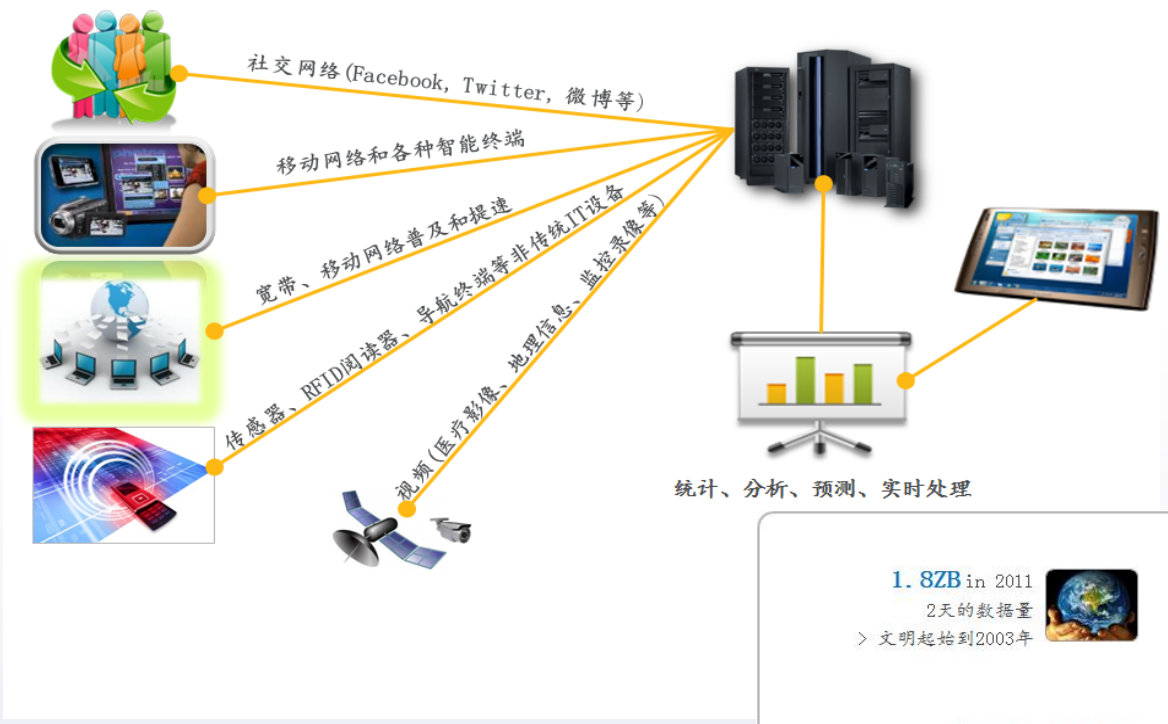
大数据 – 新一轮“信息革命”



