Hadoop大数据应用分析

1.大数据背景介绍

2. HADOOP体系架构

3. 基于HADOOP的大数据产品分析

4. 基于HADOOP的大数据行业应用分析

5. 基于HADOOP的大数据应用建议

大数据定义及特点

IDC定义：为了更为经济的从高频率获取的、大容量的、不同结构和类型的数据中获取价值，而设计的新一代架构和技术。

- 巨大的数据量 Volume
  - 集中储存/集中计算已经无法处理巨大的数据量

新浪微博

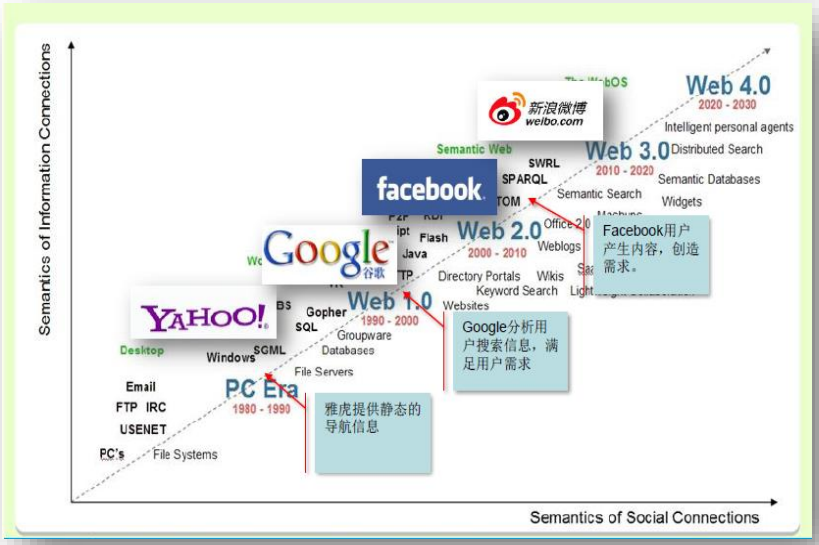
3亿+用户，高峰期一天上亿条微博

中型城市每月数十亿智能电表数据

2015年全球移动终端产生的数据量6300PB

- 多结构化数据 Variety
  - 文本/图片/视频/文档等
- 增长速度很快 Velocity
  - 海量数据的及时有效分析
  - 用户基数庞大/设备数量众多/实时海量/数据指数级别增长
- 价值密度低Value
  - 单条数据并无太多价值，但庞大的数据量蕴含巨大财富

大数据对系统的需求



- High performance –高并发读写的需求
  高并发、实时动态获取和更新数据

- Huge Storage –海量数据的高效率存储和访问的需求
  类似SNS网站，海量用户信息的高效率实时存储和查询

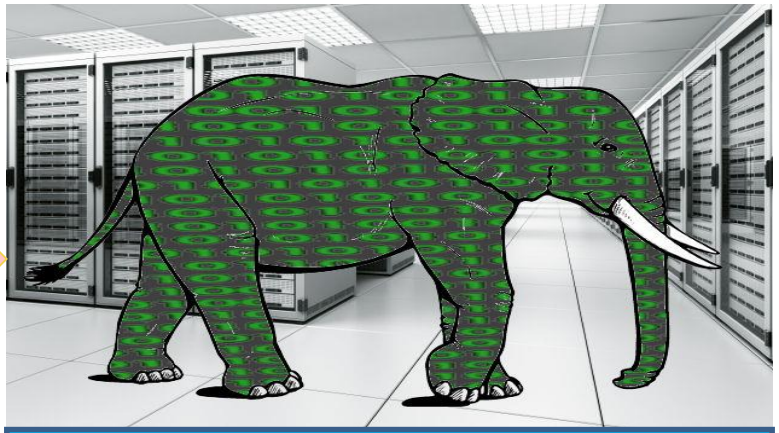
- High Scalability && High Availability –高可扩展性和高可用性的需求
  需要拥有快速横向扩展能力、提供7*24小时不间断服务

大数据和云计算的关系

云计算

商业模式驱动

大数据



应用需求驱动

- 云计算改变了IT,而大数据则改变了业务
- 云计算是大数据的IT基础，大数据须有云计算作为基础架构，才能高效运行
- 通过大数据的业务需求，为云计算的落地找到了实际应用

大数据市场分析

1 2011年-2016年中国大数据市场规模

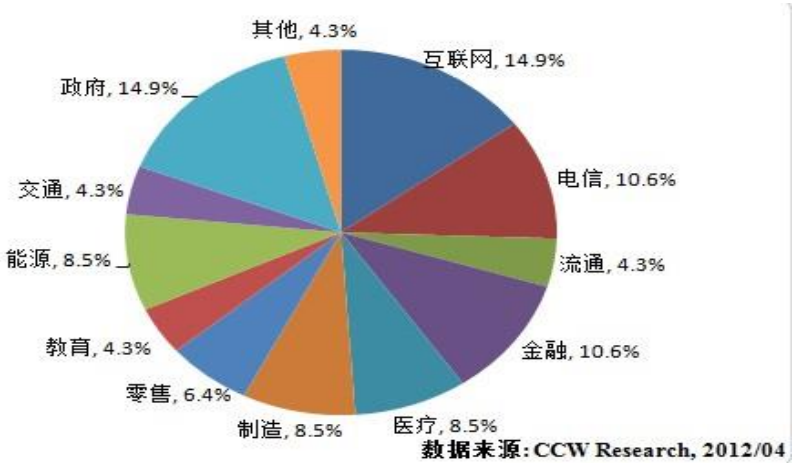
●2011年是中国大数据市场元年，一些大数据产品已经推出，部分行业也有大数据应用案例的产生。2012年-2016年，将迎来大数据市场的飞速发展。

●2012年中国大数据市场规模达到4.7亿元，2013年大数据市场将迎来增速为138.3%的飞跃，到2016年，整个市场规模逼近百亿。

2 各行业大数据市场规模

●政府、互联网、电信、金融的大数据市场规模较大，四个行业将占据一半市场份额。

●由于各个行业都存在大数据应用需求，潜在市场空间非常可观。



1.大数据背景介绍

2. HADOOP体系架构

3. 基于HADOOP的大数据厂商分析

4. 基于HADOOP的大数据行业应用分析

5.基于HADOOP的大数据应用建议

大数据主要应用技术——Hadoop

据IDC的预测，全球大数据市场2015年将达170亿美元规模，市场发展前景很大。而**Hadoop作为**新一代的架构和技术，因为有利于并行分布处理 “大数据”而备受重视。

Apache Hadoop 是一个用java语言实现的软件框架，在由大量计算机组成的集群中运行海量数据的分布式计算，它可以让应用程序支持上千个节点和PB级别的数据。 Hadoop是项目的总称，主要是由分布式存储（HDFS）、分布式计算（MapReduce）等组成。

