戴尔 激发无限

戴尔Hadoop解决方案和针对大数据而优化的PowerEdge

戴尔技术论坛

优化IT 激发无限

中国，广州 | 2013, 11.10-13

目录

何为大数据

数据世界的变化日新月异

4.3 → 每位成年人所连接的设备数

85% → 新数据类型所占的比例

10倍 → 每五年的增幅

27% → 社交媒体的使用比例

到2015年，构建现代信息管理系统的组织在财务业绩上将超出同行20%。

Gartner发布的“21世纪的信息管理”

人们正在以飞快的速度创建和使用数据...

2320亿美元

截至2016年，投入在大数据上的资金将达到2320亿美元

70%

的数据由消费者创建。但其中80%的数据都由企业负责存储和管理。

2470亿

每天发送的电子邮件数量达到2470亿封。其中80%是垃圾邮件。

440万

全球将创造440万个IT工作机会来支持大数据。只有1/3能够聘用到员工。

6000亿美元

每年因数据错误或数据质量低下而浪费6000亿美元。

48小时

每分钟就有长达48小时的视频上传到YouTube，这样每天的内容需要8年才能播放完毕。

37.5%

37.5%的大型组织表示，分析大数据是其最大的挑战。

1.8ZB

2011年使用的业务数据达到1.8ZB，比2010年增长了30%。

2亿

每日上传到Facebook的照片数量达到2亿张。这样算下来，每月将上传60亿张照片。

...带来了新的问题

为什么我们的产品更受青少年的青睐?

社交媒体活动将对产品发布带来什么影响?

下一季度，季风是否会影响我在印度尼西亚的销售以及我供应商的部件供应情况?

高级分析

社交网站和Web分析

实时数据馈送

如何捕获、分析和管理所有这些数据?

如何将这些数据转化为运营智能?

如何建立联系?

大数据在各行业中的需求

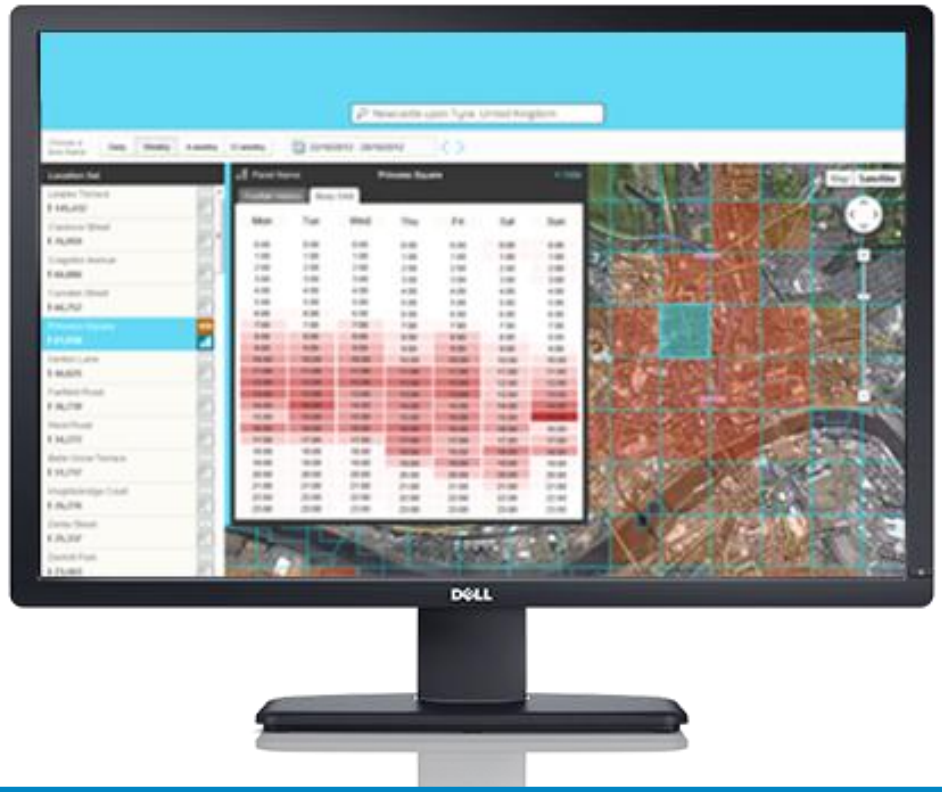
电信 | 零售 | 金融服务 | 制造 | 医疗 | 物联网 | 智慧城市

体量	GB 至 TB	TB 至 PB 以上
速度	数据量稳定，增长不快	持续实时产生数据， 年增长率超过60％
多样性	主要为结构化数据	结构化，半结构化， 非结构化，多维数据
价值	统计和报表	预测分析，机器学习， 图形算法，统计建模

出现以下情况时，应考虑采用大数据...

- 采用现有的技术堆栈执行数据分析不可行/不切实际
- 存在来自多个数据源并采用各种不同格式的相关数据
- 数据流源源不断地产生，但在捕获、存储和处理方面存在难题
- 高昂的扩展成本令人望而却步
- 大量有用的存档数据存在于磁带上（经过特定的时间后便不可恢复）
- 需要分析的数据占大多数，而不是仅占一小部分

大数据正显身手：移动用户QoS



衡量、比较并了解哪些因素影响在任意时刻访问某一位置的人数
运用分析来提升用户的服务质量

大数据正显身手：IP电视用户建议引擎

- 收集用户的点击流和观看历史记录
- 从基于Web的影片数据库添加用户元数据
- 向用户提供观看建议

点击流

EPG

VoD

大数据正显身手：金融服务

克服日益棘手且妨碍分析的数据量扩展（达到PB级）难题

通过将作业转移到设计为处理多种数据类型的技术来提升运营效率

使业务用户能够提出不同的问题来改进决策

Hadoop和大数据

- Hadoop是一种高度可扩展的开源平台，适合整合式数据存储(HDFS)和处理(MapReduce)
- 为管理“大数据”数据集和数据量而设计

是什么使Hadoop与众不同?

