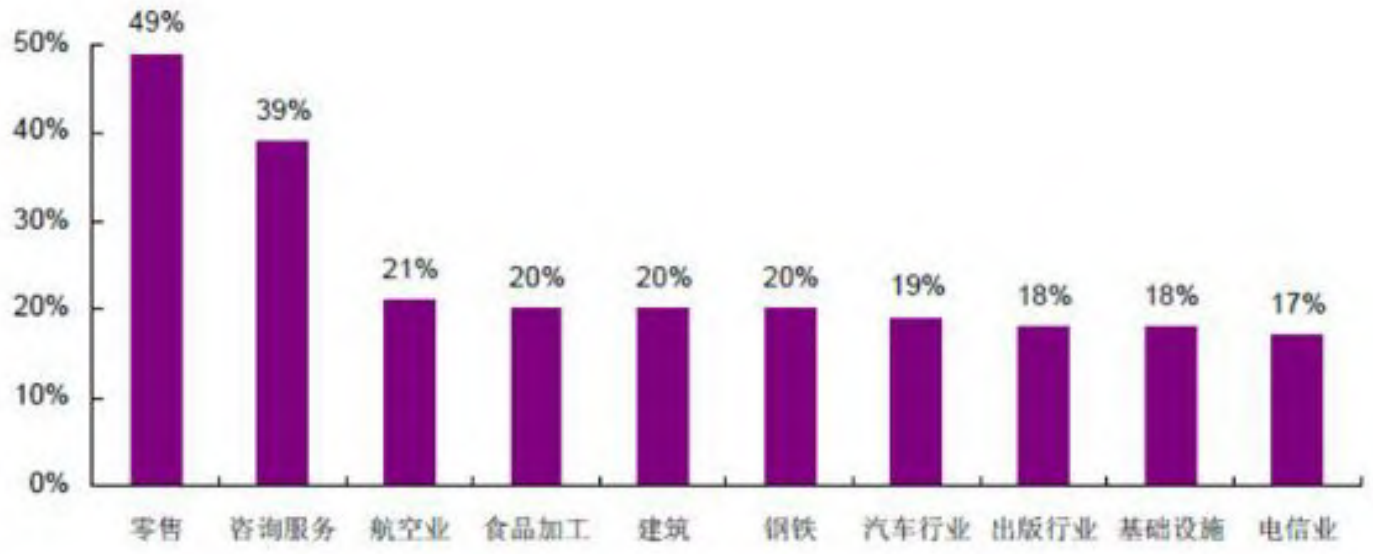
数据采集 → 数据存储 → 数据管理 → 数据挖掘 → 价值

数据利用比例直降

IBM实体分析首席科学家Jeff Jonas

- 计算速度越来越快，企业却越来越笨。
- 今天很多企业能弄懂7%的企业数据，但这个数字很快会下降到4%，然后继续螺旋式下降。

数据使用率提升10%的影响



资料来源：《Measuring the Business Impacts of Effective Data》

大数据成为全球新的经济增长点

- 预测2020年，大数据应用**市场规模将达到近2600 亿美元**

各国政府高度重视

美国：奥巴马政府3.29宣布 **“Big Data Research and Development Initiative”**

- 将投入超过2亿美元推动大数据提取、存储、分析、共享、可视化等领域的研究，并将其与历史上对**超级计算和互联网**的投资相提并论

中国：工信部物联网十二五规划

- 信息处理技术作为4 项关键技术创新工程之一被提出
