# 英特尔Hadoop解决方案介绍

英特尔亚太研发有限公司

# 海量数据应用发展趋势

# 海量数据的时代正在到来

## 巨大的数据量

- 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)
  - 宽带普及和提速(直接导致访问量、网络访问日志、通讯记录等迅猛增加)
  - 社交网络(Facebook, Twitter, 微博等)
  - 视频(视频通讯、医疗影像、地理信息、监控录像等)
  - 移动网络和各种智能终端
  - 传感器、RFID阅读器、导航终端等非传统IT设备

## 数据集特点

- 超过80%的数据是非结构化的
- 数据量在持续增加
- 数据需要长时间存储，非热点数据也会被随机访问

## 传统技术无法胜任大数据集的分析、管理和挖掘

- 传统关系数据库以及一些桌面BI分析软件处理的结构化数据在GB级别，无法从更大的数据集中发现有意义的信息。
- 需要处理的目标数据量一直在增长，传统技术无法适应这种扩展性

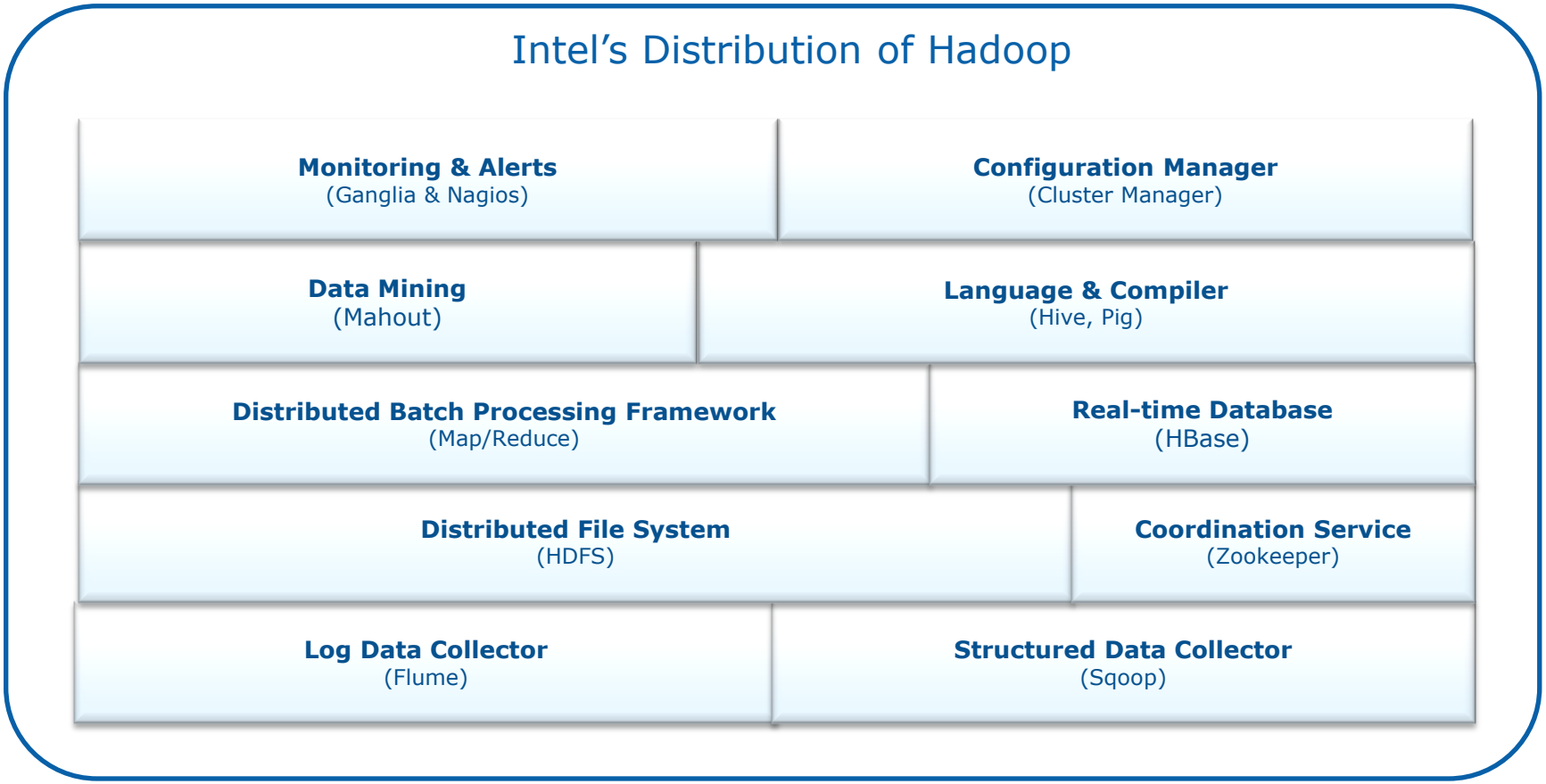