IDC预测全球的数据使用量到2020年会增长44倍，达到35.2ZB（1ZB = 10亿TB）

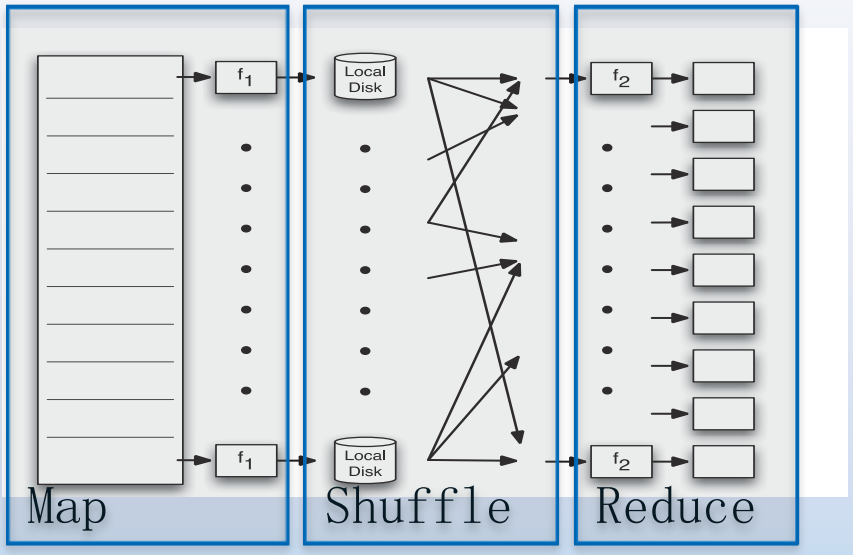
英特尔为企业大数据应用提供端到端的支撑

火龙果•整理
uml.org.cn

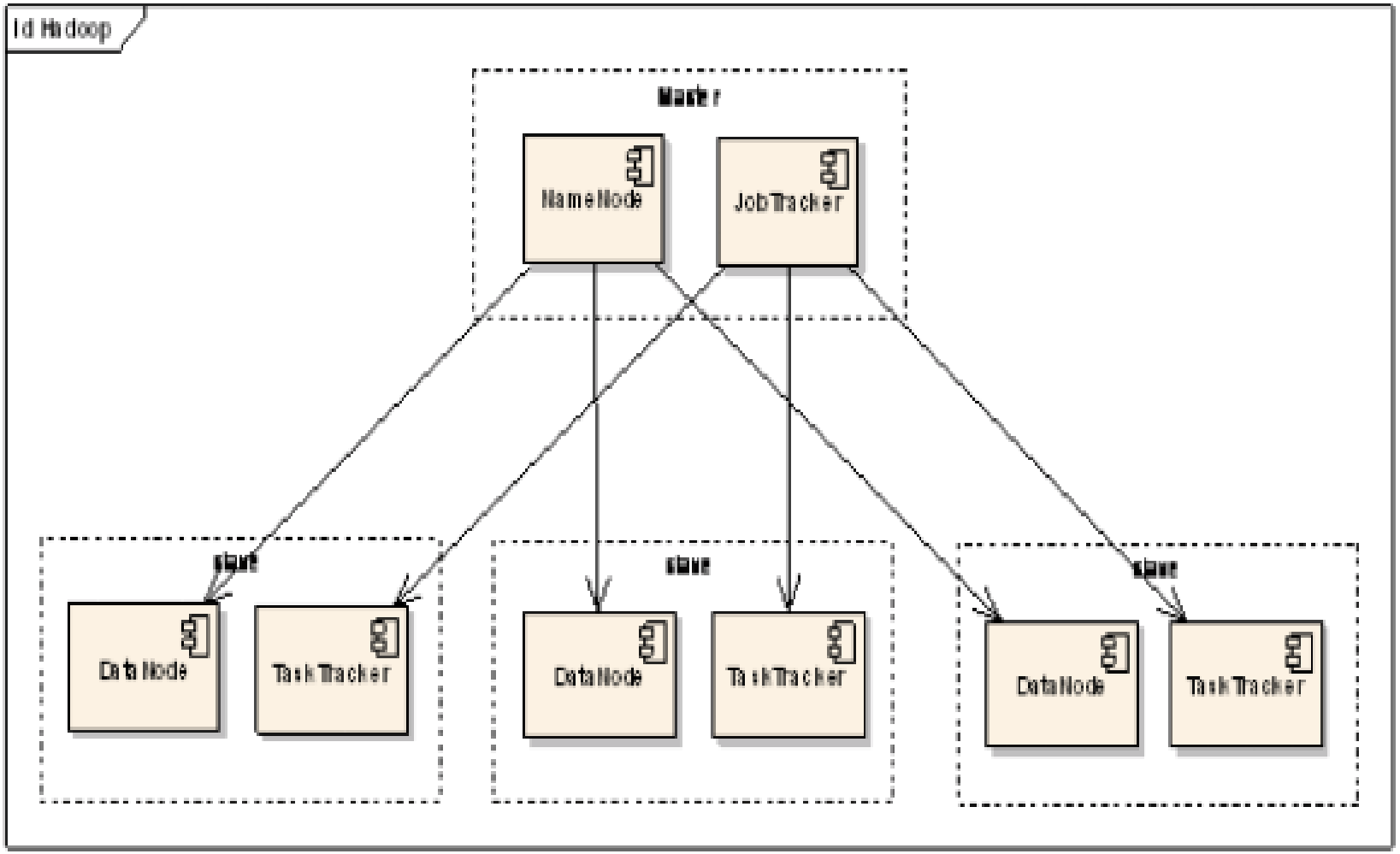
数据中心方案选择			
大数据存储	- HDFS	实时数据库	- HBase
大数据分析	- Hadoop	高性能存储	- Lustre

什么是Hadoop?