- 包括海量数据存储、数据挖掘等

涉及用户隐私事件频出，对隐私日益重视

- 如何保障数据安全？
- 如何保护用户隐私？

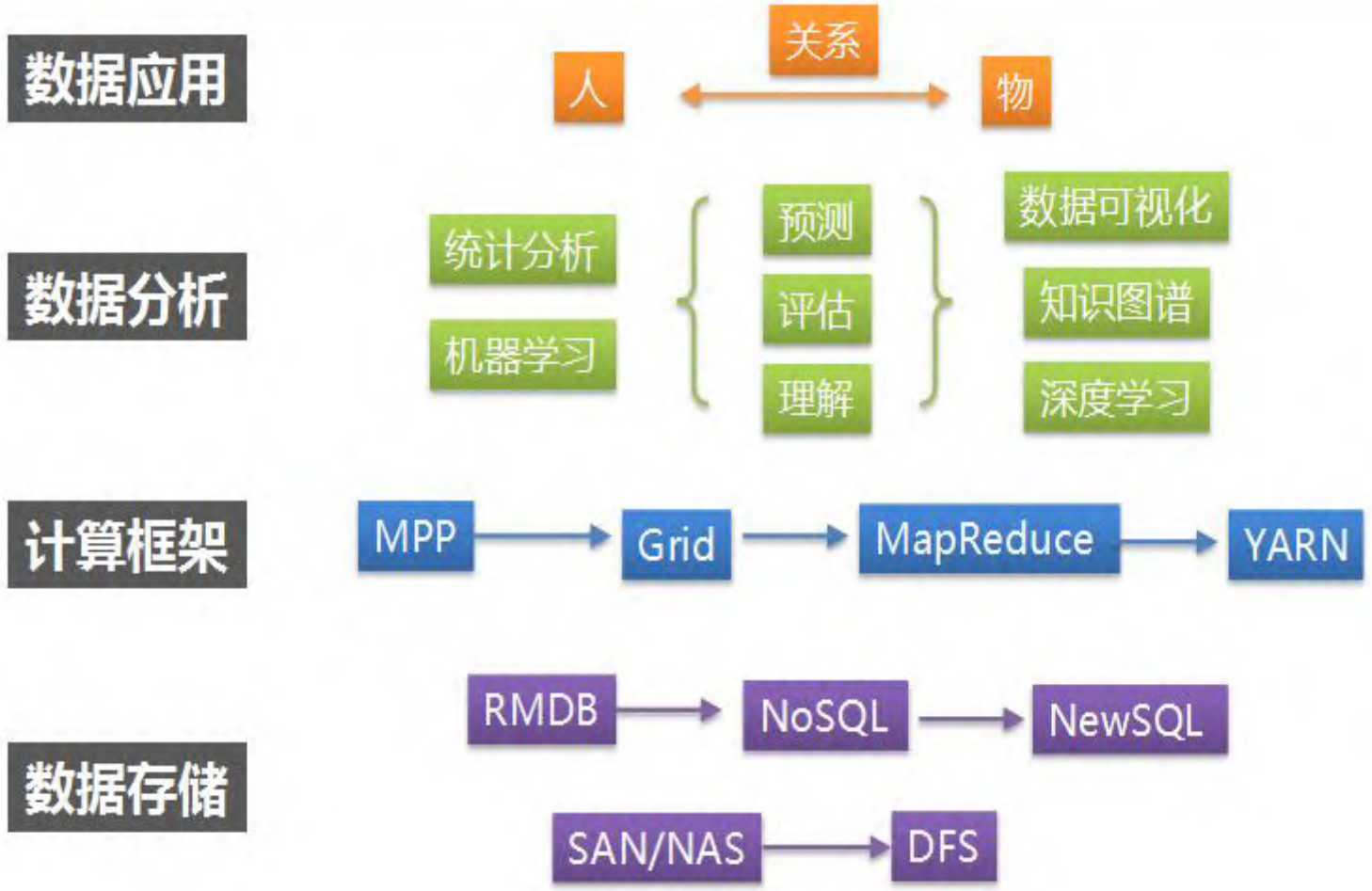
棱镜门

数据开放与隐私安全成为大数据发展的主要障碍

各国高度重视用户隐私保护和大数据安全，但目前缺少明确的法律规定，处于模糊地带

- **白宫大数据白皮书主要观点**
  - “去识别化”并不总是有效
  - “完美的个性化”也会造成微妙的或是不明显的歧视
  - “小”数据造成更大的隐私威胁
  - 改进消费者隐私权力法案
  - 通过关于国家数据外泄的立法
  - 修正电子通信隐私法
- **欧盟有关数据安全的法规**
  - 1995年10月24日，欧洲议会和欧盟制定通过《关于涉及个人数据处理的个人保护以及此类数据自由流动的指令》
  - 1997年9月15日，欧盟委员会通过了《电信部门个人数据处理和隐私保护指令》
  - 1999年，欧盟通过了《关于在信息高速公路上收集和传递个人数据的保护指令》