# 什么应用适合大数据产品方案


大量的结构化和非结构化数据、要求可变的数据结构和高效的数据导入、查询、统计等

*Source: McKinsey Global Institute Analysis
SG Cross Asset Research, PwC

# 英特尔Hadoop发行版介绍

# 英特尔Hadoop发行版组成



# 英特尔Hadoop发行版优势

**更高性能**
- 基于Hadoop底层的大量优化算法，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

**稳定运行**
- 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

**易于管理**
- 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

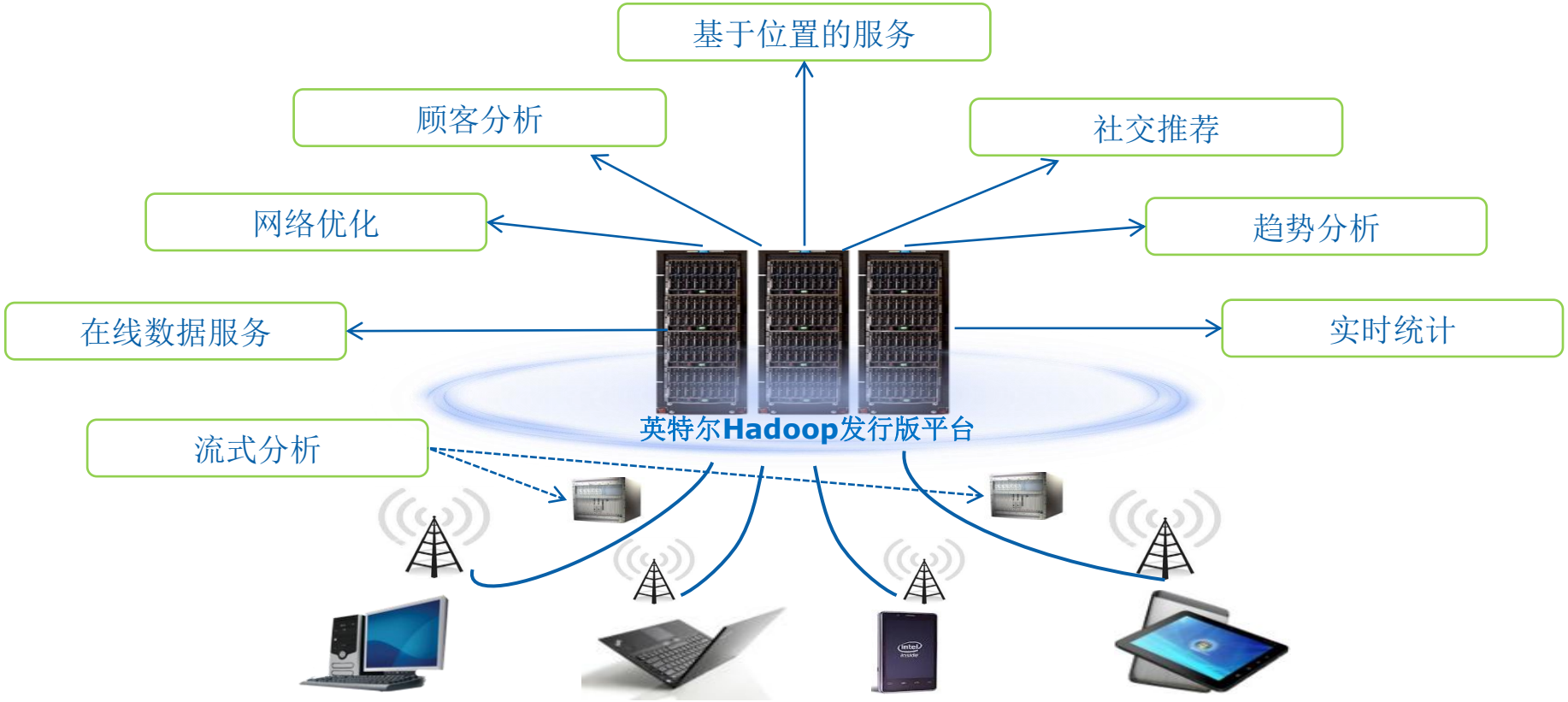
**功能增强**
- 提供跨数据中心的HBase数据库虚拟大表功能
- 实现HBase数据库复制和备份功能
- 其他针对企业用户需要的增强功能