- 整合一切 – 所有数据都存放在同一位置并存储在同一文件系统(HDFS)中
- 擅长复杂分析 – 可跨多个节点大规模进行并行分析
- 经济实惠地进行扩展 – 安装在标准服务器上并开放源代码

Hadoop与传统数据库之对比

传统数据库 “写时创建架构”

- 必须先创建架构，然后才能加载任何数据
- 必须执行显式的加载操作才能将数据转换成数据库内部结构
- 必须先显式添加新列，然后才能将这些列的新数据加载到数据库中

Hadoop “读时创建架构”

- 数据直接复制到文件存储(HDFS)，无需进行转换
- 将数据读取到HDFS中时，在此过程中会提取所取的列
- 新数据随时都可开始流动，因为架构是在此过程中创建的

优点

1. 读取速度快
2. 符合标准和监管要求

1. 加载速度快
2. 具有灵活性和敏捷性

Hadoop/大数据使用案例

预测分析
（更大的问题）

对客户的全面了解	内容优化	建议引擎	网络分析	欺诈检测

EDW扩充	ETL卸载	批处理	数据储备库	日志处理

运营数据处理
（数据棘手问题）

戴尔：针对Hadoop/大数据推出的端到端企业解决方案

戴尔大数据解决方案的方法论……

简化客户体验

缩短投入生产所需的时间

优化解决方案性能

提供最佳的投资回报

我们如何实现?

- 与合作伙伴协作
- 结合软硬件和服务
- 参考架构和规模确定
- 更加深入的售前咨询
- 整合的售后技术支持

戴尔大数据解决方案包括

Hadoop 发行版

- 英特尔 Hadoop 发行版
- Cloudera Hadoop 发行版

戴尔Crowbar工具

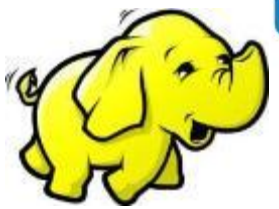
- Chef、Ganglia、Nagios、IPMI

戴尔PE-C6220、PE-C8000及PE-R720/R720XD服务器

戴尔PC-6248、Force10 S60、S4810 以太网交换机

解决方案通过以下方式提供：

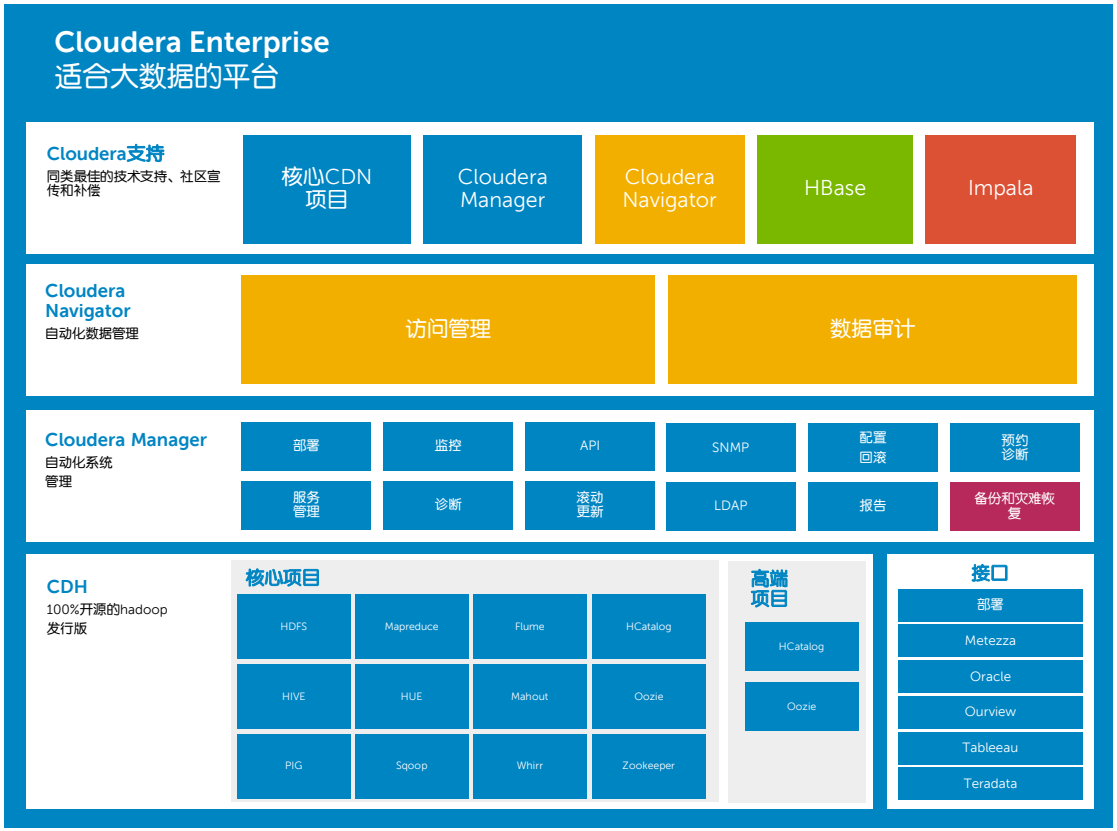
- 参考架构
- 部署向导
- 戴尔部署服务



Dell | Cloudera Hadoop
行之有效且适合企业的Hadoop发行版

更快地达成决策

- 与现有的数据仓库解决方案集成
- 采用经戴尔验证且适合您环境的PowerEdge配置
- 使用Crowbar软件框架快速部署Cloudera Hadoop
- 通过Cloudera Manager主动管理Hadoop应用程序
- 使用Cloudera Impala快速执行搜索