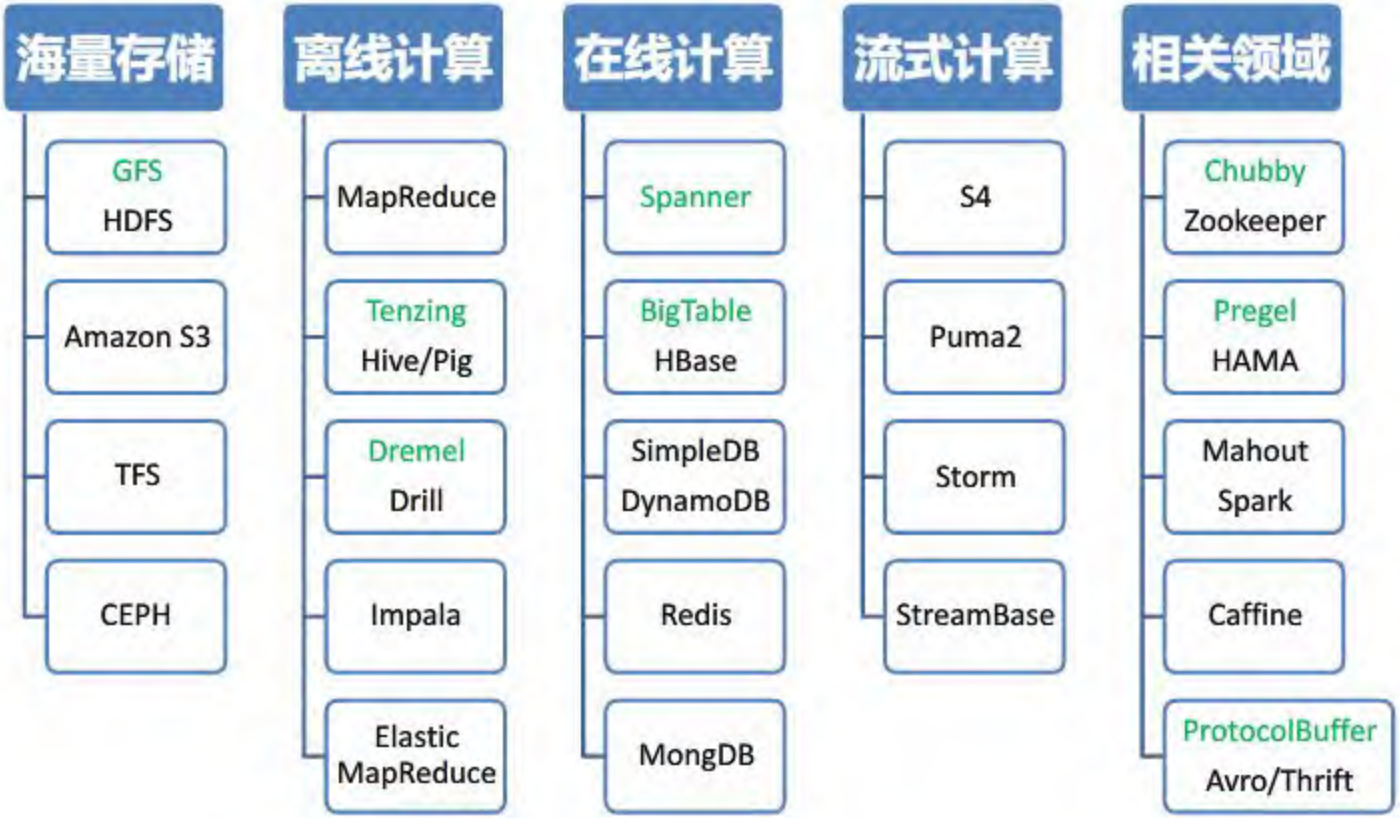
大数据技术体系


数据应用
人
关系
物
数据分析
统计分析
机器学习
预测
评估
理解
数据可视化
知识图谱
深度学习
计算框架
MPP
Grid
MapReduce
YARN
数据存储
RMDB
NoSQL
NewSQL
SAN/NAS
DFS