优点：

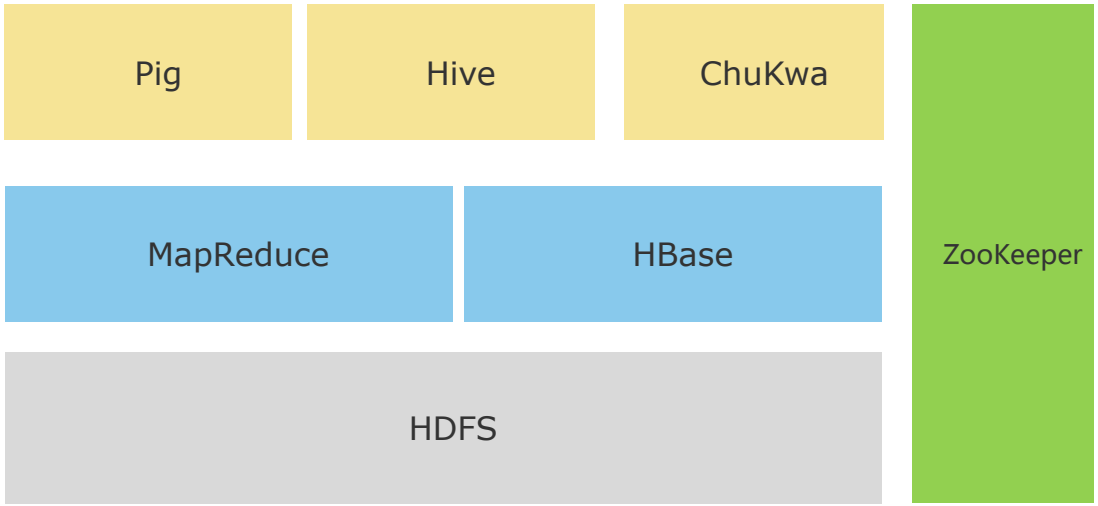
可扩展：不论是存储的可扩展还是计算的可扩展都是Hadoop的设计根本。

经济：框架可以运行在任何普通的PC上。

可靠：分布式文件系统的备份恢复机制以及MapReduce的任务监控保证了分布式处理的可靠性。

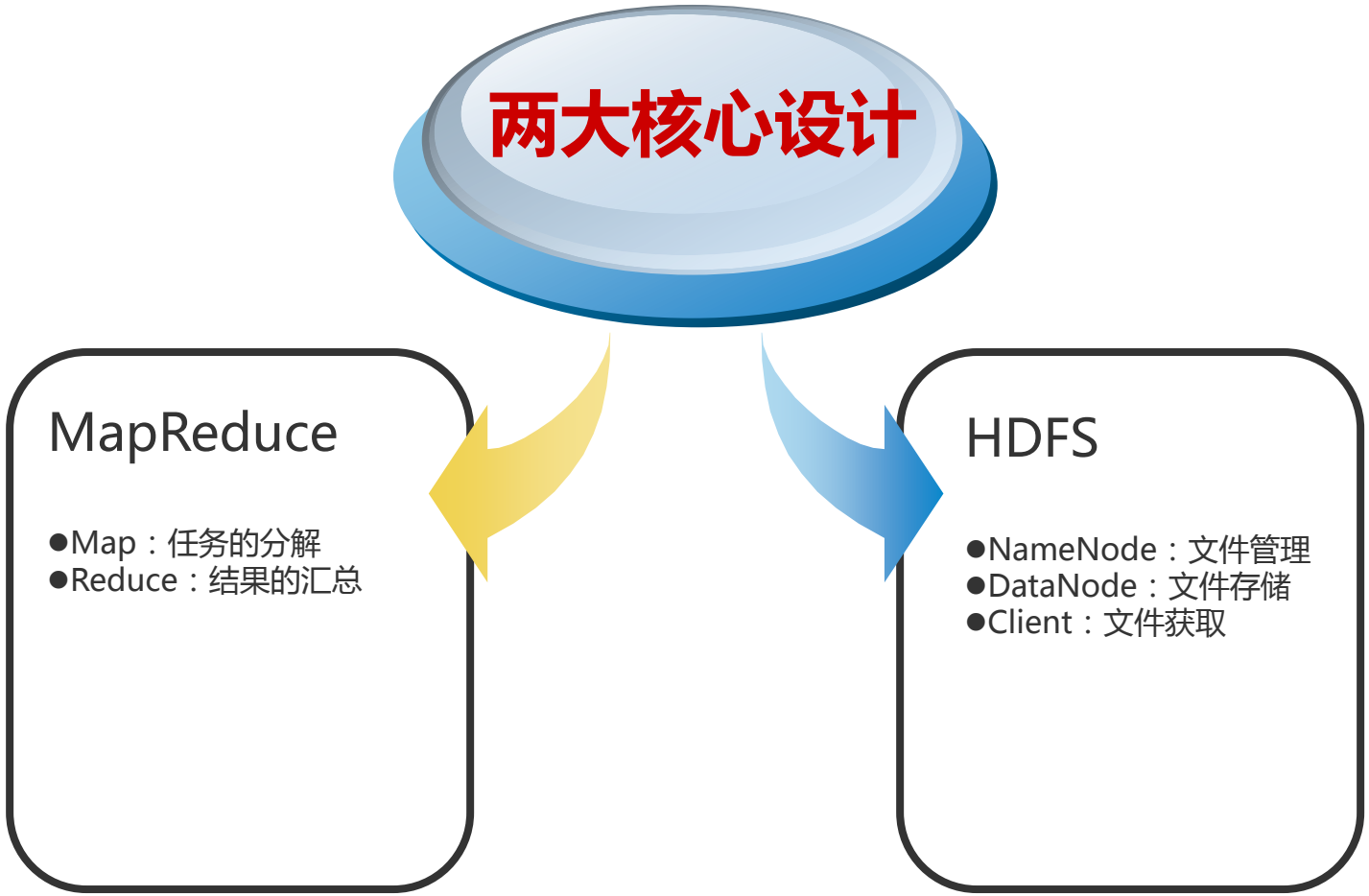
高效：分布式文件系统的高效数据交互实现以及MapReduce结合Local Data处理的模式，为高效处理海量的信息作了基础准备。

Hadoop体系架构



- Pig是一个基于Hadoop的大规模数据分析平台，Pig为复杂的海量数据并行计算提供了一个简易的操作和编程接口
- Chukwa是基于Hadoop的集群监控系统，由yahoo贡献
- hive是基于Hadoop的一个工具，提供完整的sql查询功能，可以将sql语句转换为MapReduce任务进行运行
- ZooKeeper：高效的，可扩展的协调系统,存储和协调关键共享状态
- HBase是一个开源的，基于列存储模型的分布式数据库
- HDFS是一个分布式文件系统。有着高容错性的特点，并且设计用来部署在低廉的硬件上，适合那些有着超大数据集的应用程序
- MapReduce是一种编程模型，用于大规模数据集（大于1TB）的并行运算

Hadoop核心设计


两大核心设计
MapReduce
●Map：任务的分解
●Reduce：结果的汇总
HDFS
●NameNode：文件管理
●DataNode：文件存储
●Client：文件获取

HDFS——分布式文件系统

HDFS是一个高度容错性的分布式文件系统，能提供高吞吐量的数据访问，非常适合大规模数据集上的应用。

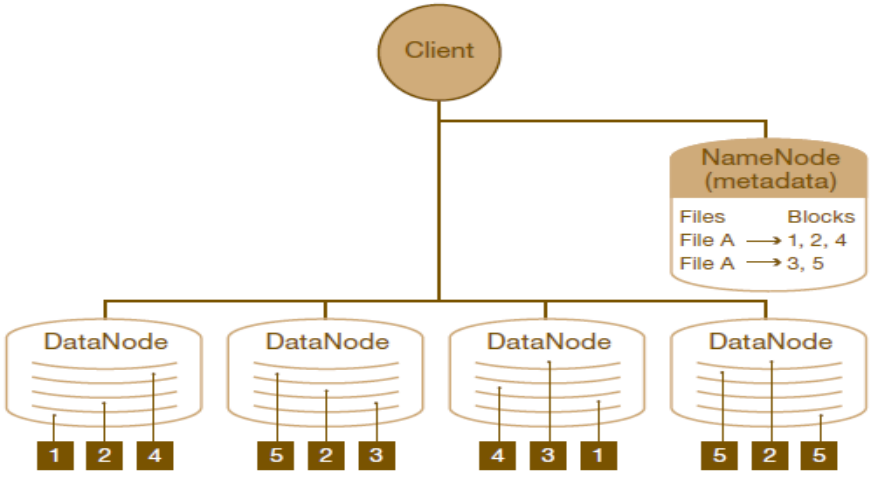
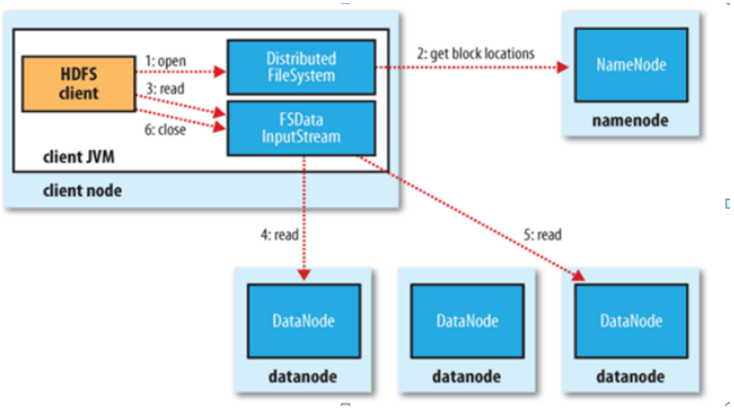


Figure 2: **The Hadoop Distributed File System, or HDFS**

Source: Apache Software Foundation, IBM, and PricewaterhouseCoopers, 2008

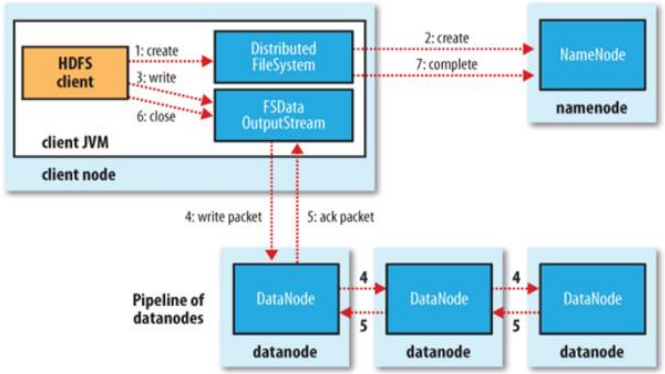
- **NameNode**
  可以看作是分布式文件系统中的管理者，存储文件系统的meta-data，主要负责管理文件系统的命名空间，集群配置信息，存储块的复制。
- **DataNode**
  是文件存储的基本单元。它存储文件块在本地文件系统中，保存了文件块的meta-data，同时周期性的发送所有存在的文件块的报告给NameNode。
- **Client**
  就是需要获取分布式文件系统文件的应用程序。

HDFS具体操作



文件写入：

1. Client向NameNode发起文件写入的请求
2. NameNode根据文件大小和文件块配置情况，返回给Client它所管理部分DataNode的信息。
3. Client将文件划分为多个文件块，根据DataNode的地址信息，按顺序写入到每一个DataNode块中。