行业领先的商用Apache Hadoop发行版

英特尔Hadoop发行版
出众的安全性、可管理性和性能

为卓越性能而设计

- 通过英特尔扩展到HBase和Hive的功能加快事务处理性能
- 提供各项作业级指标来分析群集中部署的特定工作负载
- 利用英特尔贡献自动完成基础架构配置

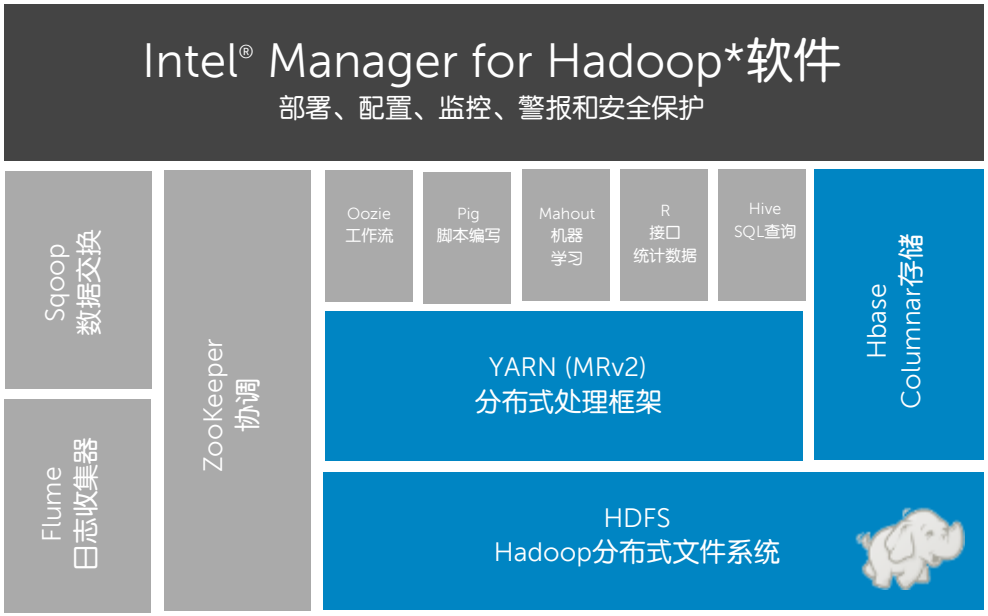
满怀信心地管理安全群集

- 配备监控、报告和警报功能
- 利用增强的加密和解密功能增强安全保护和访问控制

更快地达成决策

- 采用由支持Hadoop的业务分析软件解决方案组成、内容十分丰富的库

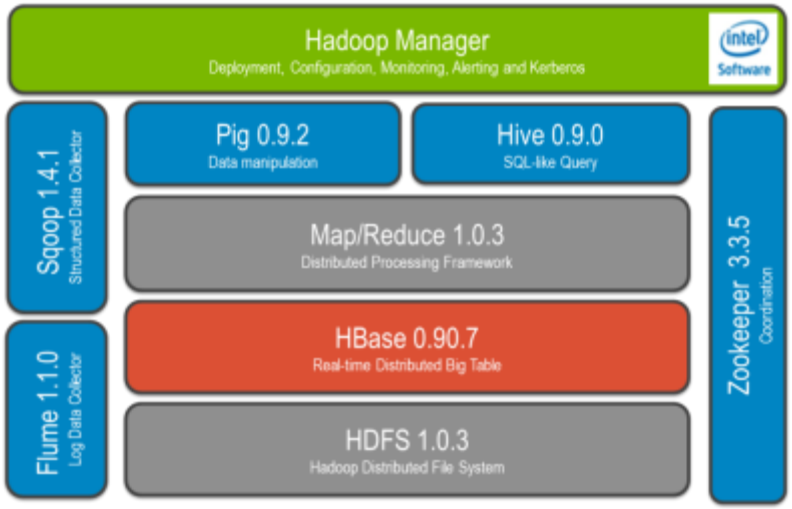
英特尔发行版的组件



提供安全且可管理的Hadoop来实现高性能

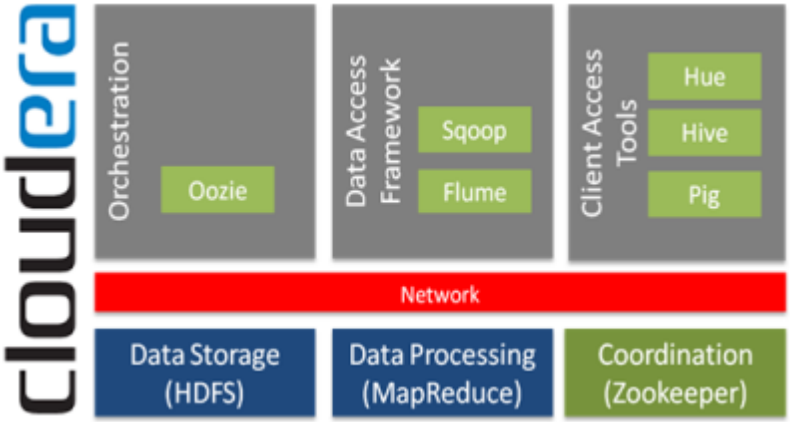
戴尔基于Hadoop的大数据解决方案

Intel Hadoop



或

Cloudera Hadoop



价值、性能和灵活性

戴尔 Crowbar - Hadoop集群部署的自动化

从“裸机”开始

- 系统BIOS更新和配置
- RAID/SAS配置
- 操作系统部署
- 部署Hadoop软件
- Hadoop软件配置
- 监控和报警的集成
- Hadoop持续运维