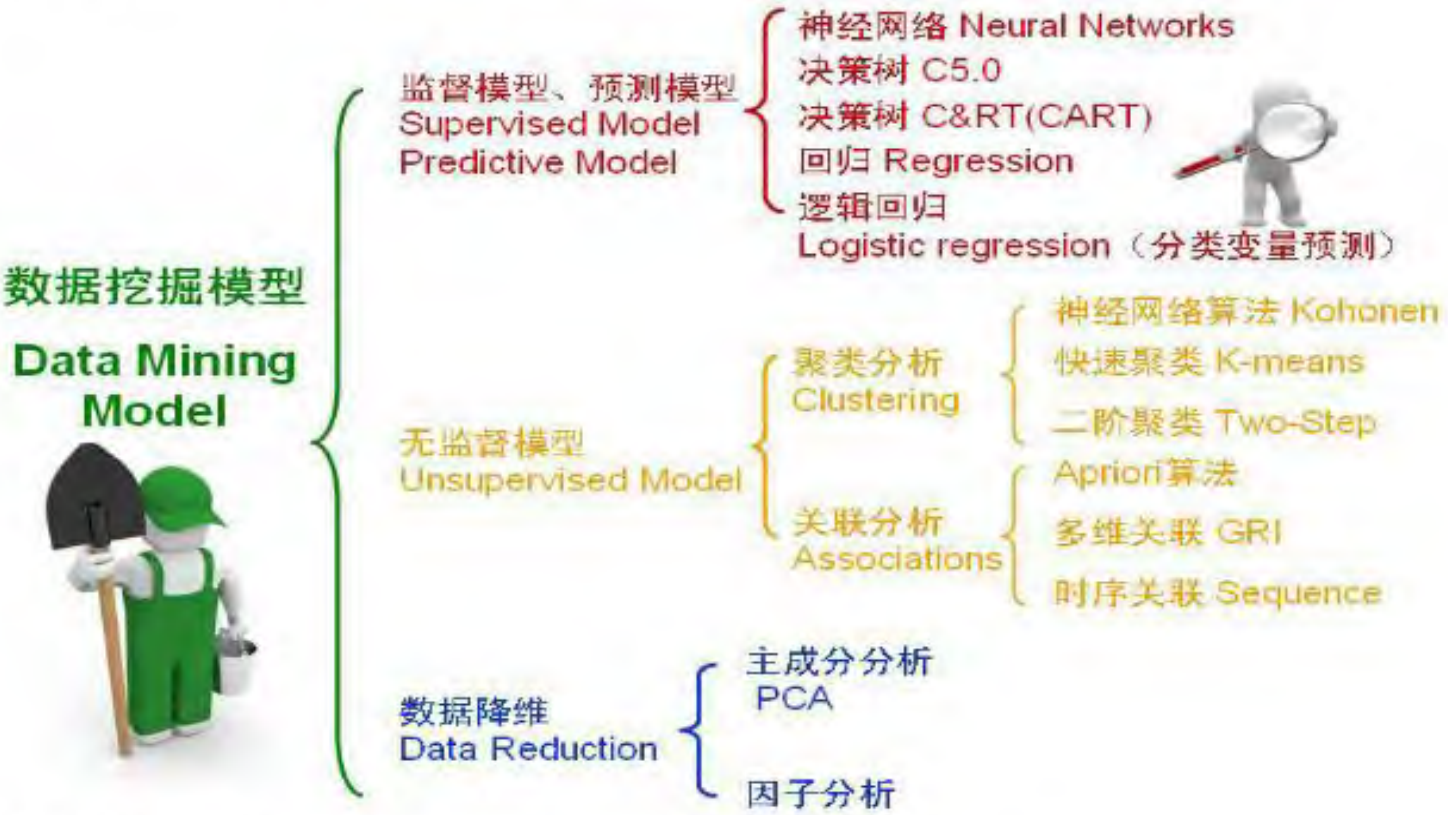
大数据涉及的关键技术

	需求	关键技术	技术描述
大数据采集处理	海量数据分布式处理	Hadoop 生态系统	针对大量数据进行分布式处理的系统框架
	非结构化数据处理	文本处理技术：自然语言理解； 多媒体处理技术...	文本内容分词与分析；图像、音视频分析
	实时数据处理	Streaming Data	流计算引擎
大数据分析	可视化交互界面	交互式可视化探索分析技术	通过交互式可视化界面辅助用户进行分析
	智能数据分析	大规模机器学习技术	计算机模拟人类学习行为，包括特征提取、图形生成等
存储、组织、管理	数据隐私保护	数据隐私防范保护措施与数据安全技术	保护隐私数据与信息个体的对应关系等安全技术
	高效存储和管理大规模数据	数据云备份技术、数据放置和调度技术、数据溯源	存储、放置、调度大规模的数据

大数据的技术领域-分布式领域



大数据的技术领域-数据分析与挖掘



相关工具:

SAS，R，Rhive/Rhadoop, Mahout，Xlib，OpenMPI...