- 最重要的大数据平台之一
- 开源Apache 项目，灵感来源Google的MapReduce白皮书和Google文件系(GFS)，Yahoo完成了绝大部分初始设计和开发（ Doug Cutting ）
- Hadoop 核心组件包括:
  - ✓ 分布式文件系统
  - ✓ Map/Reduce – 分布式计算
- 用Java编写
- 运行平台:
  - ✓ Linux, Mac OS/X, Solaris
  - ✓ 普通的X86硬件平台



英特尔Hadoop发行版数据平台整体架构



Hadoop作为建设集中化大数据平台的标准之一得到广泛支持

火龙果•整理
uml.org.cn

2007

YAHOO!
Powerset
last·fm
intel

2008

Google
able grape
ImageShack
Cascading
IBM
facebook
ENORMO Every property. Everywhere.
A9
THE UNIVERSITY OF EDINBURGH
krugle
rackspace HOSTING
Lookery Control freaks welcome
The New York Times
joost
Zvents Discover Things To Do
FORMATION CIENCES NSTITUTE
News Corporation
Cornell University Computing and Information Science
Visible MEASURES
LOTAME Locate, Target, & Message with Social Media
NetSeer
parc Palo Alto Research Center
SECURITY ENHANCED DOMAIN NAME SYSTEM
veoh

2009

AOL
cloudera
deepdyve BETA
cooliris
eyealike
TEXTMAP THE ENTITY SEARCH ENGINE
PSG College of Technology
iterend
tailsweep
hulu
RapLeaf
USCMS
Ning
quantcast
amazon web services
pressflip
detikSearch
WorldLingo
Systems@ETH zürich
VK SOLUTIONS Global Solutions Provider
TARAGANA Innovation • Quality • Simplicity
HOSTING HABITAT
HOLA SERVERS
Terrier
adknowledge
stampede beta

2010

SAMSUNG
rubicon PROJECT
BERKELEY LAB LAWRENCE BERKELEY NATIONAL LABORATORY
VISIBLE TECHNOLOGIES
APOLLO GROUP
ADSDAQ
rackspace HOSTING
RapLeaf
wordnik All the words.
MOBIGEN
comSCORE
trulia real estate search
Accela COMMUNICATIONS
Forward3D
LinkedIn
Microsoft
Infochimps Find the world's data
Pharm2Phork
ADMELD
gumgum
BrainPad
Pronux
The Datagraph Blog
NETFLIX
mobileanalytics.tv
markt24.de
twitter
media6degrees
BEEBLER
SLC Security When Experience Matters...
ebay

Hadoop在互联网行业最先得到广泛应用

火龙果•整理
uml.org.cn

淘宝网

淘宝Hadoop集群（2012年数据）

规模：3000 台左右

总数据量30PB

数据每天增长100T

Baidu百度

百度（2012年数据）

Hadoop集群规模超过1万台

服务器大部分标配12个1TB硬盘，32GB内存

facebook

Facebook（2010 数据）

超过2250台机器和23000个核心

数据容量超过36PB

每天运行超过25000个Hive任务

YAHOO!