# 英特尔Hadoop发行版与开源版本功能比较

| 英特尔Hadoop发行版增强功能 | 开源系统原始实现 |
| --- | --- |
| 针对HDFS数据节点的读写选取提供高级均衡算法，提高系统扩展性，适合不同配置服务器组成的集群 | 简单均衡算法，容易在慢速服务器或热点服务器上产生读写瓶颈，最慢服务器成为系统性能瓶颈 |
| 根据读请求并发程度动态增加热点数据的复制倍数，提高Map/Reduce任务扩展性 | 无法自动扩充倍数功能，在集中读取时扩展性不强，存在性能瓶颈 |
| 为HDFS的NameNode提供双机热备方案，提高可靠性 | NameNode是系统的单点破损点，一旦失效系统将无法读写 |
| 实现跨区域数据中心的HBase超级大表，用户应用可实现位置透明的数据读写访问和全局汇总统计 | 无此功能，无法进行跨数据中心部署 |
| 可将HBase表复制到异地集群，并提供单向、双向复制功能，实现异地容灾 | 没有成熟的复制方案 |
| 在HBase中，根据数据局部性、服务器Region数、表的Region数来实现负载均衡，适合多用户共享集群创建多张大表的应用 | 只根据Region数量进行负载均衡，容易产生系统不均衡 |
| 基于HBase的分布式聚合函数，比传统方式提高10倍以上效率 | 无成熟方案 |
| 实现对HBase的不同表或不同列族的复制份数精细控制 | 无此功能 |
| HBase的Major Compaction精细控制 | 简单算法，容易产生合并风暴 |

# 行业解决方案

# 海量数据电信解决方案


基于位置的服务
顾客分析
社交推荐
网络优化
趋势分析
在线数据服务
实时统计
流式分析
英特尔Hadoop发行版平台

# 案例一：电信详单查询系统 - 某运营商省公司

业务问题

- 提供所有手机用户的详单在线查询系统
- 提供七大种类信息
  - 套餐及固定费、通话、短/彩信、上网、增值业务、代收费用业务扣费、其他扣费
- 高峰时期提供千万并发用户在线查询请求

已有方案

- 使用两台IBM P5 570小型机作为数据库服务器
- 使用某关系数据库
- 只存放3个月数据
- 最多提供100查询/秒查询
- 需要限制每个用户每天查询次数来保证系统稳定服务