云计算与大数据技术开始融合

大数据与计算融合成为业界发展趋势

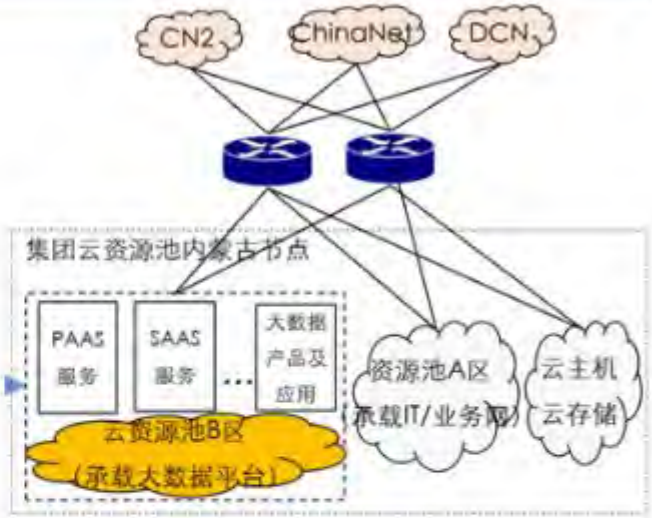
- 平台化成为大数据发展趋势
- 云计算有效降低大数据平台的管理复杂度
- 大数据有助于提升云计算平台运行的高效性
- Google基于大数据云平台推出DataFlow服务

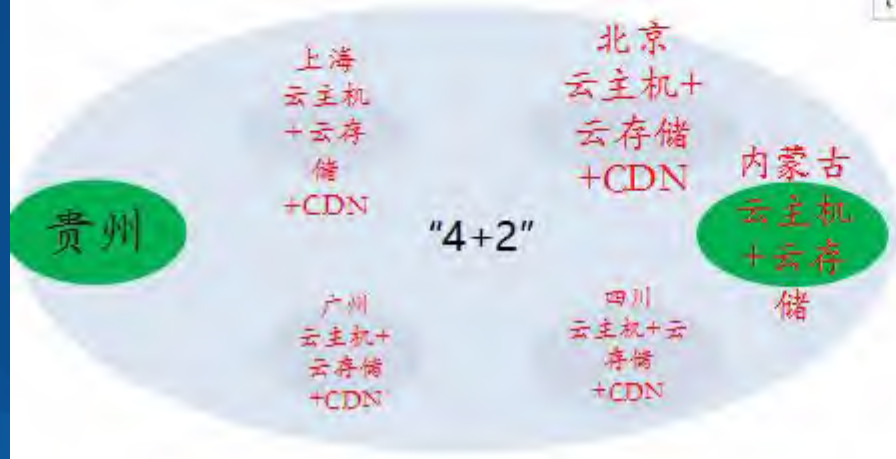
- 关键词
  - 聚合
  - 开放
  - 跨界
  - 创新

中国电信大数据

云数据中心资源池整体布局

- **云数据中心（4+2）**：按照“4+2”整体布局，已完成“4+1”部署（包括上海、广东、四川、北京云数据中心和内蒙古低成本云数据中心），另一个低成本云数据中心（贵州）正在建设中，预计年底建成





- **云资源池大数据区：**在云数据中心总体布局下，根据大数据业务的特点，今年将在内蒙古基于原有的大数据能力平台硬件建设云数据中心资源池的大数据区域，用于承载特定的大数据业务系统