戴尔针对大数据优化PowerEdge配置

工作负载优化：Hadoop的工作负载千差万别

- 在工作负载优化过程中需要进行分析和基准测试
- HBase与单纯的Map/Reduce有所不同
  - I/O模式不相同
  - Hbase需要更多内存
  - Cloudera RTQ (Impala)属于I/O密集型
- Map Reduce的使用情况各异
  - 从I/O密集型到CPU密集型
- 接收和传输影响边缘（网关）节点
- 异构群集与专用群集相比呢?
  - Cloudera增加了对异构群集和节点的支持
  - 如果工作负载一致，采用专用群集是合适的
    - 主要用于“数据”业务

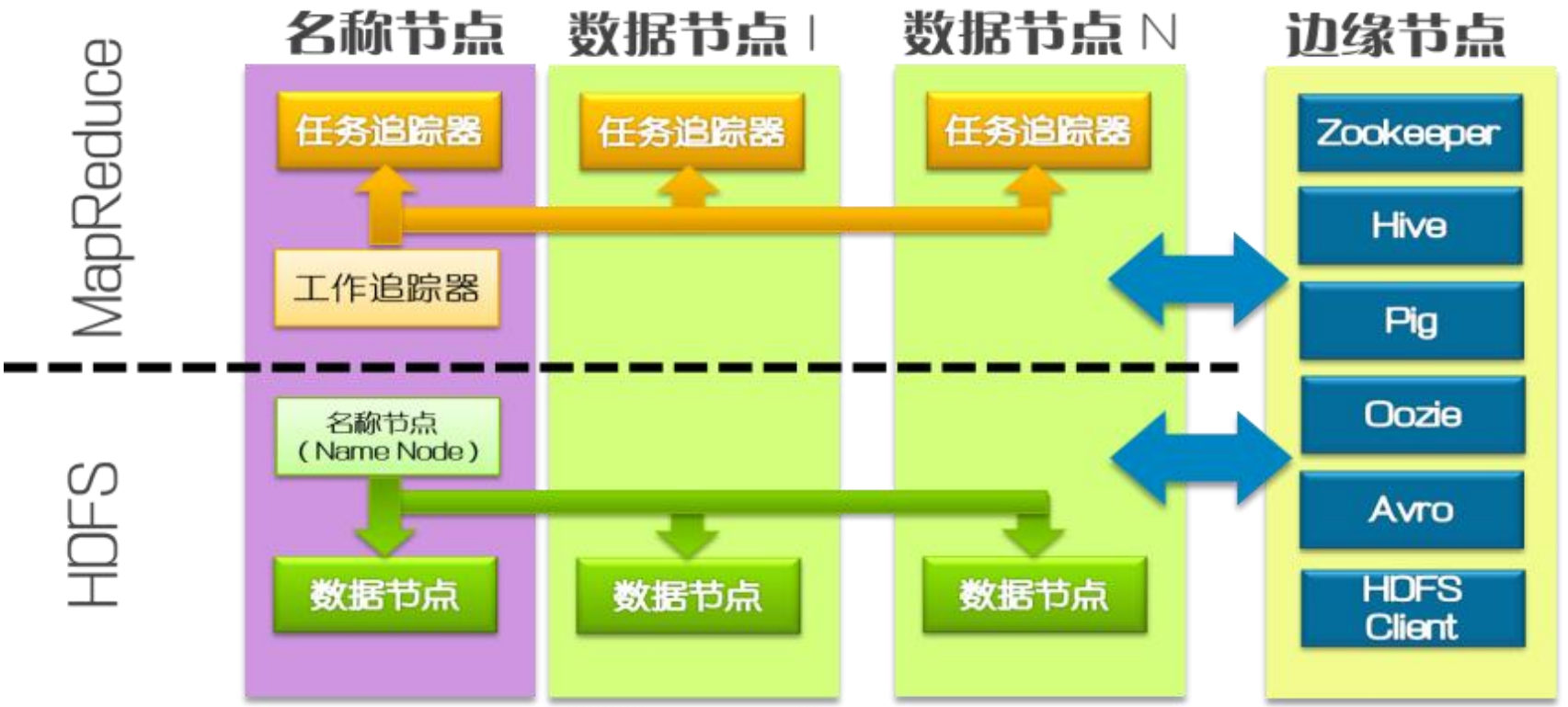
参考体系结构选项

- 高可用性
  - 网络配置
  - 主/辅助名称节点配置
- 备用交换机
  - 可行
  - 如需咨询，请与我们联系
- 群集大小
  - 参考体系结构可轻松扩展到大约720个节点
  - 此外，网络工程师还需要进行仔细的考察
- 节点大小
  - 内存建议只是个起点
  - 磁盘/内核之间的平衡从无休止

实际应用 – 戴尔客户的Hadoop配置

型号	数据节点配置	备注	RA
R720/R720 Xd	双路，16核， 最多24个2.5英寸磁盘轴	最受欢迎的Hadoop平台	✓
C8000	双路，16核， 最多48个3.5英寸磁盘轴	常用于磁盘轴/内核比较高、每TB成本较低的Hadoop应用程序	✓
C6220	双路，16核， 6个2.5英寸磁盘轴	常用于内核/磁盘轴比较高、密度较高的Hadoop应用程序。	
C2100	双路，12核， 12个3.5英寸磁盘轴	备受欢迎，硬件已停产，但常常改用于Hadoop	✓

戴尔 Hadoop 部署参考架构




R720


C8000


C6220

为Hadoop优化的服务器平台 – PEC8000

通过PowerEdge C平台，从数据存储、报告和分析系统中获得价值，提供规模、速度、丰富性和易用性

- **PowerEdge C8000 – 12代产品**
- **高密度计算配置：**
- C8000（4U机箱）+ 8 台服务器节点C8220 + 2个双冗余电源节点。
- 特点：与一般2U单节点双路机架服务器相比，计算密度为4倍，节点功耗更低。