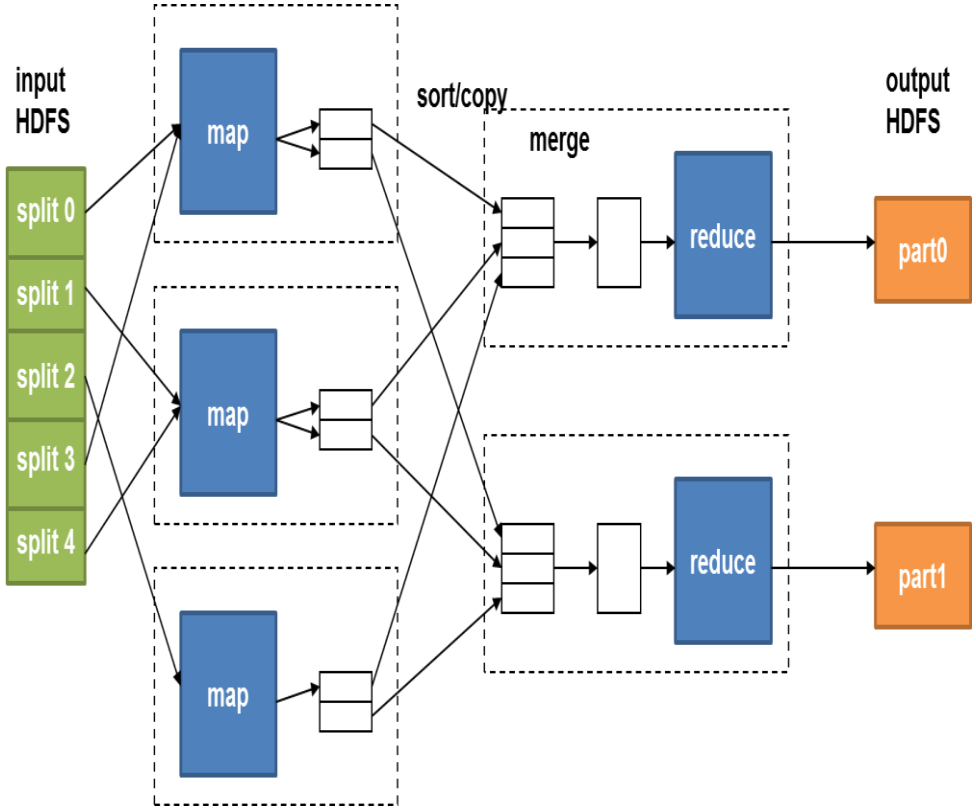


文件读取：

1. Client向NameNode发起文件读取的请求
2. NameNode返回文件存储的DataNode的信息。
3. Client读取文件信息。

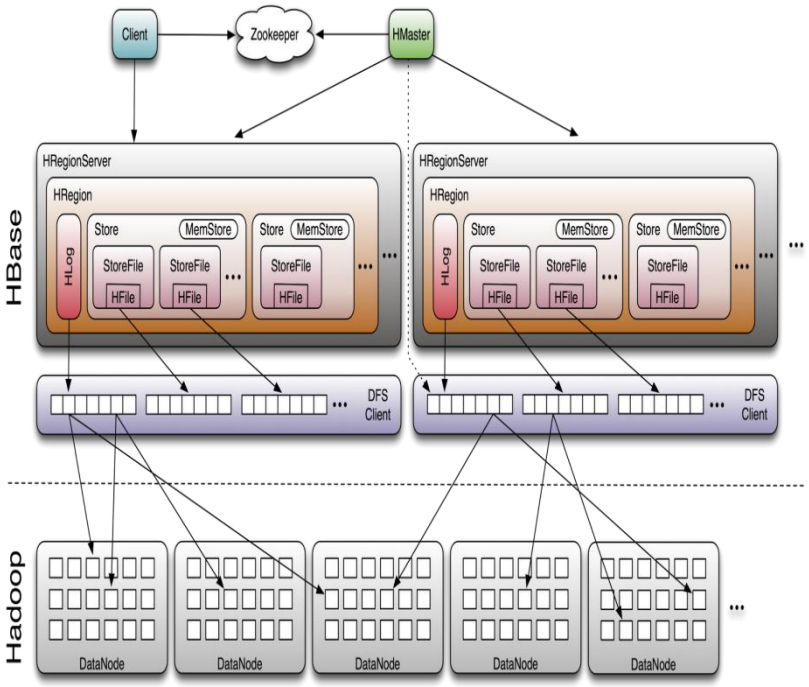
MapReduce——映射、化简编程模型

MapReduce是一种编程模型，用于大规模数据集的并行运算。Map（映射）和Reduce（化简），采用分而治之思想，先把任务分发到集群多个节点上，并行计算，然后再把计算结果合并，从而得到最终计算结果。多节点计算，所涉及的任务调度、负载均衡、容错处理等，都由MapReduce框架完成，不需要编程人员关心这些内容。



1. 根据输入数据的大小和参数的设置把数据分成splits，
   每个split对于一个map线程。
2. Split中的数据作为Map的输入， Map的输出一定在Map端。
3. Map的输出到Reduce的输入的过程(shuffle过程)：
   第一阶段：在map端完成内存->排序->写入磁盘->复制
   第二阶段：在reduce端完成映射到reduce端分区->合并->排序
4. Reduce的输入到Reduce的输出
   最后排好序的key/value作为Reduce的输入，输出不一定是在reduce端。

HBASE——分布式数据存储



HBase – Hadoop Database，是一个高可靠性、高性能、面向列、可伸缩的分布式存储系统；

HBase位于结构化存储层，HDFS为HBase提供了高可靠性的底层存储支持,MapReduce为HBase提供了高性能的计算能力，Zookeeper为HBase提供了稳定服务和failover机制；

Pig和Hive还为HBase提供了高层语言支持，使得在HBase上进行数据统计处理变的简单。

1.大数据背景介绍

2. HADOOP体系架构

3. 基于HADOOP的大数据产品分析

4. 基于HADOOP的大数据行业应用分析

5.基于HADOOP的大数据应用建议

Hadoop主要开发厂商

大型企业和机构在寻求解决棘手的大数据问题时，往往会使用开源软件基础架构Hadoop的服务。由于Hadoop深受欢迎，许多公司都推出了各自版本的Hadoop，也有一些公司则围绕Hadoop提供解决方案。Hadoop的发行版除了社区的Apache hadoop外，cloudera，IBM，ORACLE等都提供了自己的商业版本。商业版主要是提供Hadoop专业的技术支持，这对一些大型企业尤其重要。

ORACLE 甲骨文

Oracle

Cloudera

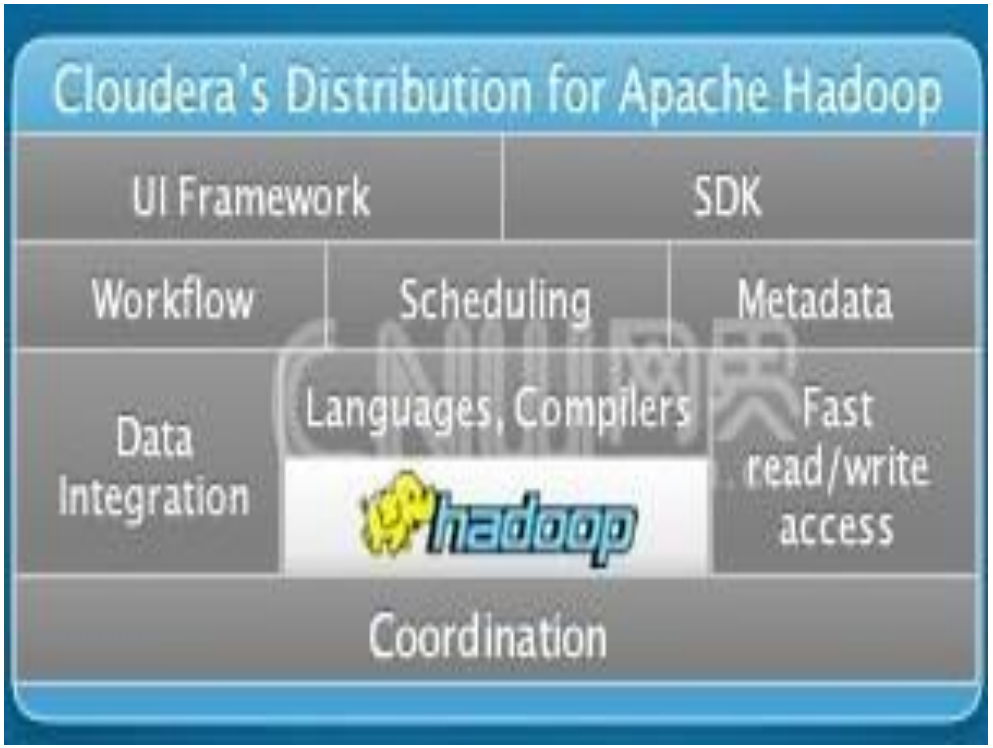
• ◦ ◦ ◦

EMC

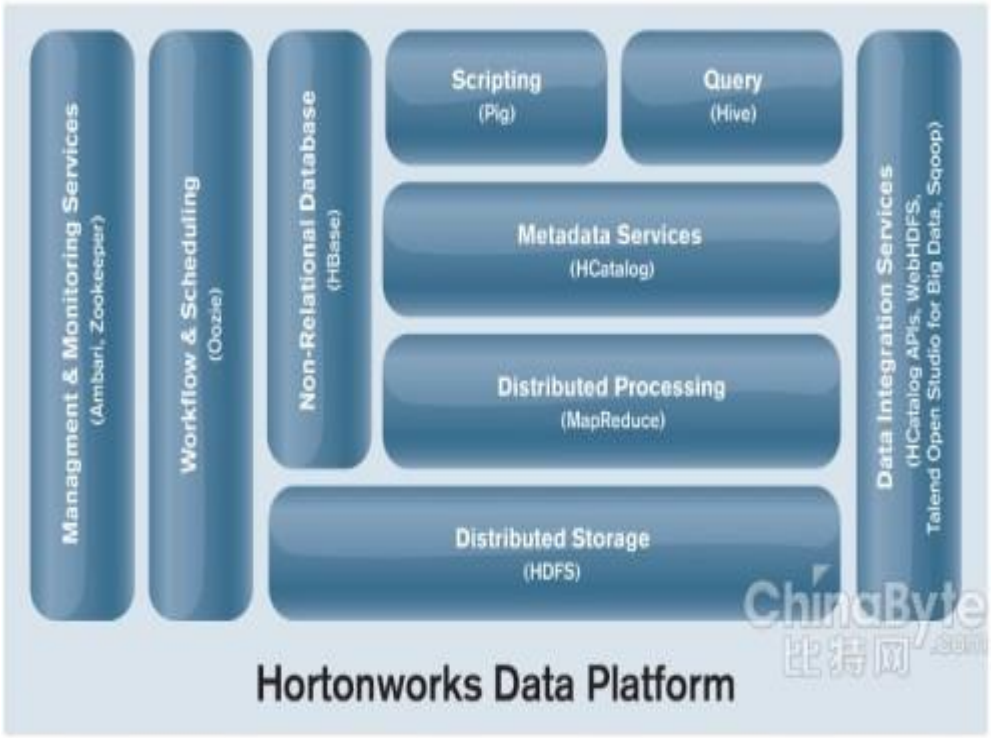
EMC²
where information lives®

Hadoop主要开发厂商—— CLOUDERA

在Hadoop生态系统中，规模最大、知名度最高的公司则是Cloudera。2008 年成立的 Cloudera 是最早将 Hadoop 商用的公司，为合作伙伴提供 Hadoop 的商用解决方案，主要是包括支持，咨询服务和培训。Cloudera的客户中倒是有很多知名公司，如AOL、哥伦比亚广播公司、eBay、Expedia、摩根大通、Monsanto、诺基亚、RIM和迪士尼等。Cloudera 企业解决方案包括Hadoop软件发行版、Cloudera管理器。