优势1：超大规模数据资源

开放平台 · 中国电信数据优势

中国电信 CHINA TELECOM

独特的数据价值

- 数据广度
- 数据跨屏
- 数据中立

海量的数据基数

- 移动用户：2亿
- 宽带用户：1.5亿
- 固话用户：1.3亿
- ITV用户：4000万
- 公共WIFI热点：102万
- 家庭WIFI热点：3000万

上千节点的数据处理能力

每日汇聚500亿条数据

全国数据中心近 377个	星级中心 80个
运营商市场份额 70%	网络能力 10TB+
存储能力 EB级	计算能力 百万 物理核

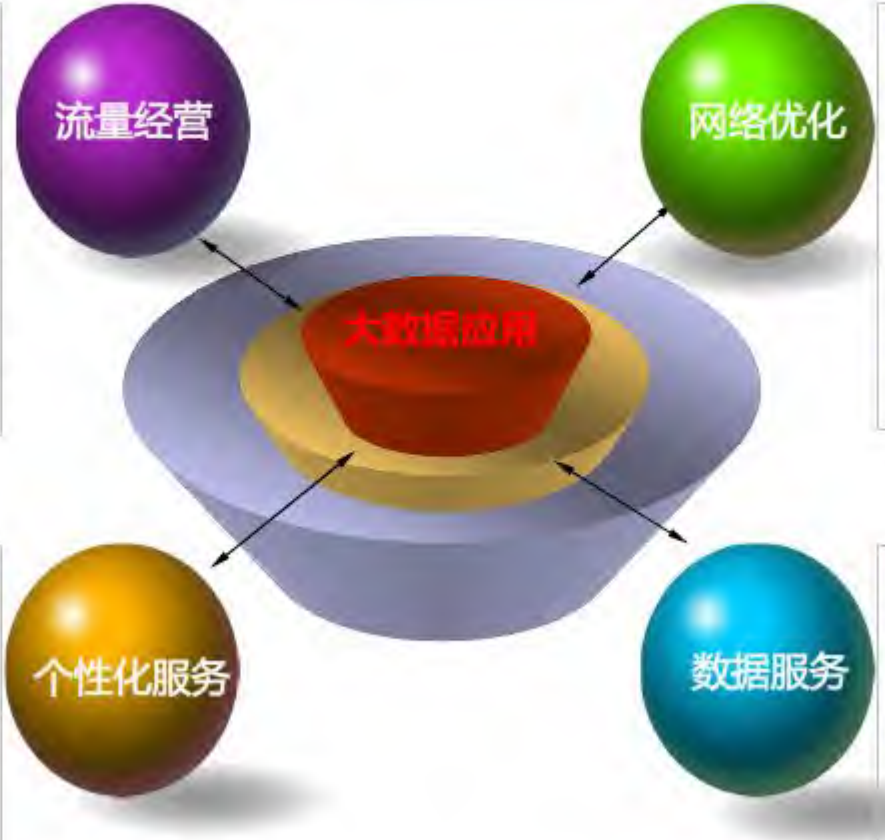
优势2：行业和客户资源优势

1、积累的政企客户数量全国超过3000万个
2、精耕重点行业：如平安城市、电子政务等