C8000高性能计算和存储服务器

- **高密度存储配置：**
- C8000（4U机箱）+ 1 台服务器节点C8220 + 1 个内置冗余电源节点 + 4 个存储节点C8000XD。4个存储节点则共有48块3.5''硬盘。
- 特点：与一般2U12块3.5''硬盘存储服务器相比，存储密度为2倍，每TB功耗更低。

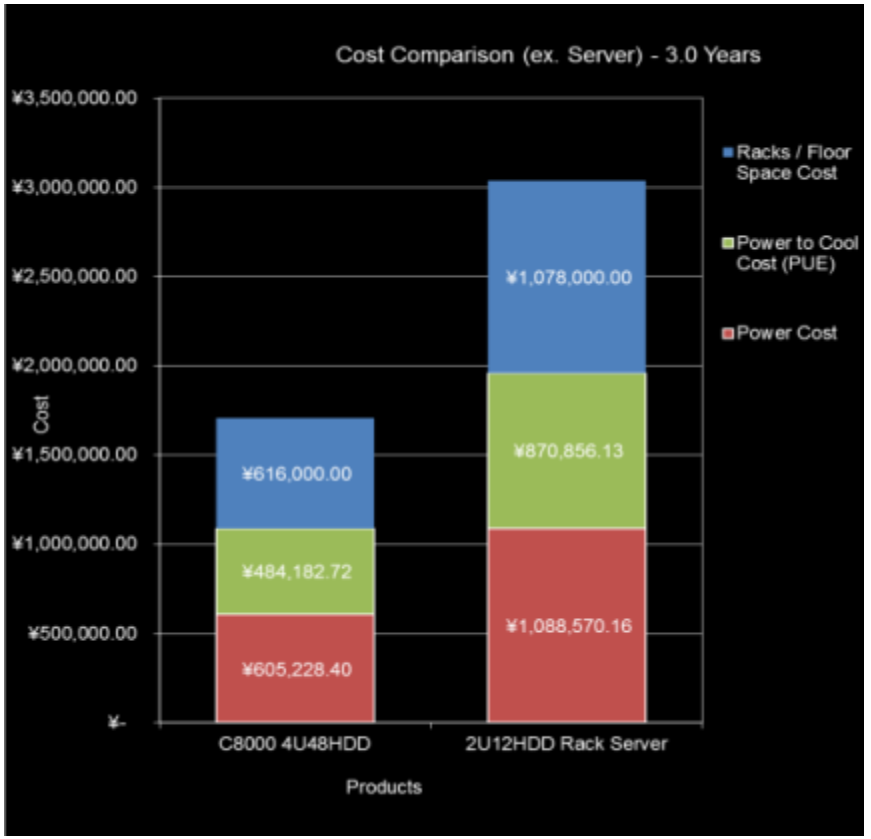
数据节点，高性价比存储型配置

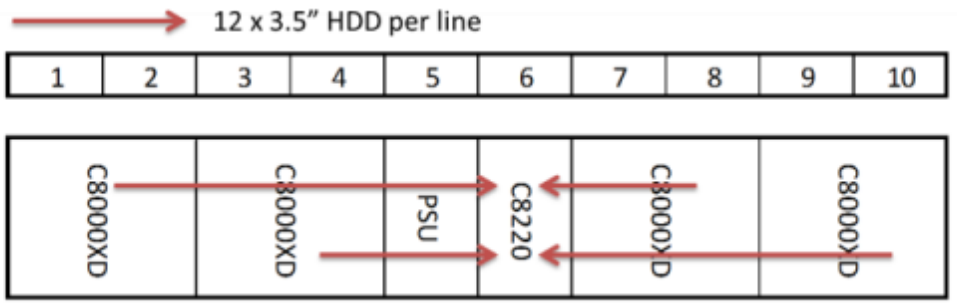
最优化的每TB成本，每数据节点48块3.5"盘

HDFS存储容量要求非常大，计算要求适中的场景

C8000 4U1Node48HDD：

- 1个C8000机箱，带1个电源装置(1400W * 2)
- 1个C8220：E5-2620 * 2 / 64G / 500G SATA 2.5英寸 * 2 / LSI9202
- 4个C8000XD：3T 3.5英寸 SATA * 12