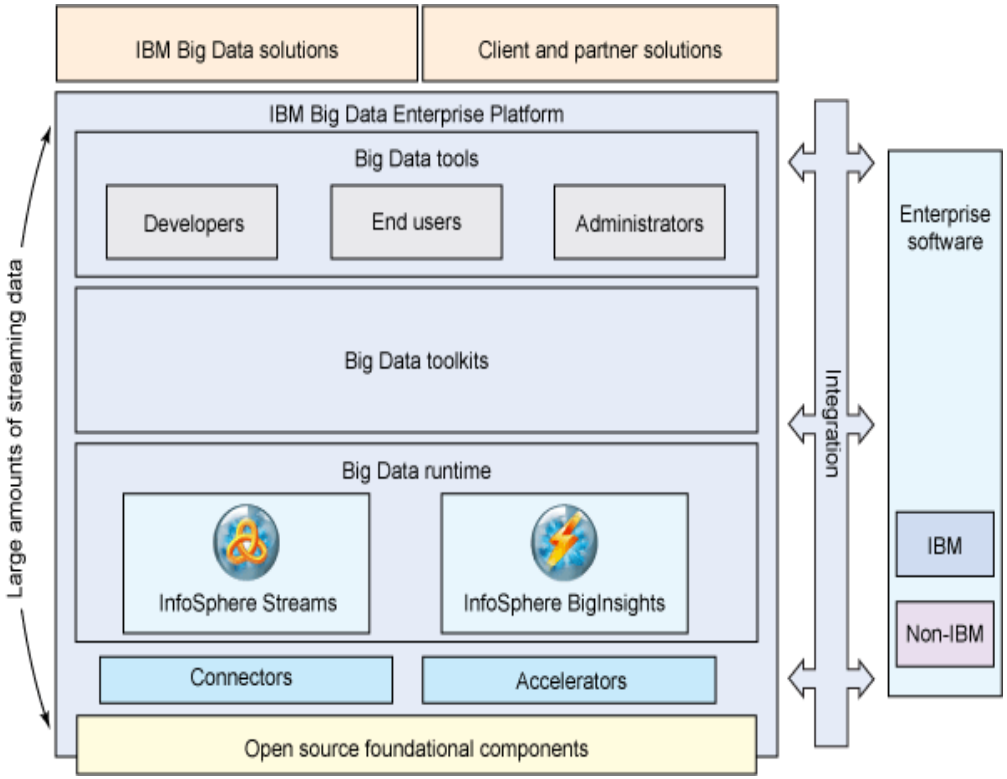


Hadoop主要开发厂商—— Hortonworks



2011年成立的Hortonworks是雅虎与硅谷风投公司Benchmark Capital合资组建的公司。公司成立之初吸纳了大约25名至30名专门研究Hadoop的雅虎工程师，上述工程师均在2005年开始协助雅虎开发Hadoop，这些工程师贡献了hadoop 80%的代码。 Hortonworks 的主打产品是Hortonworks Data Platform (HDP)，包括稳定版本的Apache Hadoop的所有关键组件。

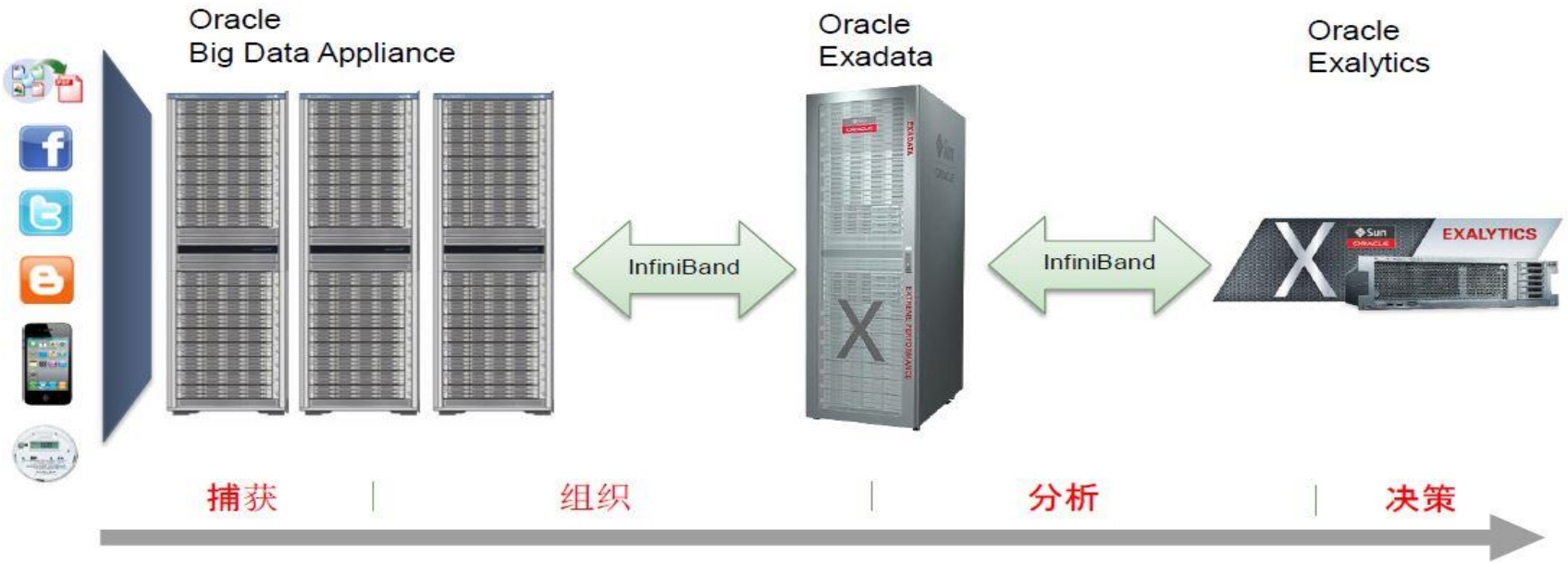
Hadoop主要开发厂商——IBM



InfoSphere BigInsights 是一个软件平台，旨在帮助企业从大量不同范围的数据中挖掘商机并进行分析，如日志记录、点击流、社会媒体数据、新闻摘要、电子传感器输出，甚至是一些事务数据等。BigInsights 包括Apache Hadoop发行版、面向MapReduce编程的Pig编程语言、针对IBM的DB2数据库的连接件以及IBM BigSheets。

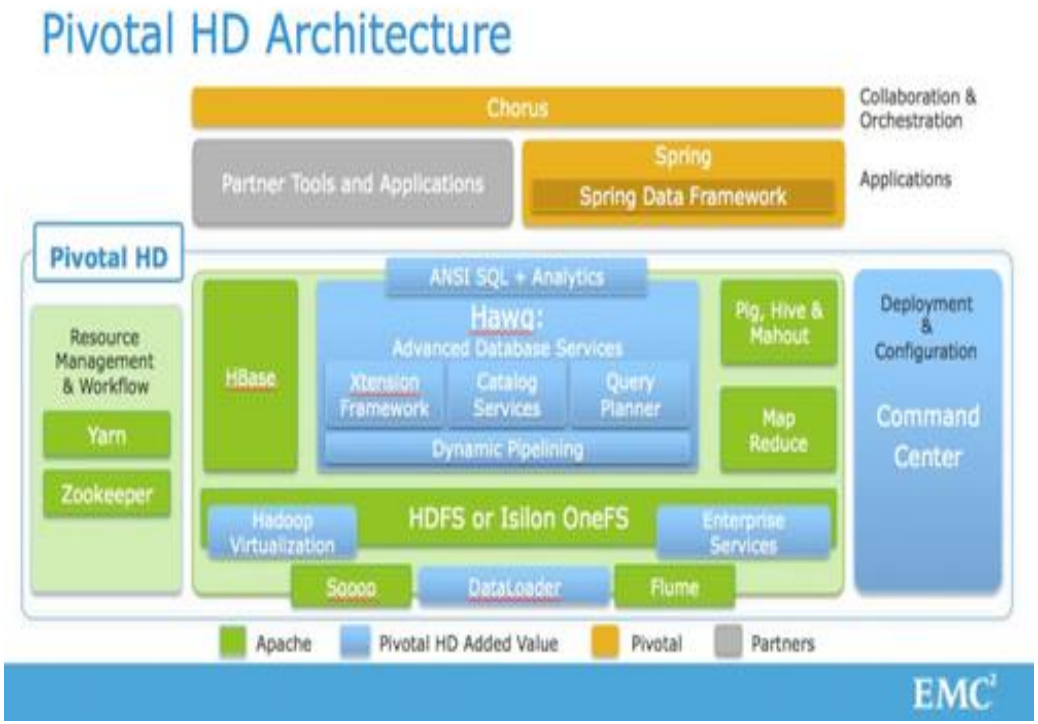
IBM通过其智慧云企业（SmartCloud Enterprise）基础架构，将BigInsights和BigSheets作为一项服务来提供。客户不必购买支持性硬件，也不需要IT专门知识，就可以学习和试用大数据处理和分析功能。据IBM称，客户用30分钟就能搭建起Hadoop集群，并能将现有数据转移到集群里面。

Hadoop主要开发厂商—— ORACLE



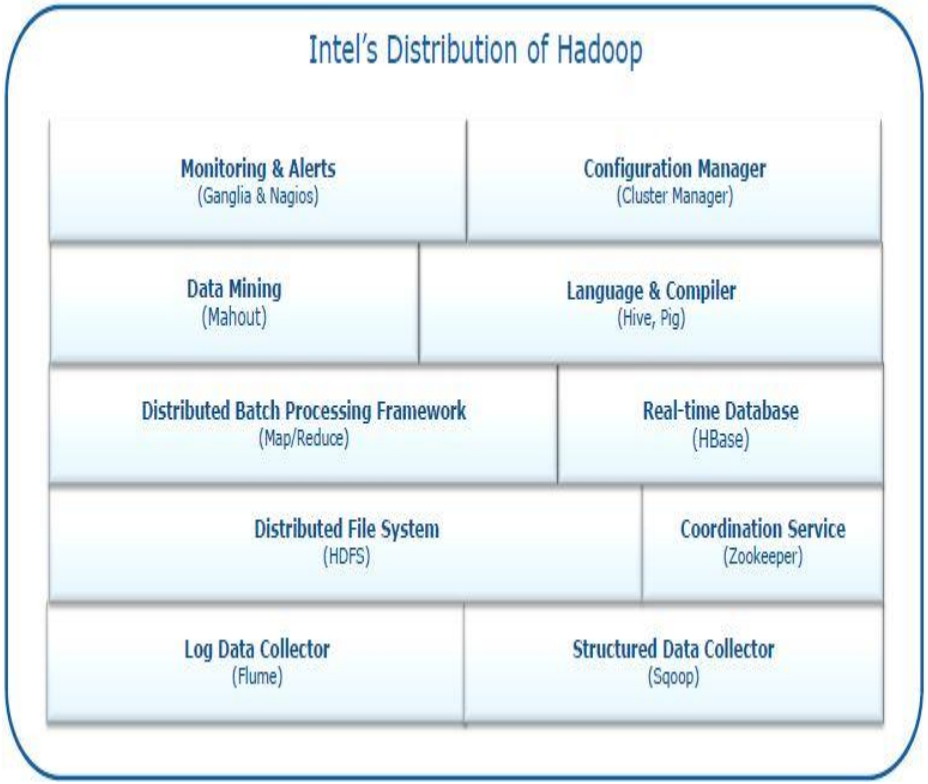
Oracle Big Data机与Oracle Exadata数据库云服务器以及新推出的Oracle Exalytics商务智能云服务器，为客户提供了一个端到端的大数据解决方案，从而为客户在企业内获取、组织、分析大数据以及最大限度地挖掘大数据的价 值提供了所需要的一切条件。 Oracle Big Data机是一款集成设计的系统，并且针对获取、组织以及将非结构化数据加载到Oracle数据库11*g*之 中的整个流程进行优化。Oracle Big Data机包括开源Apache Hadoop、Oracle NoSQL数据库、Oracle数据集成Hadoop应用适配器、Oracle Hadoop装载器。