3、跨界融合、跨屏应用引领潮流：开放合作，创新生态发展新机制

电信大数据应用的主要场景

中国电信 CHINA TELECOM

- 获取并处理DPI数据，分析用户行为特征等
- 根据用户行为偏好，推送相关业务
- 按照流量价值分级经营



- 实时采集处理信令据，监控网络状况
- 实现网络、应用和用户的智能指配
- 指导网络规划

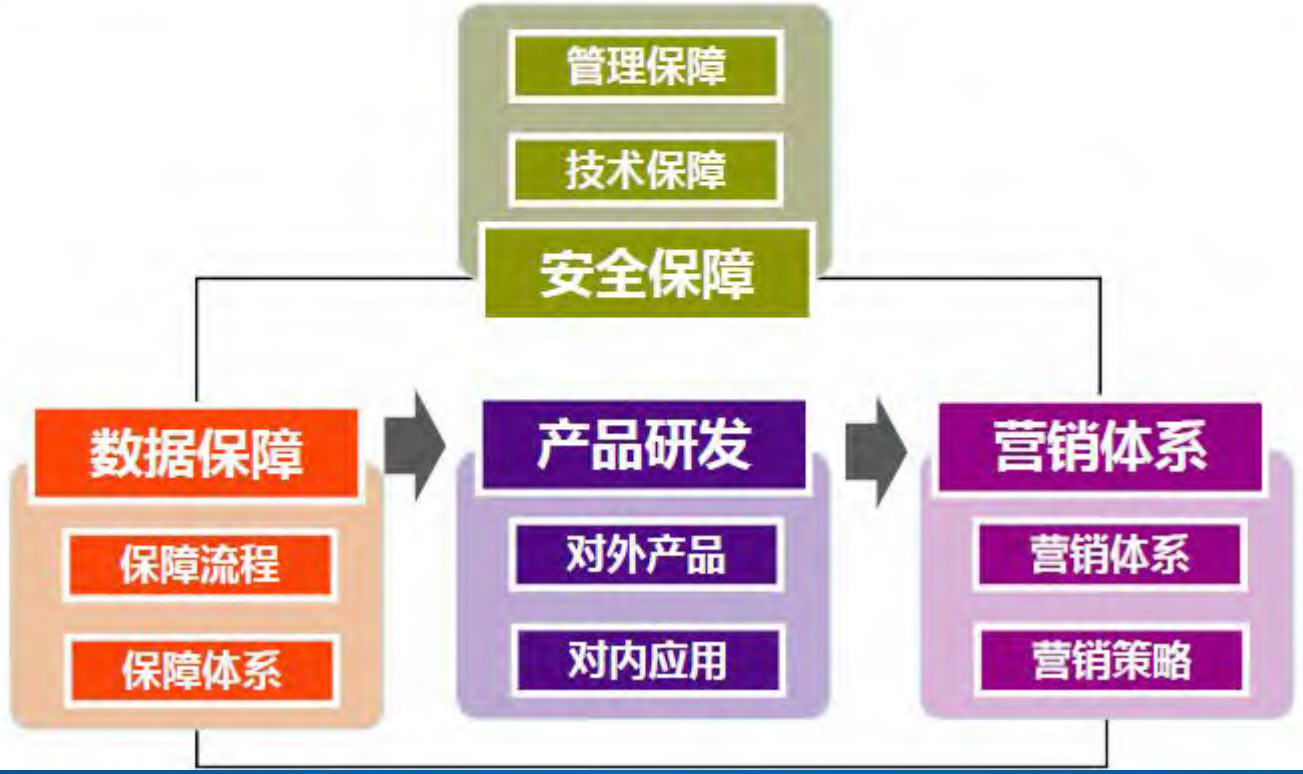
- 优化产品、套餐等，提供个性化定制能力
- 根据用户等级提供差异化服务

- 将数据封装成服务，提供给企业所有用户
- 提供数据分析开放能力

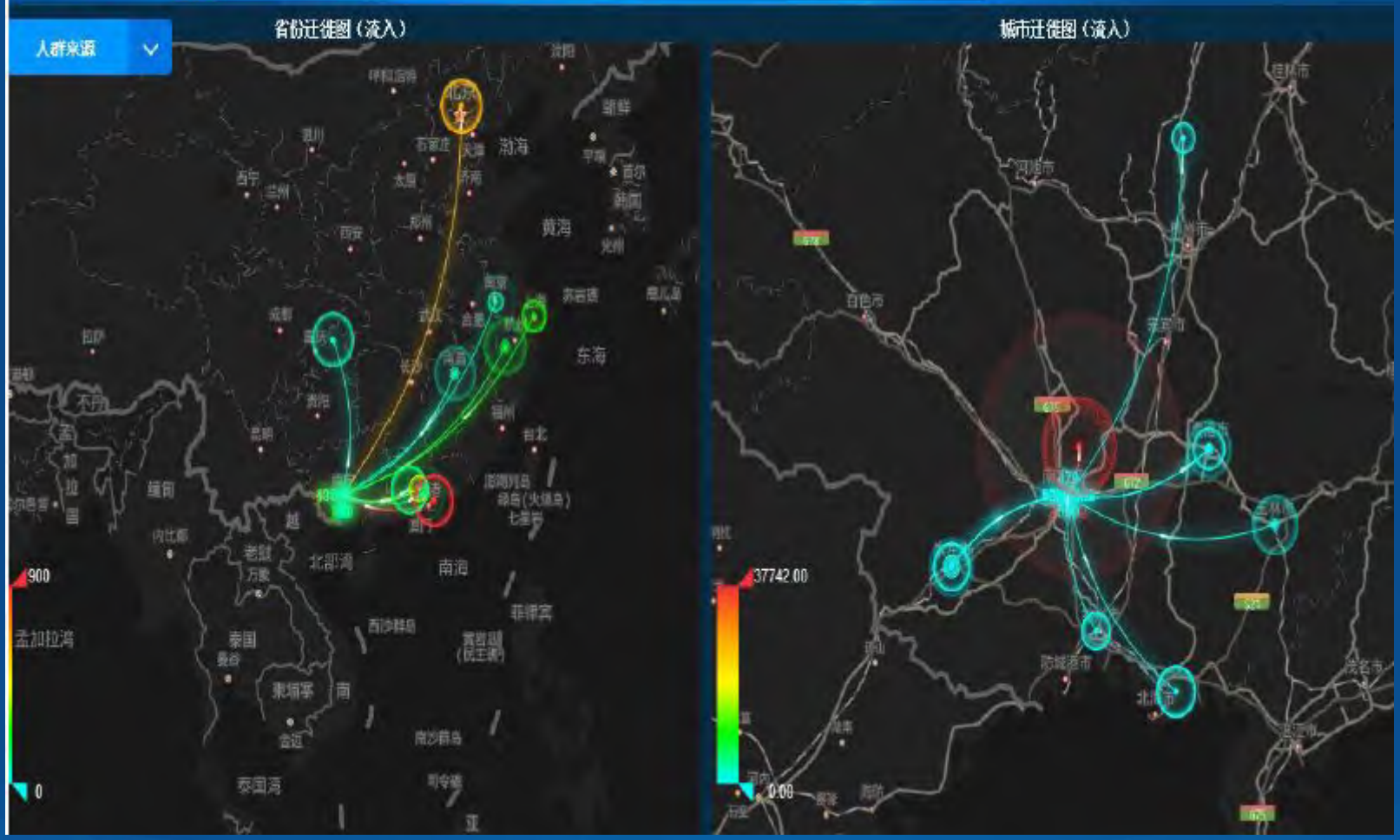
中国电信大数据运营体系

中国电信 CHINA TELECOM 世界触手可及