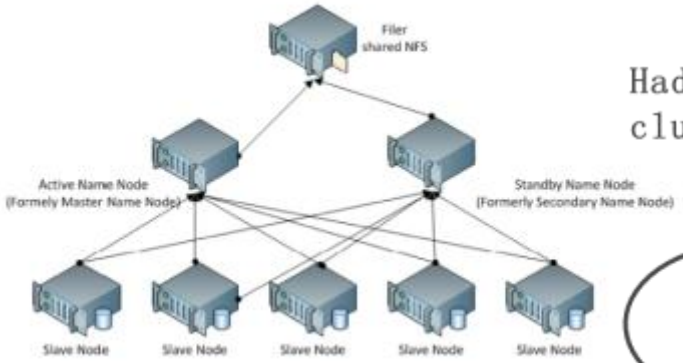
为Hadoop优化的服务器平台 – PEC6220

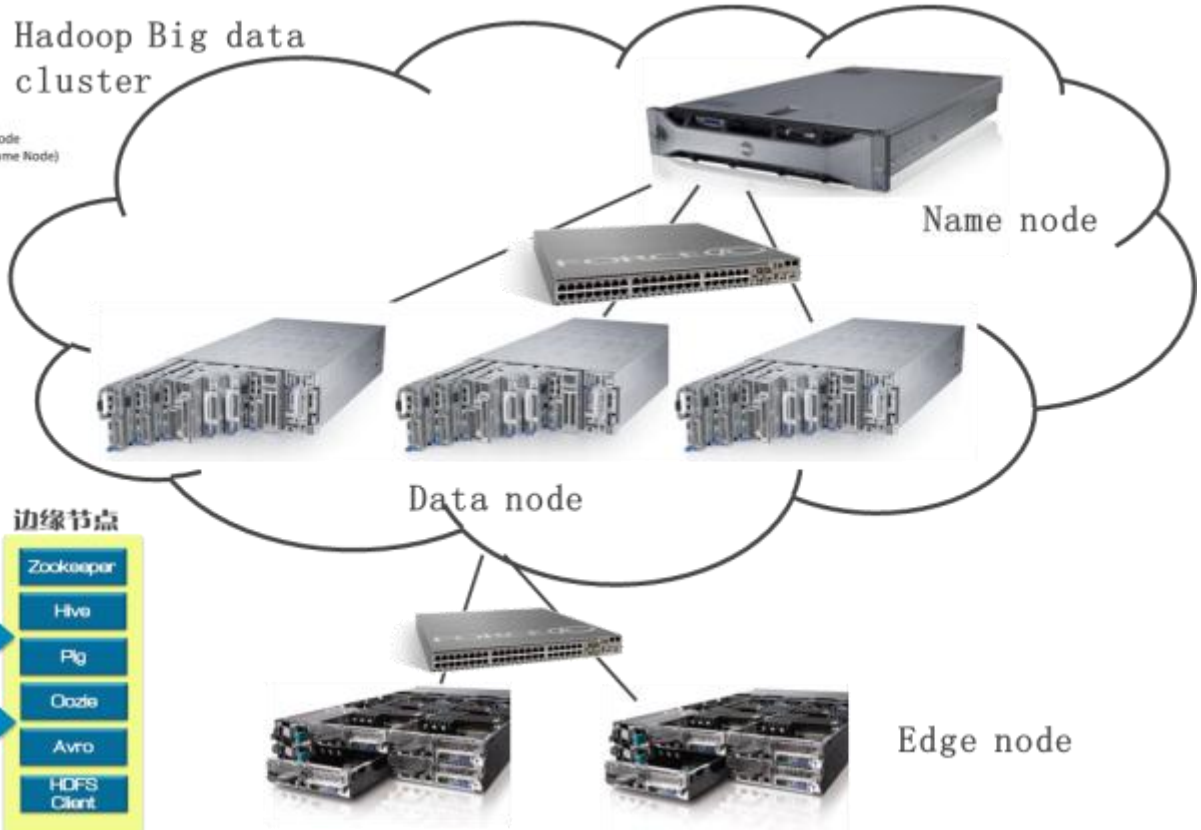
通过PowerEdge C平台，从数据存储、报告和分析系统中获得价值，提供规模、速度、丰富性和易用性

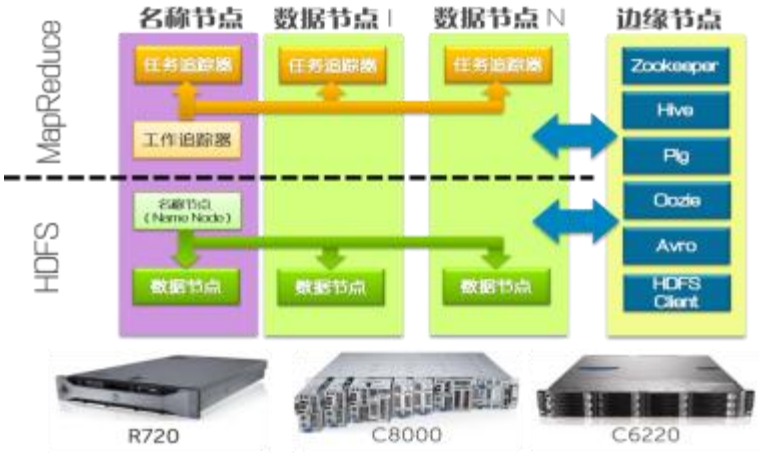
- **PowerEdge C6220 – 12代产品**
- C6000（2U机箱）+ 4 台服务器节点C6220 + 2个冗余电源。
- 2颗 Intel E5-26XX CPU，16根内存槽位，12块3.5″硬盘 或 24块2.5″硬盘
- 服务器密度为传统1U服务器的两倍，同时保留热插拔硬盘灵活性
- 所有节点都可进行独立维护，管理员可随时对任何一个节点进行停机维护，不影响其他节点的正常运行
- 区别于传统2U4节点云服务器，更提供前置硬盘灵活分配技术，支持多种类型业务搭配部署

C6220高性能和多节点服务器

Hadoop部署参考







戴尔专业服务和支持

Dell Hadoop咨询

IT咨询 | 配置和部署 | 支持

- 项目持续时间有限，但成效显著
  - 研讨会
  - 大数据评估
  - 概念验证
  - 生产实施
- 由经过认证的英特尔Hadoop专家提供
- 在每一步骤进行知识分享

通过使用Hadoop充分利用大数据的商业价值

确定在您组织的什么地方适合采用Hadoop

组建您自己的Hadoop专家团队

Hadoop解决方案
安装和实施

全包式解决方案部署服务

IT咨询 | 配置和部署 | 支持

- 从下单到客户验收的项目管理
- 采用经戴尔验证的参考体系结构
- 硬件和软件上门安装
- 在戴尔工厂进行的可选的机架集成
- 对已安装解决方案的验证
- Crowbar软件知识传授

节约资金、最大限度减少中断、优化性能。

按照戴尔最佳做法执行

快速的上线准备

Dell ProSupport

24 x **7** x **365**

全天候提供高级硬件和软件支持

- 由专家全程负责特定于解决方案的培训
- 为Crowbar提供支持
- 为Cloudera Enterprise软件提供协作支持
- 下一工作日上门服务，可选择四小时/八小时部件和人员响应
- 采用客户设定的严重性级别选项进行上报管理

在上门支持响应时间方面，戴尔服务名列第一[1]