Hadoop主要开发厂商——EMC

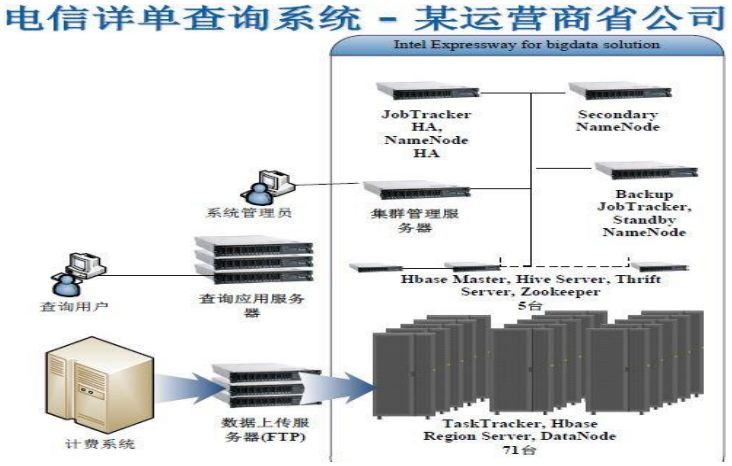


EMC公司于2013年发布了自身的Apache Hadoop发行版——Pivotal HD，同时发布的还有一个名为HAWQ的技术，通过HAWQ能够将Greenplum分析型数据库与Hadoop分布式架构进行紧密地融合。Pivotal HD对Apache Hadoop进行了全面的改造，同其他一些Hadoop发行版相比，其最大的优势就是能够与Greenplum数据库进行整合，Pivotal HD和HAWQ让EMC在Hadoop领域更进一步，同时将成为EMC大数据战略中的一个重要里程碑。

Hadoop主要开发厂商——INTEL



基于在大数据领域的长期技术积累和应用经验，英特尔推出成熟的企业级 Hadoop 发行版，为企业和政府部门实现大数据应用提供强有力的平台支持。英特尔在 Hadoop 上的改进和功能增强为用户提供了一个高性能、高稳定性和可管理的大数据应用实施平台，并提供全面的专业支持。在 Hadoop软件的英特尔分发版在中国推广的两年多时间里，已经在电信行业、智能交通行业有多个成功应用。