总体纲领：构建完整的中国电信大数据运营体系，在**安全**基础上，贯穿**数据**、**产品**、**营销**全过程，夯实持续运营能力基础。

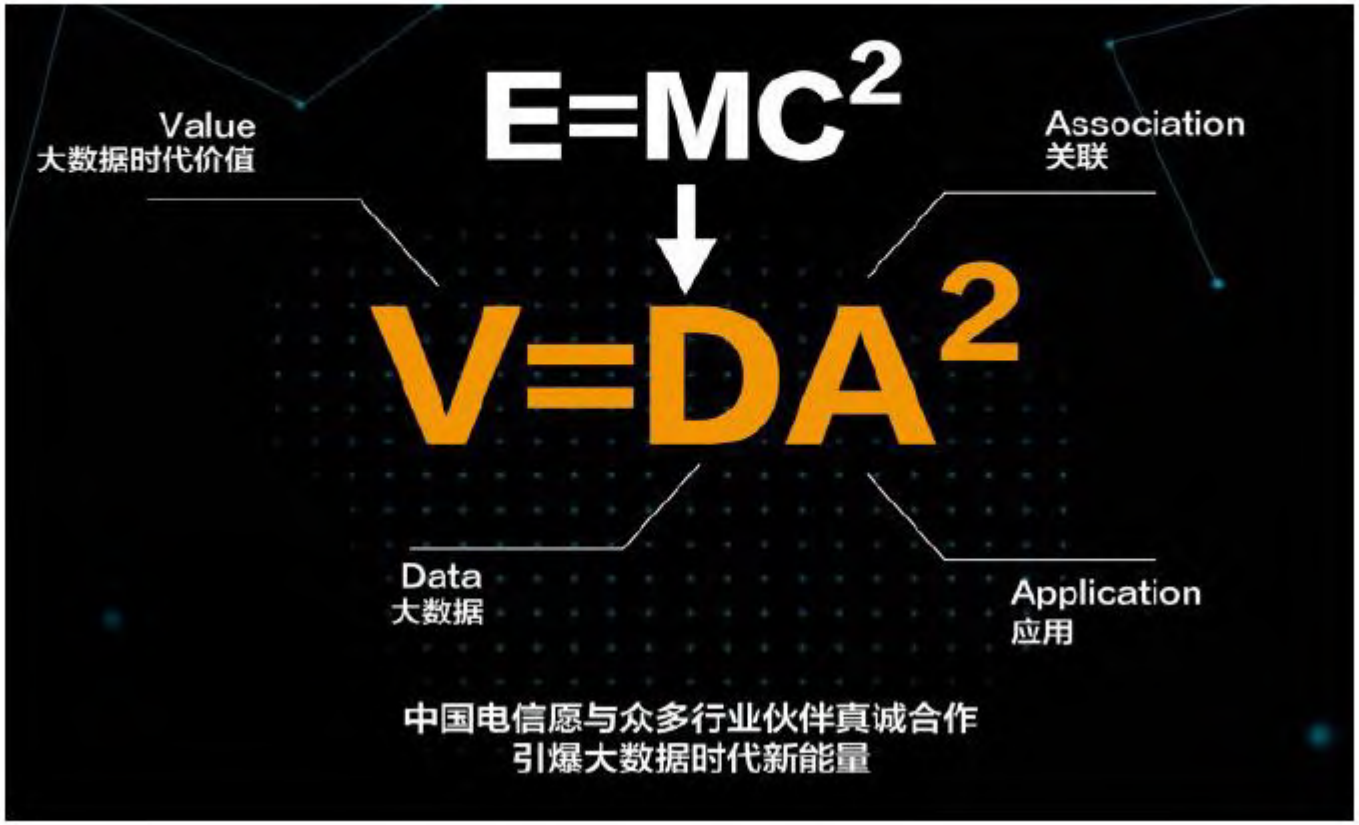


大数据应用平台——人流量热图监控应用



大数据生态：合作共赢

中国电信 CHINA TELECOM 世界触手可及



我们的结论

大数据与云计算是相辅相成的，云计算为大数据提供了有力的工具和途径，大数据为云计算提供了很有价值的用武之地。

而大数据相关的云计算，除了具备大存储、大内存以及高IO性能的要求，对比于一般意义上的云计算还需具备以下特殊性：

- **存储与处理一体化**
- **本身资源的复用度低**
- **因为存储核心数据，需要做适度安全隔离**

The 8th China Cloud Computing Conference

Thank you