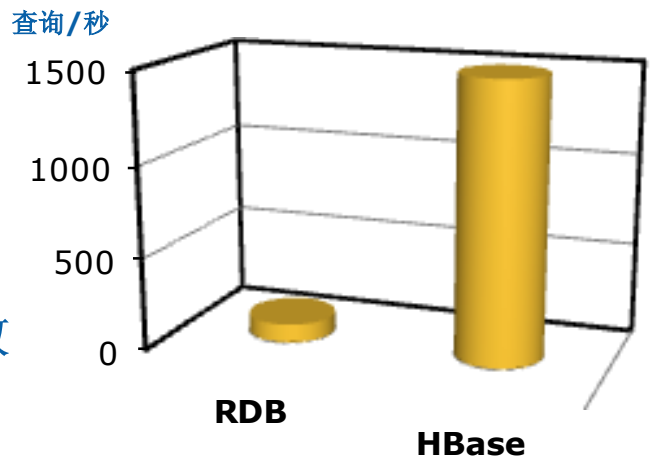
# 案例一：电信详单查询系统 - 某运营商省公司（续）

新方案数据规模

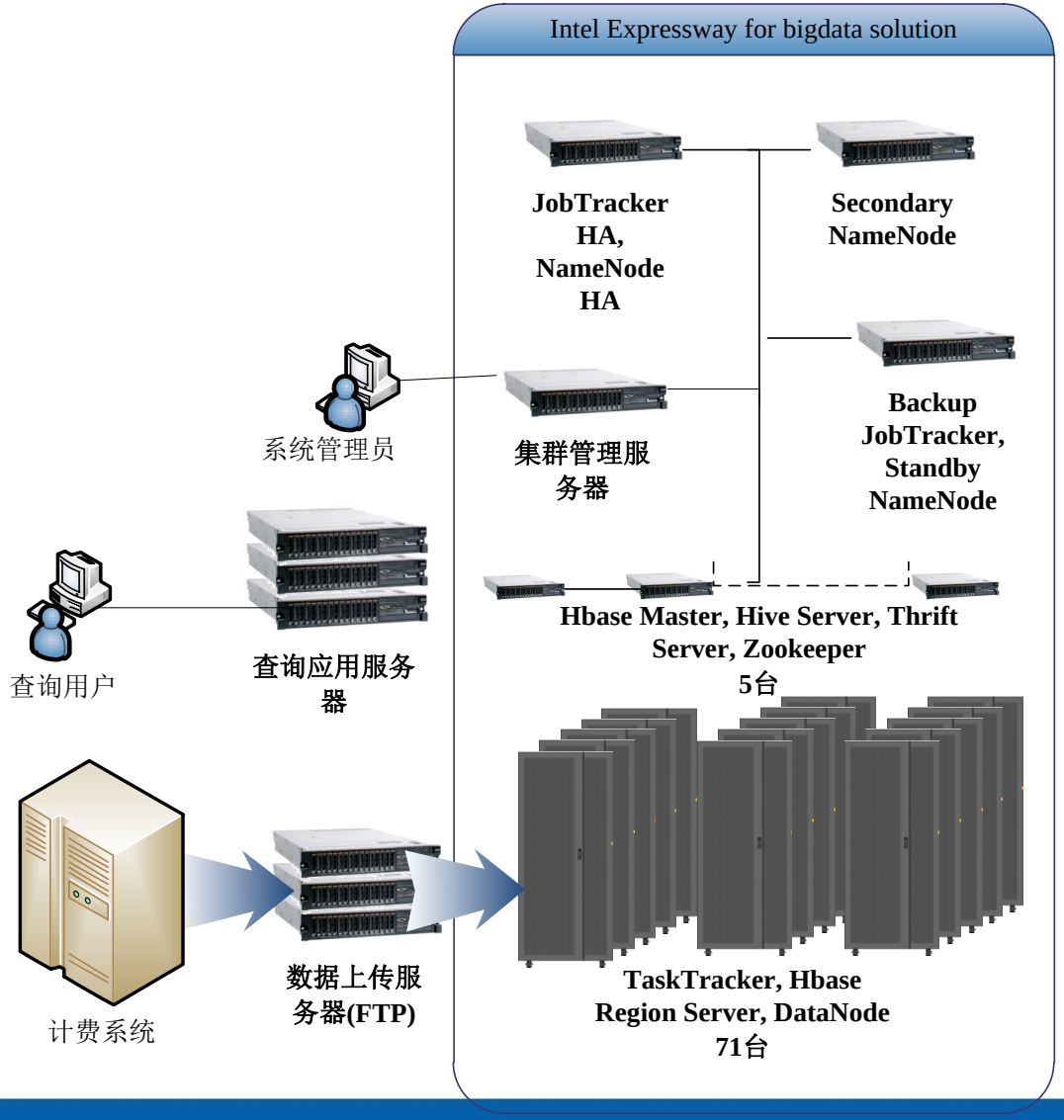
- 可容纳360TB原始数据
- 存放半年七大种类详单数据
- 平均每天2TB新增数据导入

新方案

- 构建80台双路IA服务器集群，安装英特尔Hadoop发行版构建分布式数据库集群
- 共提供400TB详单数据存储容量
- 集群提供每秒80万条详单数据插入
- 集群可以保证每秒2万条月详单查询请求，满足在线用户需要



# 案例一：电信详单查询系统 - 某运营商省公司（续）



# 案例二：电信手机上网记录及挖掘 - 某运营商研究院

业务问题

- 手机用户2G/3G上网详细信息在线查询
- 运营商上网详细信息统计分析
- 流量分析、用户行为分析等数据挖掘

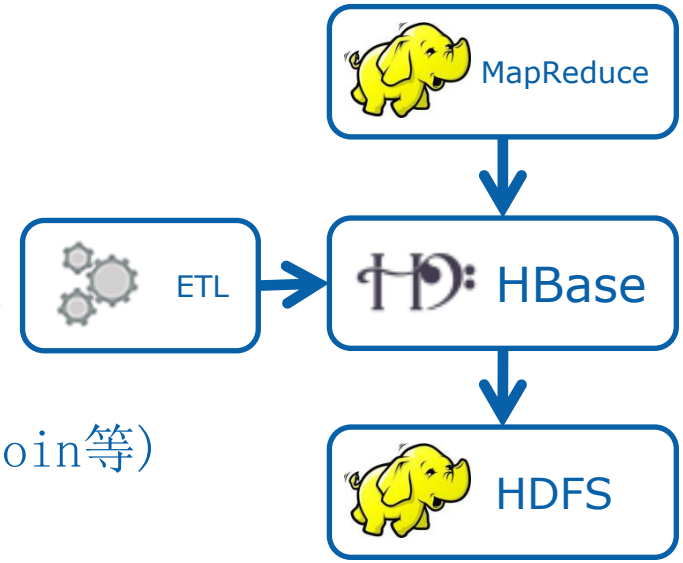
系统要求

- 原提出的基于关系型数据的解决方案不具可行性
- 预计每个省份平均一年30TB上网记录数据
- 建立全国数据中心，规划存储2PB原始数据

# 案例二：电信手机上网记录及挖掘 - 某运营商研究院

新方案

- 使用现有188台X86服务器
- 共支持2.5PB上网记录数据
- 高速数据装载，每秒导入十几万记录
- 秒级低延时查询,可开放给公众查询
- 分钟级别的统计分析(sum, count, join等)
- Map/Reduce实现数据挖掘平台




intel
Amazing things happen with Intel inside®