1.大数据背景介绍

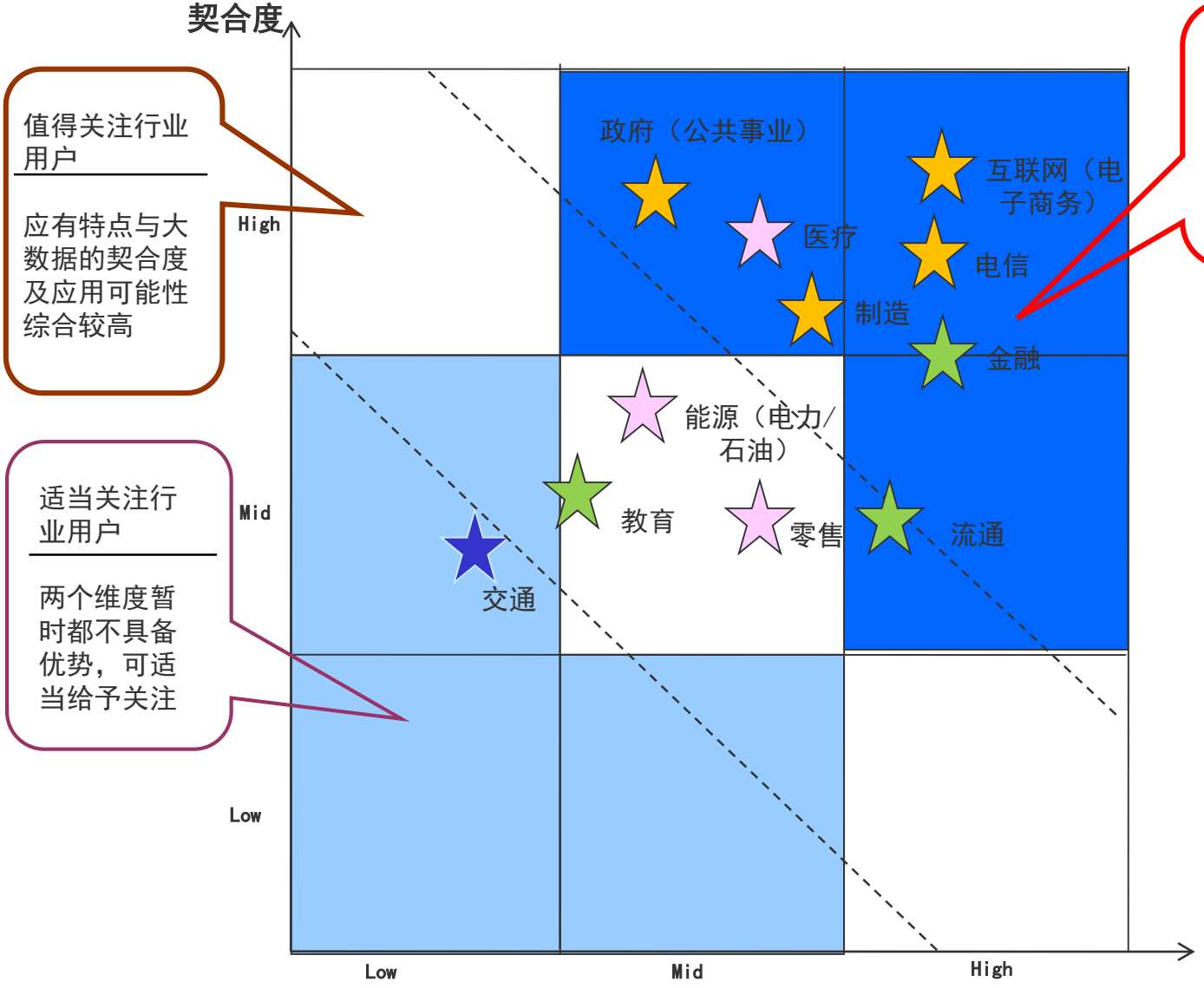
2. HADOOP体系架构

3. 基于HADOOP的大数据产品分析

4. 基于HADOOP的大数据行业应用分析

5.基于HADOOP的大数据应用建议

大数据应用行业分析



- ***纵轴契合度：***

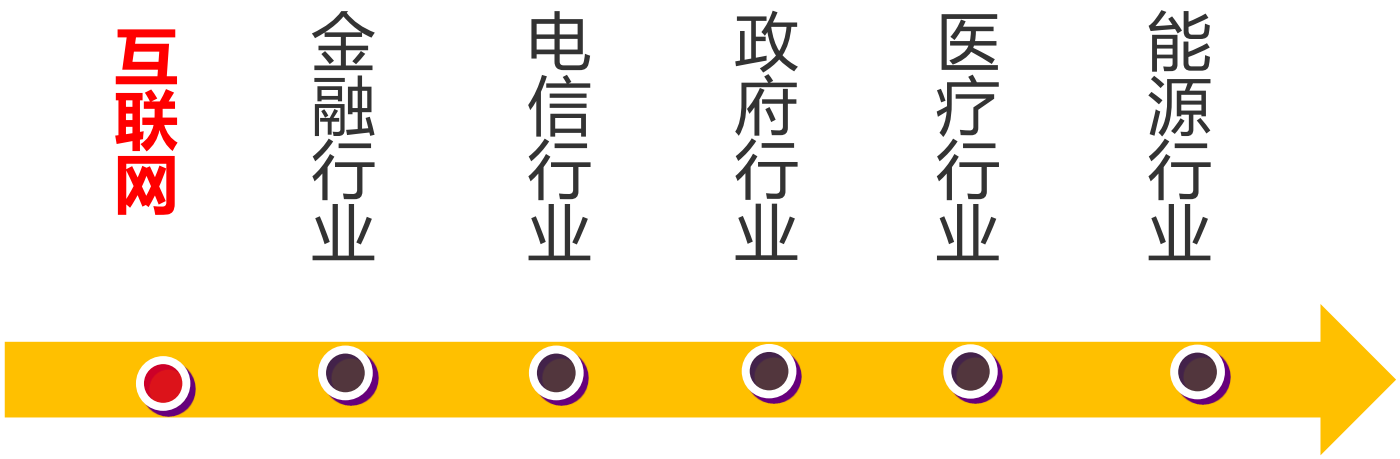
  表示该用户的IT应用特点与大数据特性的契合程度；

- ***横轴应用可能性：****表示该用户出于主客观因素在短期内投资大数据的可能性；*

- ***注：***

  该位置为分析师访谈的综合印象，为定性分析，图中位置不代表具体数值

大数据行业应用分析——互联网行业



互联网行业大数据需求分析

互联网行业拥抱大数据的关键因素

网络终端设备

- 网络技术的升级和终端设备的爆发，使今天的用户能够使用多种设备、从不同位置、通过多种手段来接入互联网，并在这一过程中不断创造新内容

在线应用和服务

- 越来越丰富的在线应用和服务，不断激励用户创造和分享信息，尤其是社会化媒体业务，带动图片、视频等非结构化数据飞速增长

与各垂直行业的融合

- 互联网作为一个高渗透力的行业，正在与各垂直行业发生深度的融合，原本隐藏于先下的孤岛信息，源源不断的输入到线上。

互联网大数据技术的应用，会首先带动社会化媒体、电子商务的快速发展，其他的互联网分支也会紧追其后，整个行业在大数据的推动下将会蓬勃发展。

互联网行业大数据分析面临的主要问题

●互联网行业对数据实时分析要求较高，例如广告监测、B2C业务，往往要求在数秒内返回上亿行数据的分析，从而达到不影响用户体验和快速准确营销的目的。

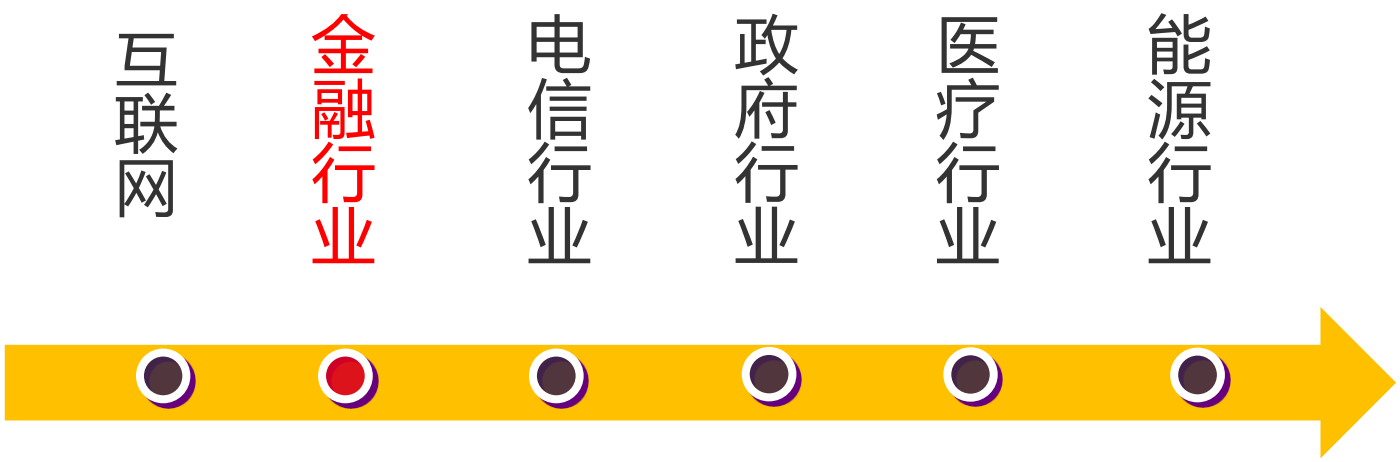
●目前互联网企业面对大数据，会普遍感觉到实时分析能力差、海量数据处理效率低、缺少分析方法、分析软件能力差等问题。

数据来源：CCW Research. 2012/4

互联网行业Hadoop应用

公司	具体应用
HADOOP在阿里巴巴	用于处理商业数据的排序，并将其应用于阿里巴巴的ISEARCH搜索引擎，垂直商业搜索引擎。节点数： 15台机器的构成的服务器集群 服务器配置： 8核CPU，16G内存，1.4T硬盘容量
HADOOP在百度	HADOOP主要应用日志分析，同时使用它做一些网页数据库的数据挖掘工作。 节点数：10 - 500个节点。 周数据量： 3000TB
HADOOP在Facebook	主要用于存储内部日志的拷贝，作为一个源用于处理数据挖掘和日志统计。 主要使用了2个集群：一个由1100台节点组成的集群，包括8800核CPU（即每台机器8核）,和12000TB的原始存储(即每台机器12T硬盘),一个有300台节点组成的集群，包括2400核CPU（即每台机器8核），和3000TB的原始存储(即每台机器12T硬盘),由此基础上开发了基于SQL语法的项目：HIVE
HADOOP在TWITTER	使用HADOOP用于存储微博数据，日志文件和许多中间数据 使用基于HADOOP构件的Cloudera's CDH2系统，存储压缩后的数据文件（LZO格式）
HADOOP在雅虎	主要用于支持广告系统及网页搜索 机器数：25000，CPU：8核 集群机器数： 4000 个节点 (2*4cpu boxes w 4*1TB disk & 16GB RAM)

大数据行业应用分析——金融行业



金融行业大数据需求分析

金融行业大数据需求背景

IDC研究显示，数据是重要资产的理念已经在中国金融行业形成共识，数据的真正价值在于能够洞察企业内部规律，数据的洞察力成为金融企业的核心竞争力。在中国金融行业信息化建设中，与信息加工密切相关的大数据管理正逐渐成为与核心业务系统建设、渠道建设和前置建设同等重要的领域。

经过多年的发展与积累，目前中国的大型商业银行和保险公司的数据量已经达到100TB以上级别，并且非结构化数据量在迅速增长。

金融行业大数据发展分析

从未来几年看，金融行业在“十二五”时期面临发展方式转型的挑战，转型主要集中在三大方面：一，建立全面的风险管理体制，向严监管转型；二，从粗放式管理向精细化管理转型；三，从“利润为中心” 向“客户为中心”转型。

大数据在加强风险管控、精细化管理、服务创新等转型中别具现实意义，是实现向信息化银行转型的重要推动力。金融行业应首先在战略层面对大数据进行规划，积极应对大数据时代的挑战，推进并建立数据驱动型发展方式。

摩根大通基于Hadoop的大数据应用

- 已经开始使用Hadoop技术以满足日益增多的用途，包括诈骗检验、IT风险管理和自助服务。
- 150PB在线存储数据、30,000个数据库和35亿个用户登录账号。

- Hadoop能够存储大量非结构化数据，允许公司收集和存储Web日志、交易数据和社交媒体数据。
- 数据被汇集至一个通用平台，以方便以客户为中心的数据挖掘与数据分析工具的使用。

Zions银行基于Hadoop的大数据应用

美国地区性银行Zions Bancorp(ZIONS)

- 数据仓库存储了120多个不同类型的数据，包括交易日志，日志，欺诈警报，服务器日志，防火墙日志和IDS日志
- 跨整个企业进行数据挖掘，加快取证调查并提高欺诈侦测，以及整体安全性
- 利用**Hadoop**来存储所有数据，并对客户交易和现货异常进行判断，对可能存在欺诈行为提前预警的
- 基于Hadoop的安全数据仓库,迅速对来自各种源头的恶意软件威胁作出响应并对抗它们

中信银行信用卡中心基于Hadoop的大数据应用

大数据挑战

- 发卡量增长迅速：2008年发卡约500万张，2010年增加了一倍。
- 业务数据增长迅速：随着业务的迅猛增长，业务数据规模也线性膨胀。
- 数据存储、系统维护、数据有效利用都面临巨大压力。

需求

可扩展、高性能的数据仓库解决方案

能够实现业务数据的集中和整合；可以支持多样化和复杂化数据分析提升信用卡中心的业务效率；通过从数据仓库提取数据，改进和推动有针对性的营销活动。

EMC Green-plum

→ 未来和基于Hadoop的Pivotal HD相融合

采用大数据方案后价值体现

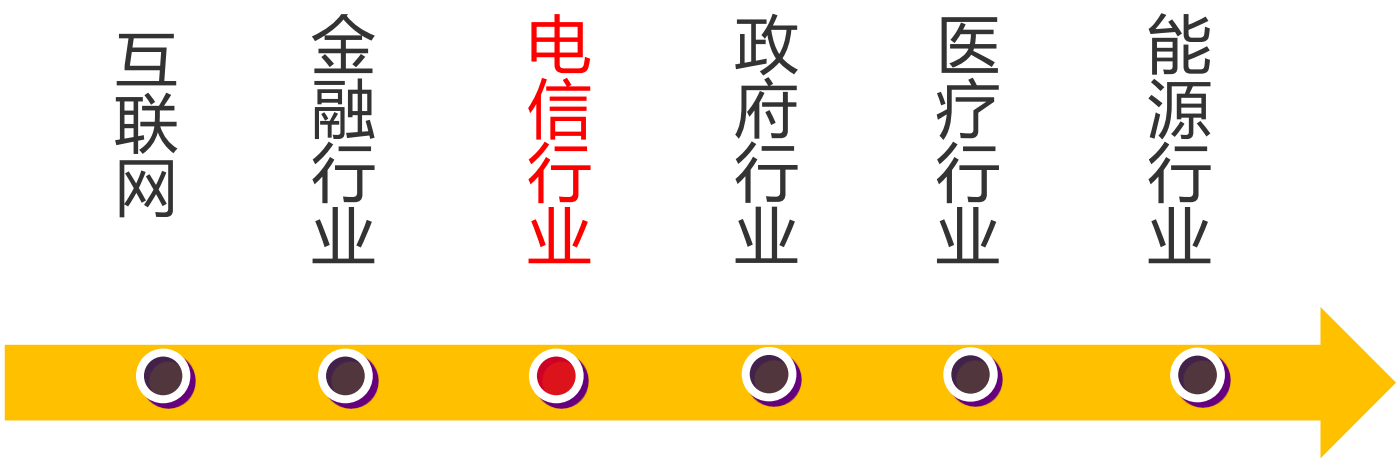
实时的商业智能

可以结合实时、历史数据进行全局分析,风险管理部门现在可以每天评估客户的行为，并决定对客户的信用额度在同一天进行调整；原有内部系统、模型整体性能显著提高

秒级营销

Greenplum数据仓库解决方案提供了统一的客户视图，更有针对的进行营销。2011年，中信银行信用卡中心通过其数据库营销平台进行了1286个宣传活动，每个营销活动配置平均时间从2周缩短到2-3天。

大数据行业应用分析——电信行业



电信行业大数据需求分析

提升网络服务质量，增强管道智能化

随着互联网和移动互联网的发展，运营商的网络将会更加繁忙，用于监测网络状态的信令数据也会快速增长。通过大数据的海量分布式存储技术，可以更好地满足存储需求；通过智能分析技术，能够提高网络维护的实时性，预测网络流量峰值，预警异常流量，有效防止网络堵塞和宕机，为网络改造、优化提供参考，从而提高网络服务质量，提升用户体验。

更加精准地洞察客户需求，增强市场竞争力

客户洞察是指在企业或部门层面对客户数据的全面掌握并在市场营销、客户联系等环节的有效应用。通过使用**大数据**分析、数据挖掘等工具和方法，**电信**运营商能够整合来自市场部门、销售部门、服务部门的数据，从各种不同的角度全面了解自己的客户，对客户形象进行精准刻画，以寻找目标客户，制定有针对性的营销计划、产品组合或商业决策，提升客户价值。判断客户对企业产品、服务的感知，有针对性地进行改进和完善。通过情感分析、语义分析等技术，可以针对客户的喜好、情绪，进行个性化的业务推荐