利用我们的全球规模和技能

将停机转变为正常运行，将问题上报转变为满意度

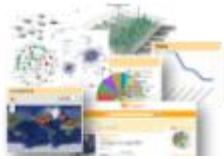
战胜复杂性，从而专注于创新

[1]Technology Business Review，“服务和支持客户满意度排行榜 | 2010年第4季度”，2011年3月，Julie Perron。戴尔服务的提供情况和条款因地区而异。有关详细信息，请访问www.dell.com/servicedescriptions。

您的完整 Hadoop/大数据解决方案提供商

- Dell Kitenga
- Datameer
- Pentaho

适用于大数据的分析软件解决方案



cloudera

一流的Hadoop合作伙伴

intel

Dell PowerEdge 第12代服务器

戴尔网络解决方案

安装和配置服务

全面的端到端实施

戴尔针对大数据的专业服务

调查 → 发现 → 规划 → 实施

客户案例

企业和大数据
它在不同行业的使用情况如何?

欧洲公共部门行政管理
- 每年带来2500亿欧元的价值
- 工作效率每年提升大约0.5%

美国医疗
- 每年带来3000亿欧元的价值
- 生产力每年提升大约0.7%

大数据在各个行业都可创造可观的财物价值

美国零售业
- 可实现的净利润增加60%以上
- 生产力每年提升大约0.5-1.0%

全球个人位置数据
- 为服务提供商创造超过1000亿美元的收入
- 为最终用户带来多达7000亿美元的价值

资料来源：麦肯锡全球研究所分析

资料来源：麦肯锡

Rapleaf

使命

- 使营销人员可以极其轻松地访问个性化客户的数据

挑战

- 旧式的关系数据库系统无法通过扩展来进行数据收集和分析

成效

- 复杂的大规模数据处理管道可使内容个性化服务实现超过 90% 的准确度
- 实现了全天候可靠性，可随时提供客户所需的数据
- 由于提升了运营效率，因此可更加智能地分配Rapleaf工程资源