Yahoo（2011 数据）

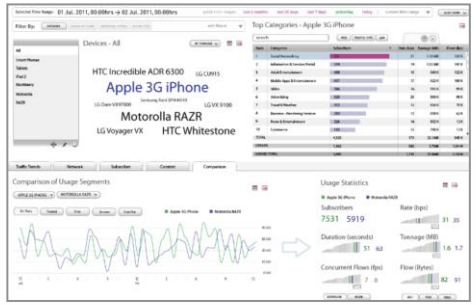
14个集群，总共超过25000台机器，最大的集群有4000台服务器

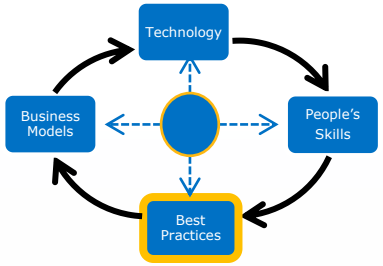
超过3PB的压缩数据

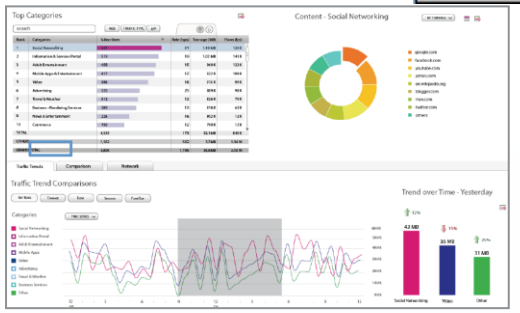
每天运行超过10000个以上作业

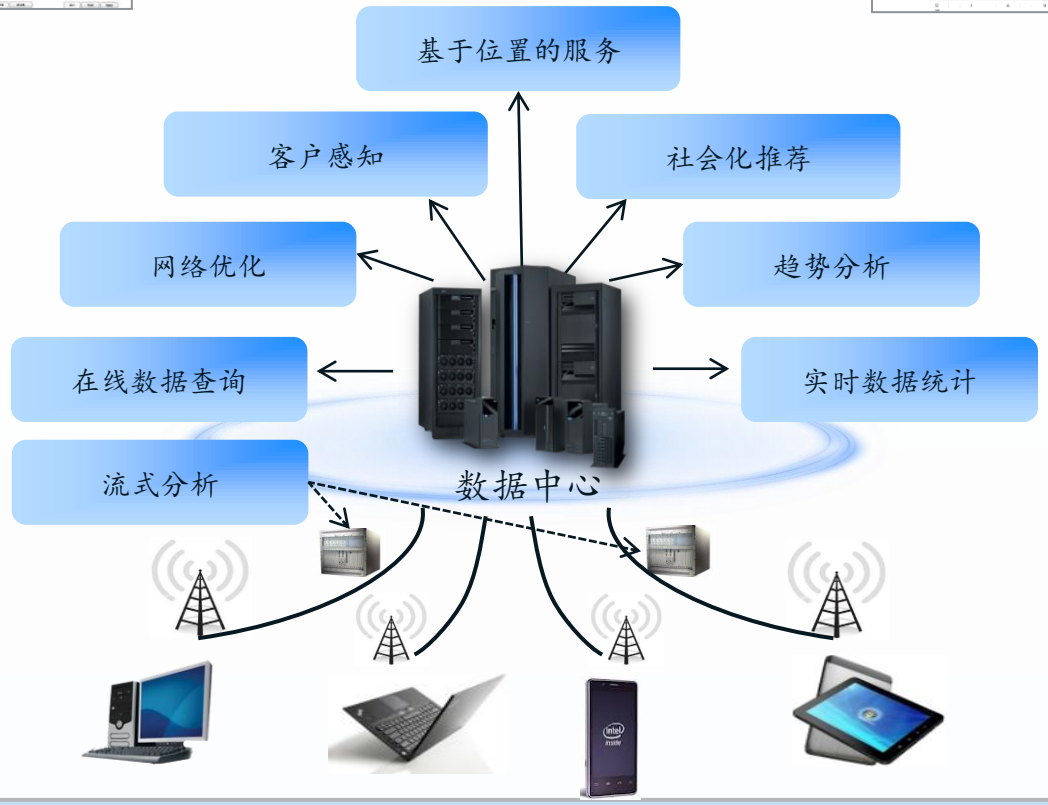
Hadoop适合电信行业大数据应用的场景举例

火龙果•整理
uml.org.cn









Hadoop正在得到传统行业的认可-医疗行业

火龙果•整理
uml.org.cn

面临挑战

数据源	结构化、半结构化（病历），非结构化（PACS影像）
数据格式	标准难制定，或不断变化
数据量	中型城市（1000万人口）存放50年会达到10PB

区域医疗健康档案系统

解决问题	居民电子病历存储、查询
解决方案	HBase分布式数据库 存放健康档案和PACS影像数据
特点优势	高速数据导入；实时数据查询； 关键字搜索； 分布式统计