升级行业信息化解决方案，提升客户价值

智慧城市的发展以及教育、医疗、交通、环境保护等关系到国计民生的**行业**，都具有极大的信息化需求。目前，**电信**运营商针对智慧城市及**行业**信息化服务虽然能够提供一揽子解决方案，但主要还是提供终端和通信管道，**行业**应用软件和系统集成尚需要整合外部的应用软件提供商，对于客户的价值主要体现在网络化、自动化等较低水平。而随着社会、经济的发展，客户及客户的客户对于智能化的要求将逐步强烈，因此运营商如能把**大数据**技术整合到**行业**信息化方案中，帮助客户通过数据采集、存储和分析更好地进行决策，将能极大提升信息化服务的价值

提供数据安全服务，在大数据市场建立差异化竞争优势

大数据也有大风险，其中之一就是客户隐私泄露及数据安全风险。由于大量的数据产生、存储和分析，数据保密和隐私问题将在未来几年内成为一个更大的问题，企业必须

中国移动基于Hadoop的大数据应用

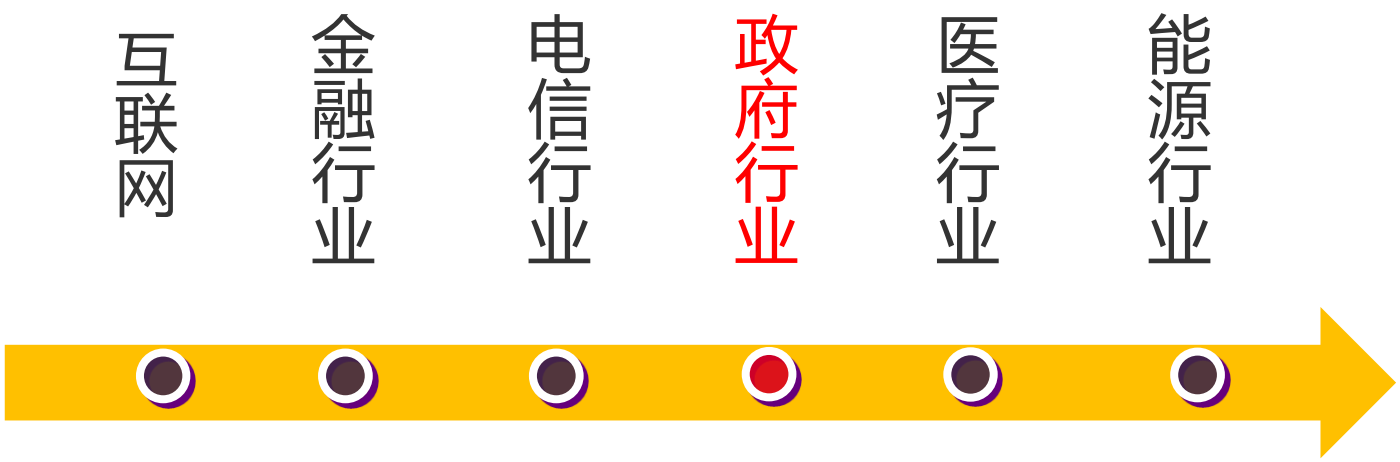
在中国移动“大云”产品总体架构中，分析型PaaS产品底层基于Hadoop数据存储和分析平台，在技术路线方面，选择数据仓库与Hadoop混搭的方式，借鉴关系型数据仓库在传统应用支持方面以及在复杂查询和分析方面的快速响应能力，同时也借鉴了Hadoop的非结构化数据处理能力以及存储的低成本。屏蔽Hadoop与数据仓库的使用细节，让用户在使用这些数据时尽量无感知；在数据的ETL采集预处理环节，尽量采用Hadoop与分布式ETL的方式，提高数据转换效率，同时降低成本。

中国联通基于Hadoop的大数据应用

中国联通已经构建了一个全国集中的一级架构海量数据存储和查询系统：通信用户上网记录集中查询与分析支撑系统，在集团公司进行统一部署，各个省分仅仅是做数据的采集，按照业务实时性将数据传送到集团公司，由集团公司统一处理，全国所有用户所有上网记录数据都放北京数据中心里，在国内**电信行业**当中也是首创的方式。

中国联通成功将大数据和**Hadoop**技术引入到‘移动通信用户上网记录集中查询与分析支撑系统’。截止到目前已经部署了4.5PB的存储空间。其中，4.5PB的存储分布在300个数据节点上，即每个节点配备15TB的存储空间。系统每天有能力处理700亿条上网记录。

大数据行业应用分析——政府行业



政府行业大数据需求分析

1、加强统筹规划，优化大数据形成机制。强化对大数据建设工作的组织协调，打破地区和部门数据壁垒，实现数据资源联合共建、广泛共享。建立政府和社会联动的大数据形成机制，以政府数据公开共享，推动公共数据资源的开发利用。

2、加强数据收集和信息感知，提高智慧城市感知水平。加强政府部门在管理和服务过程中对数据的主动采集，建立政府大数据库。鼓励制造业企业和商业机构加强对生产经营活动中的数据采集，形成覆盖生产过程和商业各环节各流程的数据库。推进无线识别技术、传感器、无线网络、传感网络等新技术的广泛应用，提高数据采集的智能化水平。

3、推进大数据应用，提高经济社会智慧化水平。推进政务信息公开。推行政府网上办事，收集分析挖掘社会政务服务需求，推进公共服务个性化和政府决策智能化。支持公共服务机构和商业机构开放与社会民生密切相关的公共数据。推进国民经济各行业和企业数据开发，发展商业智能。鼓励开展服务大众的大数据应用，提升智慧生活品质。

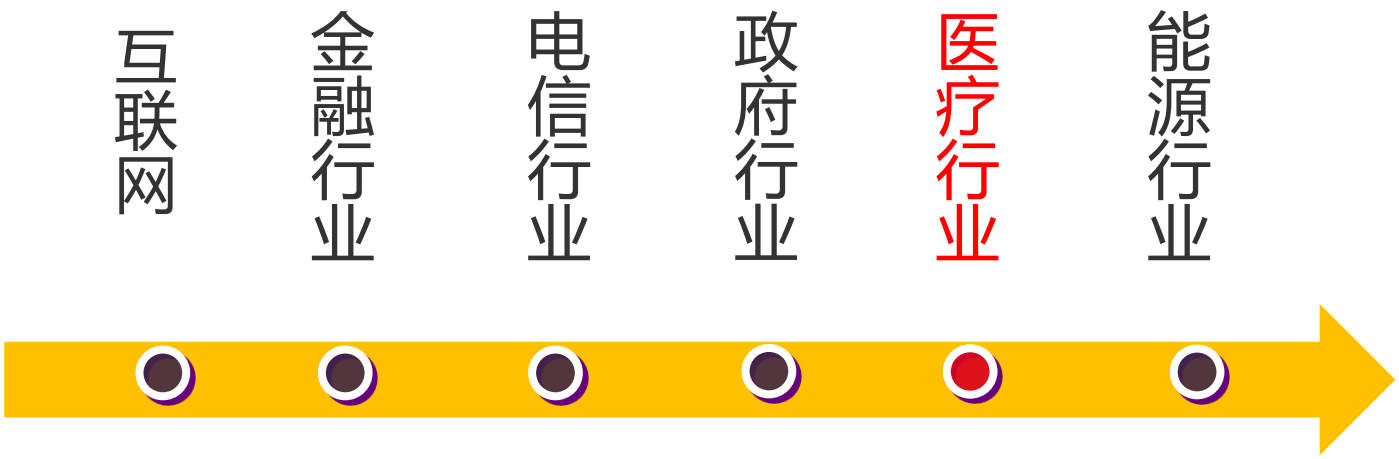
政府行业大数据应用——智慧城市

智慧城市

2013年1月29日，住房和城乡建设部公布了首批９０个国家智慧城市试点名单，试点城市的公布标志着我国智慧城市发展进入规模推广的阶段。在目前智慧城市的发展阶段，主要的应用还处于对感知设备传递的信息进行简单处理的水平，充分认识大数据对于智慧城市建设的关键作用，对于避免智慧城市建设中出现“重感知，轻智慧”的通病具有重要意义。

从智慧城市的体系结构来看，由于智慧城市的基础在于物联网技术，因此智慧城市体系架构和物联网的体系结构相类似，也可分为四层，分别为感知层、传输层、平台层、应用层。智慧城市相对于之前数字城市概念，最大的区别在于对感知层获取的信息进行了智慧的处理，因此也可以认为智慧城市是数字城市的升级版。由城市数字化到城市智慧化，关键是要实现对数字信息的智慧处理，其核心是大数据处理技术。