Nokia

使命

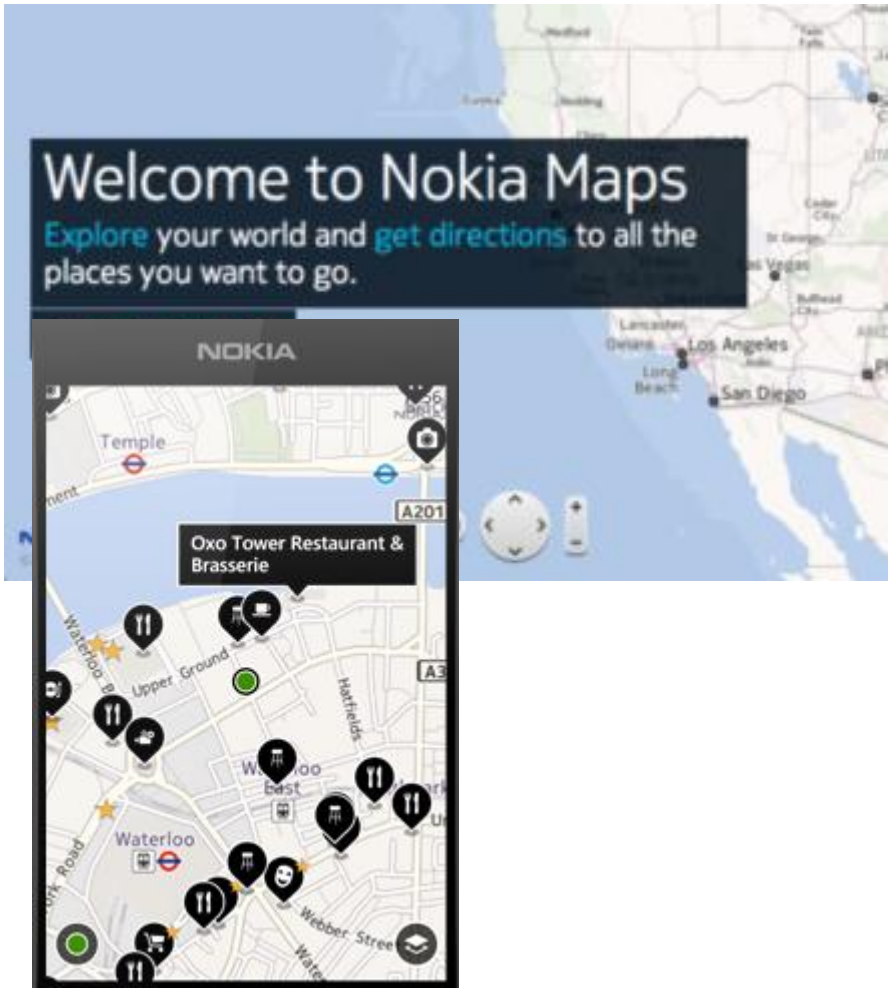
- 有效地收集和使用数据，以了解并提升用户使用手机和位置产品时的体验

挑战

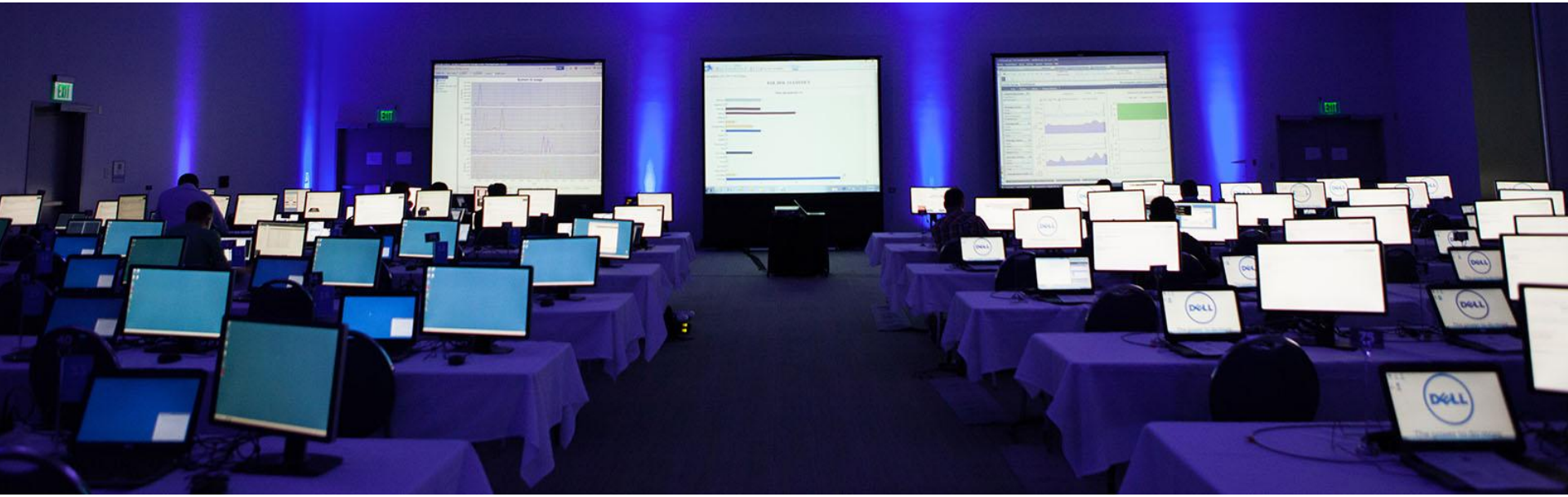
- 为支持数字地图业务而进行的数据收集需要采用专有RDBS，因而变得效率低下且成本令人难以承受

成效

- 开发出的3D数字地图引入了流量模型，从而可了解速度类别、历史流量模型、全球视频流等。
- 单个Hadoop群集现在可用作唯一的企业存储，从而取代了多个旧式系统



Dell SecureWorks



使命

全年全天候实时保护其客户资产的安全

挑战

从客户环境中收集大量数据，并对这些数据进行处理和分析

成效

- 将数据存储的成本降低到大约每GB 21美分
- 相较之前的解决方案成本降低80%
- 部署时间缩短6个月
- 不到1年收回全部投资

为何要选择戴尔的Hadoop程序

极具创新的行业领军产品：

适合大数据且可随时投入企业中使用的Hadoop解决方案

2008
戴尔DCS成立

2009
戴尔DCS交付首台大数据服务器

2011
戴尔率先推出优化的Apache Hadoop解决方案

2011
率先宣布提供Openstack支持

2011
戴尔在密度优化服务器领域排名第一

2012
戴尔推出用于部署Apache Hadoop的Crowbar

2013
英特尔和戴尔宣布支持英特尔Hadoop发行版

2008 | 2009 | 2011 | 2012 | 2013

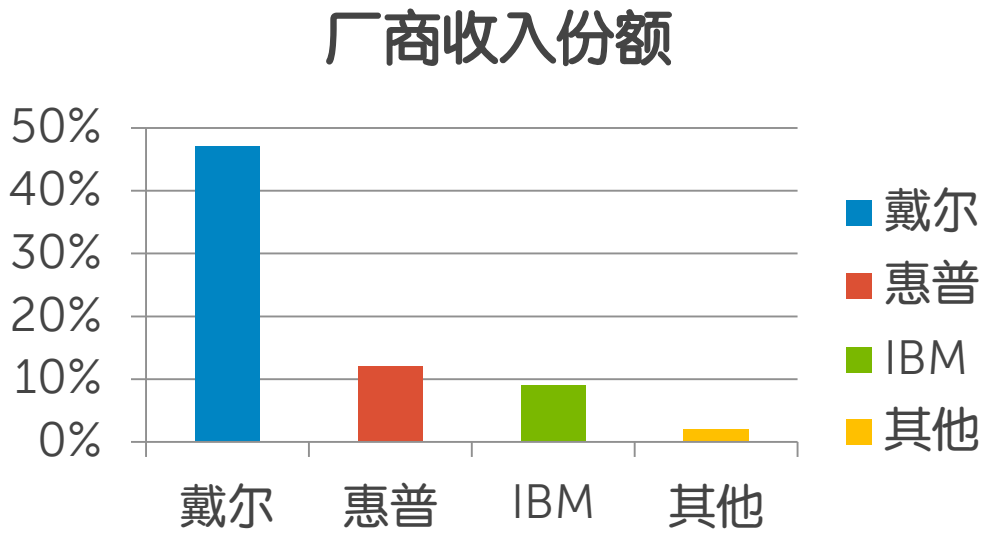
创新者 → 早期采用者 → 主流

戴尔是全球排名第一的高密度服务器提供商

并且一直以来都稳居领导者宝座

IDC
Analyze the Future

其密度优化服务器设备在全球所占的市场份额比紧随其后的三家竞争对手之和还要多



资料来源：IDC于2013年5月提供的服务器市场数据
IDC新闻稿：http://www.idc.com/getdoc.jsp?containerId=prUS24136113

戴尔大数据/Hadoop：端到端优势

使客户能够从任意规模的数据集中获得真知灼见

创建兼顾以下方面的成功业务洞见项目：数据、计算、带宽和分析

充分利用易于部署和使用的端到端业务洞见解决方案

提供相得益彰的技术和解决方案，从而可带来增量优势和价值

A ▸ **分析**和商业智能软件，可帮助获得关键的真知灼见

B ▸ 提供高**带宽**，从而可快速移动大量数据

C ▸ **计算能力**可处理和操纵信息

D ▸ **数据**存储和集成解决方案，可高效存储和管理您的数据：结构化数据、非结构化数据，等等...