信息共享提升效率

医疗信息区域内准实时共享，医生可快速调阅病人信息

决策辅助系统

趋势分析：例如，流行病扩展情况分析、癌症的历年趋势、药物效果分析

计算机辅助诊断

疾病自动分类和诊断

疾病管理

业务分析

健康管理

数据分析

基于健康档案的区域平台

基层卫生信息系统

个性化医疗

基于病史的自动医疗、饮食等建议；针对病史和病症的OTC药物参考等

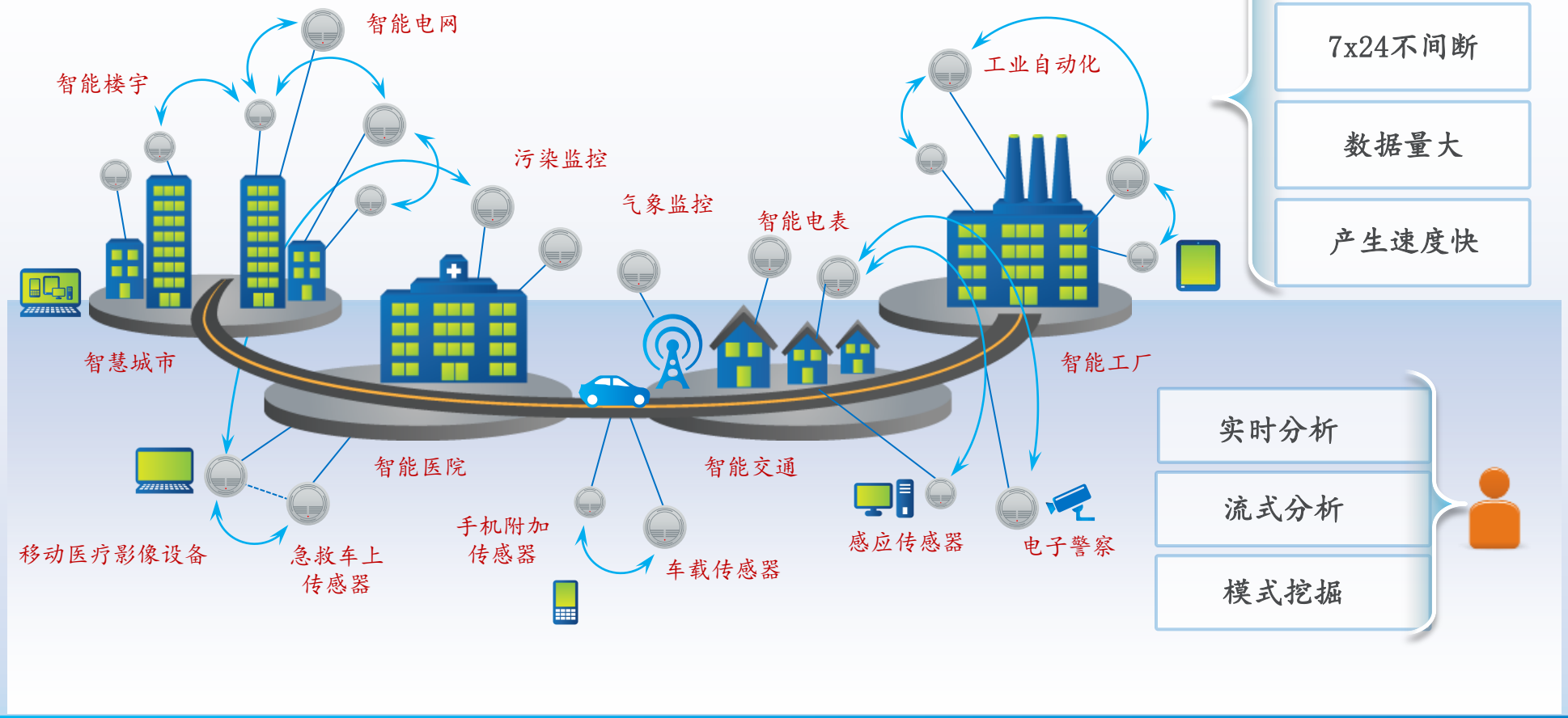
基本医疗

社区护理

家庭病床

数据采集/临床诊疗

Hadoop正在得到传统行业的认可-智慧城市



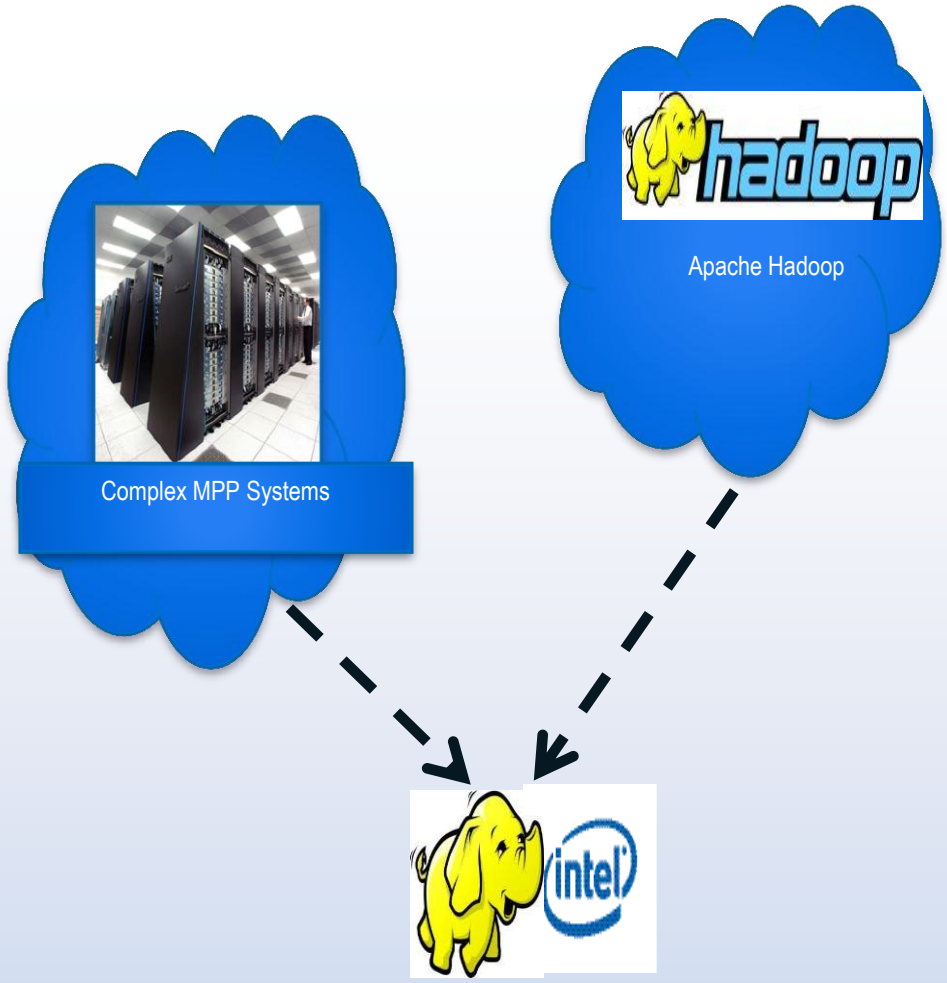
- 机器生成数据
- 7x24不间断
- 数据量大
- 产生速度快

- 实时分析
- 流式分析
- 模式挖掘

英特尔与Hadoop大数据平台

火龙果•整理
uml.org.cn

- Apache Hadoop的主要代码贡献者之一
- Intel提供符合企业级需求的 Hadoop 数据平台
  - 经过测试和验证的稳定版本，在生产环境成功部署运营
  - 包括了Intel针对现有客户在实际使用中出现问题的解决方法以及改进和优化
  - 基于Intel在云计算研发上的经验积累，提供专业的咨询服务，帮助构建高可扩展高性能的分布式系统
  - 结合Intel的硬件部门，提供全面的软硬件解决方案
  - 为Intel硬件平台进行优化，提供更高性能



英特尔Hadoop发行版数据平台堆栈

火龙果•整理
uml.org.cn

英特尔Hadoop发行版组件

稳定的企业级Hadoop发行版	利用硬件新技术进行优化
HBase改进和创新，为Hadoop提供实时数据处理能力	针对行业的功能增强，应对不同行业的大数据挑战

Intel Hadoop Manager 2.3
安装、部署、配置、监控、告警和访问控制

- Sqoop 1.4.1 关系型数据库同步工具
- Flume 1.3.0 日志收集工具
- Mahout 0.7 数据挖掘
- R - statistics 数据分析
- Hive 0.9.0 数据仓库
- Pig 0.9.2 数据流处理
- Oozie 3.3.0 工作流调度
- **Map/Reduce 1.0.3** 分布式计算框架
- **HBase 0.94.1** 实时、分布式、高维数据库
- **HDFS 1.0.3** 分布式文件系统
- Zookeeper 3.4.5 分布式协作服务