大数据行业应用分析——医疗行业

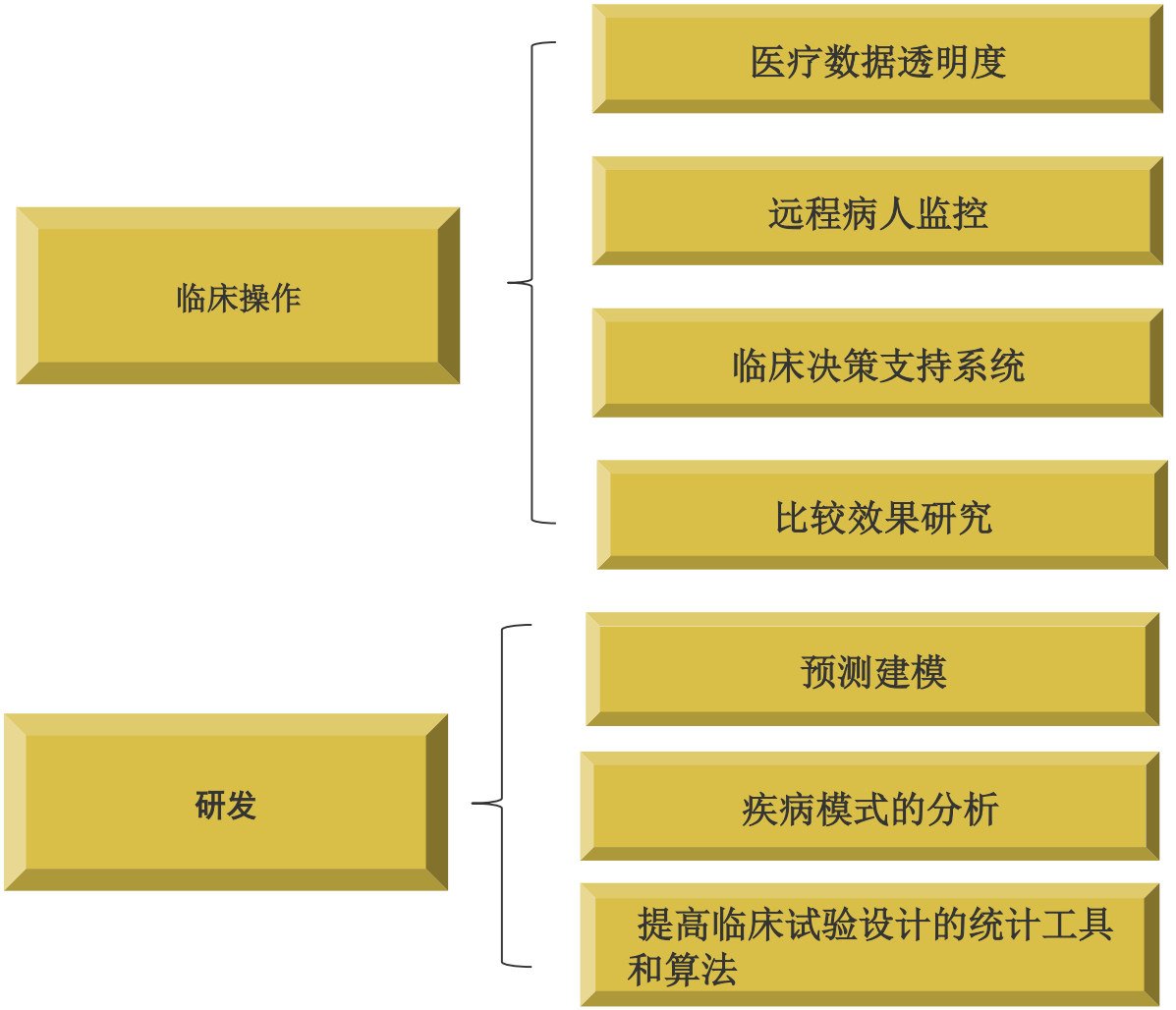


医疗行业大数据需求分析

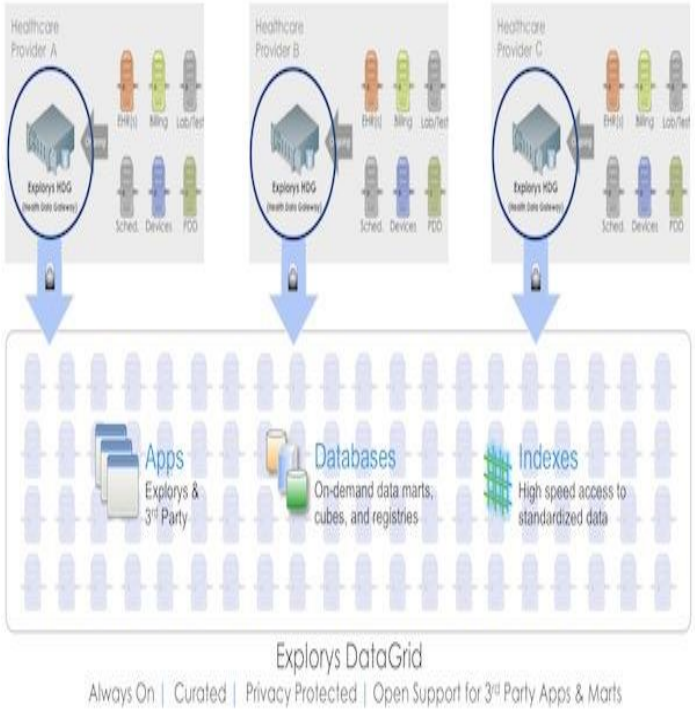
●医疗行业产生的数据量主要来自于PACS影像、B超、病理分析等业务所产生的非结构化数据。人体不同部位、不同专科影像的数据文件大小不一，PACS网络存储和传输要采取不同策略。面对大数据，医疗行业遇到前所未有的挑战和机遇。

●医疗行业大数据应用场景非常多，右图仅以临床操作和研发为例，展示医疗行业大数据应用场景。

●对于公共卫生部门，可以通过过覆盖全国的患者电子病历数据库，快速检测传染病，进行全面的疫情监测，并通过集成疾病监测和响应程序，快速进行响应。



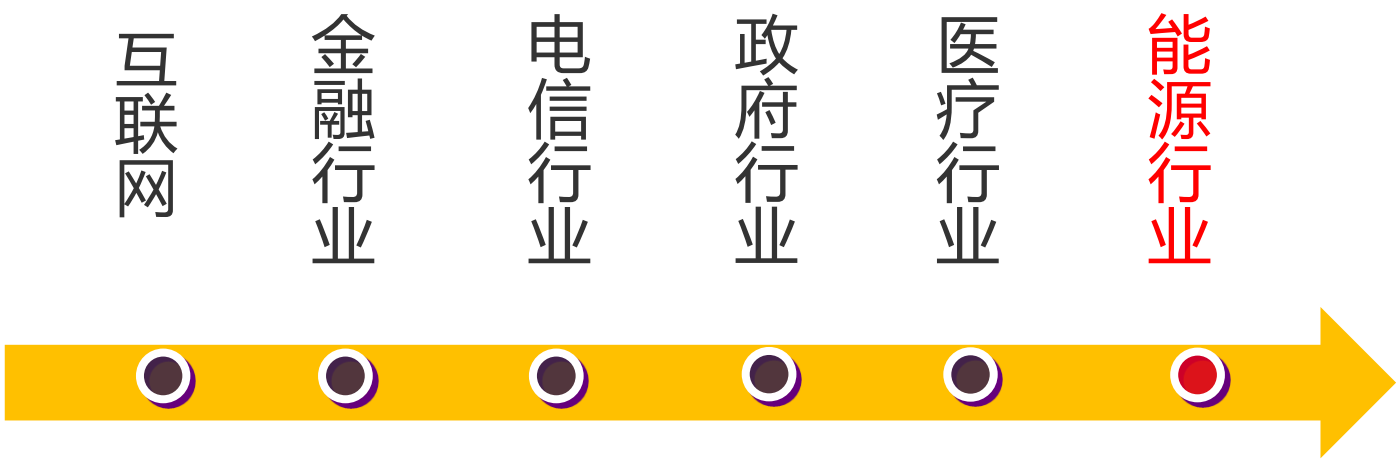
医疗行业基于Hadoop的大数据应用



Cloudera正在与西奈山医学院合作开发新的生物数据分析方法和系统。Cloudera还与FDA合作侦测多种药物组合的副作用，与埃默里大学合作帮助病历学家更准确地分析医疗影像。Cloudera的客户之一——Explorys的业务主要是聚合并分析医疗记录，而英特尔和NextBio则合作使用Hadoop处理基因数据。

Apixio利用Hadoop平台开发了语义分析服务，可以对病人的健康提供医生、护士、及其他相关人士的回答。Apixio试图通过对医疗记录进行先进的技术分析，与一个简单的基于云计算的搜索引擎来帮助医生迅速了解病人相关病史，挽救生命。

大数据行业应用分析——能源行业

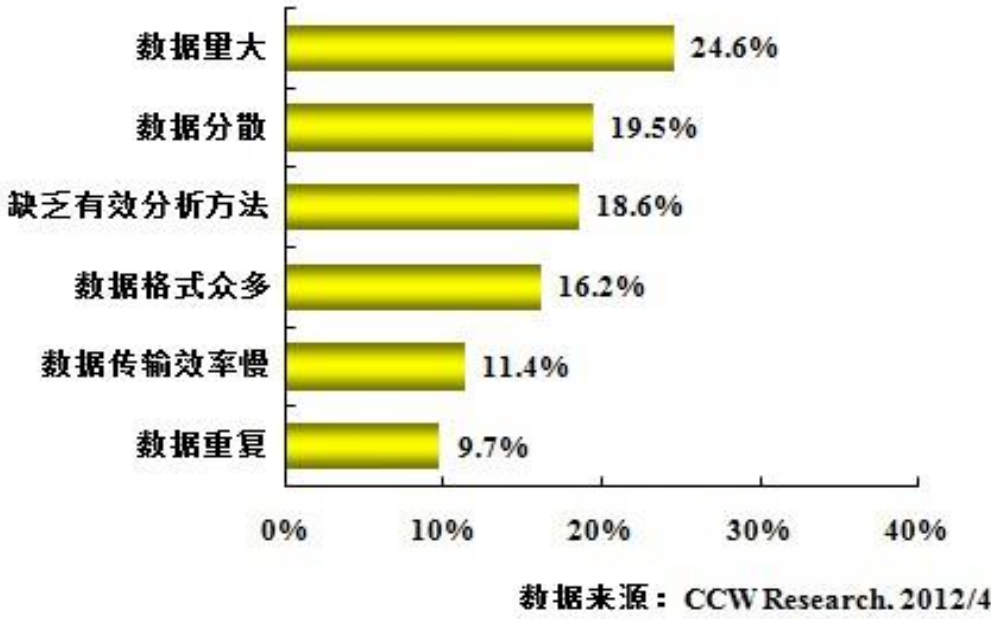


能源行业大数据需求分析

能源行业数据特征

能源勘探开发数据的类型众多，不同类型数据包含的信息各具特点，综合各种数据所包含的信息才能得出地下真实的地质状况。

能源行业面临的大数据问题



数据来源：CCW Research. 2012/4

能源行业企业对大数据产品和解决方案的需求集中体现在：可扩展存储、高带宽、可处理不同格式数据的分析方案。

能源行业基于Hadoop的大数据应用

- Opower使用Hadoop来提升电力服务，尽量为用户节省在资源方面的投入。Opower现在管理着30TB的信息，其中包括来自5000万用户（横跨60个公共事业部）能源数据，气象与人口方面的公共及私人数据，历史信息，地理数据及其他。这些都是通过超过20个MySQL数据库和一个Hadoop集群来存储和处理的。
- 采用Hadoop来对来自从海洋深处地震时产生的数据进行排序和整理，其背后有可能意味着石油储量。

谢谢观赏！！！