英特尔Hadoop发行版大数据平台 – 主要特色

火龙果•整理
uml.org.cn

提供完善的企业级产品支撑

- 全面测试的企业级发行版，保证长期稳定运行，集成最新开源的和自行开发的补丁，用户可以及时修正漏洞保证各个部件之间的一致性，使应用顺滑运行

实时数据处理的分布式大数据应用平台

- 通过对 HBase 进行改进和创新，英特尔 Hadoop 发行版提供实时数据处理功能。为企业对数据的实时监控和即时处理提供有效保障

针对企业用户开发的新的平台功能

- 提供企业关键应用程序所需的即时大数据分析，以及其他针对企业用户需要的增强功能，例如：提供跨数据中心的 HBase 数据库虚拟大表功能，实现 HBase 数据库复制和备份功能，等等。

提供底层 Hadoop 性能优化算法和稳定性增强

- 基于 Hadoop 底层的大量优化算法，配合英特尔优化架构，使应用效率更高、计算存储分布更均衡，系统安装程序计算得出的优化参数配置，适合大多数应用情况，与硬件技术相结合，提高平台性能

提供企业必须的管理和监控功能

- 提供独有的基于浏览器的集群安装和管理界面，解决开源版本管理困难的问题，提供网页、邮件方式的系统异常报警

英特尔为提供企业级的Hadoop平台支持保障

火龙果•整理
uml.org.cn

服务形式	内容描述	备注
远程技术支持	客户可以通过电话或者网络方式7x24小时联系Intel，客户需要提供问题细节描述和问题严重性等级。Intel按照服务协议提供远程响应。 服务内容包括： 1、有效服务期内，根据用户给出的问题严重性以及描述进行远程响应 2、有效服务期内，为用户提供软件补丁，或者新版本软件 3、有效服务期内，用户可以通过Intel官方网站获得相关技术资料和手册	
现场技术支持	Intel根据用户请求，按照相关服务等级，安排相关技术人员到用户现场提供技术服务。 服务内容包括： 1、Intel Hadoop软件安装配置 2、Intel Hadoop软件升级 3、Intel Hadoop平台日常巡检 4、Intel Hadoop软件问题，故障排查	
技术咨询服务	用户可根据自身项目需求，选择性购买基于Intel Hadoop软件生态系统的技术咨询服务。 服务内容包括： 1、Intel Hadoop系统设计规划咨询 2、Intel Hadoop系统开发咨询 3、Intel Hadoop管理培训 4、Intel Hadoop开发培训	

英特尔Hadoop核心研发团队在中国可以更好的支持本地客户

火龙果•整理
uml.org.cn

Intel为企业提供大数据设备参考

火龙果•整理
uml.org.cn

Specification	Detail
 Up to 20 nodes (for compute and storage)	Per node: ● CPU: 2x Intel(R) Xeon(R) CPU 8-core E5-2650 2.00GHz (Up to 40 CPUs in total for one rack) ● Memory: Maximum 384GB (Up to 7,680 GB for one rack) ● Storage: Maximum 36TB, 12x 3T 3.5″ SATA hard disks(Up to 720 TB for one rack)/ 24 x 1T 2.5″ SATA hard disks (Up to 480 TB for one rack) ● Connection ■ Ethernet: 4x 1 Gb Ethernet Ports ■ Infiniband: 2x QDR (40Gb/s) / FDR (56Gb/s) ports (with add-on module) ● Administration: ■ 1x shared or dedicated IPMI v2.0 Ethernet port ■ Remote KVM (Keyboard, Video, Mouse) ■ Support Intel (R) Intelligent Power Node Manager ● Power supply: 2x redundant 750W ● Form factor: 16.7″W x 16.5″L (or 424 mm x 419 mm)
1 24-port QDR InfiniBand switch	24x InfiniBand QDR ports
1 48-port 1GbE Ethernet switch	48x 1GbE ports
Packaging and accessories	42U rack packaging InfiniBand cables: QSFP/QSFP+ Ethernet cables: RJ45

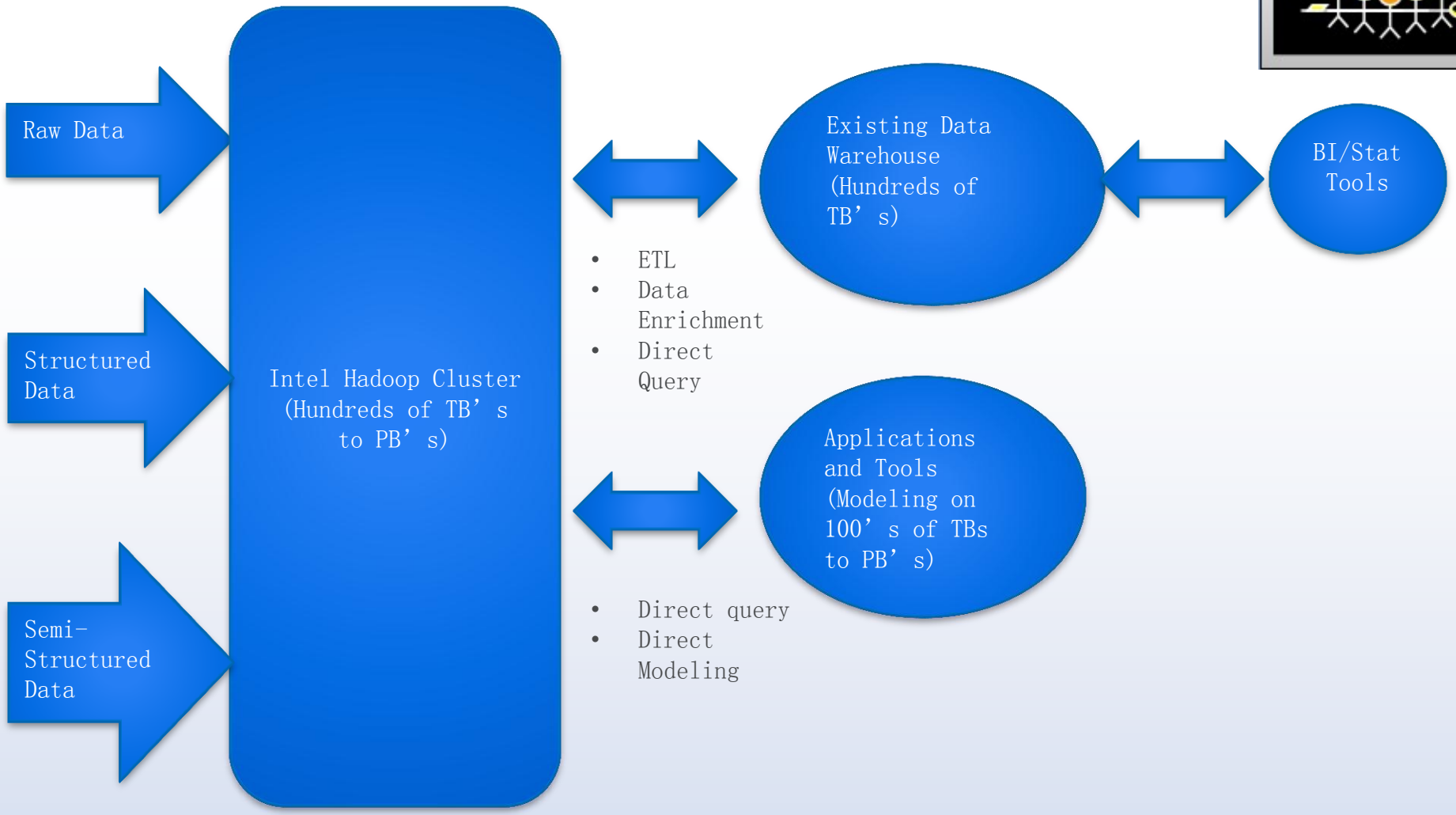
General Package (Channels)
CentOS
Java Platform Standard Edition
Intel Distribution of Hadoop

Vertical Package (ISVs)
CentOS
Java Platform Standard Edition
Intel Distribution of Hadoop
Telecom CDR analytics
Real-time ITS

[1] Expandable to 8-core E5-2680 2.70GHz or 6-core E5-2667 2.90GHz for higher computing capacity
[2] With additional accessories, provide options to support SAS hard disks.
[3] With other optional accessories, provide options to change to 4x 1Gb Ethernet, or 2x 10Gb Ethernet

IDH与原有数据平台有效整合形成完整大数据解决方案

火龙果•整理
uml.org.cn



Historical/Archival/Event Level | Aggregated/Enriched | Summary/Interactive

谢谢