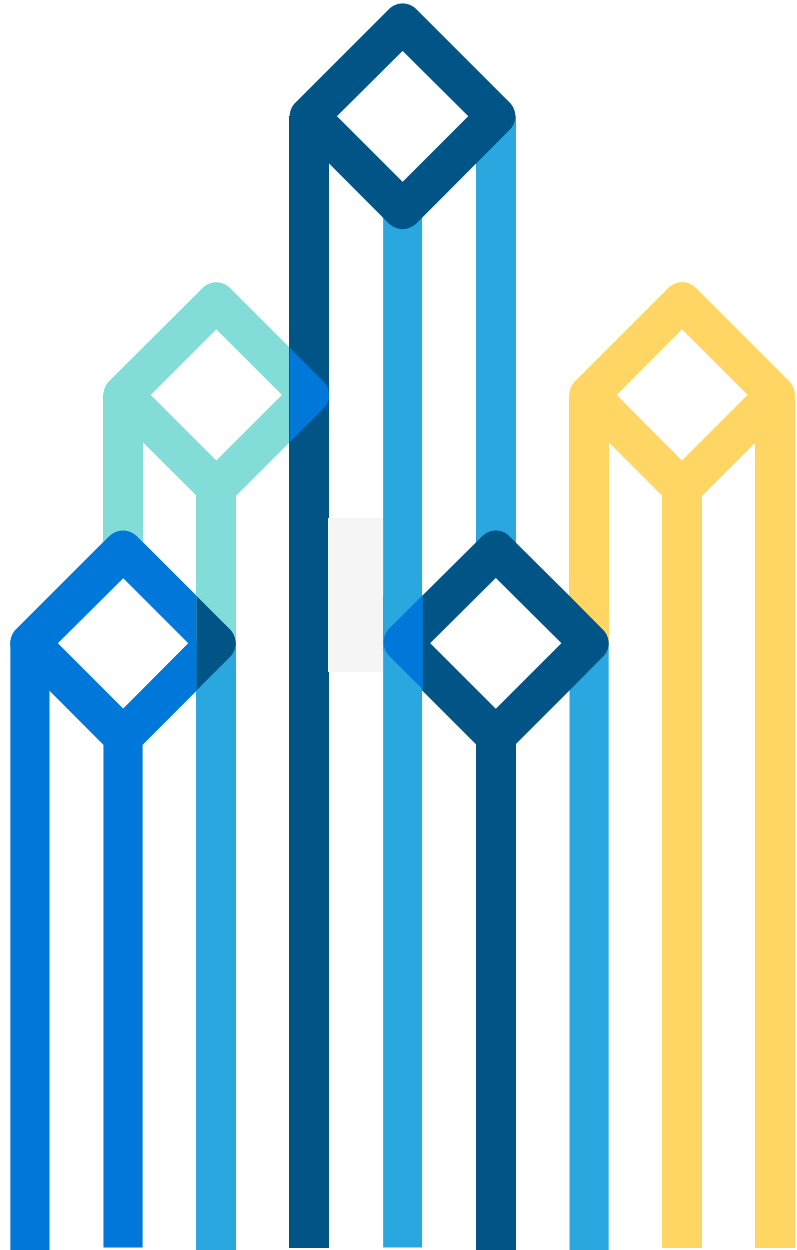
cloudera®



Cloudera Enterprise Introduction

-- From Hadoop to Enterprise Data Hub

Presenter's Name | Position

提纲

- 数据正在驱动行业的发展
- 以Hadoop为核心的大数据平台
  - 企业数据平台（Enterprise Data Hub）
- Cloudera Enterprise
  - CDH
  - Cloudera Security
  - Cloudera Manager
  - Cloudera Director
  - Cloudera Support
- 总结

cloudera

无所不在的数据

物联网及智能终端数据

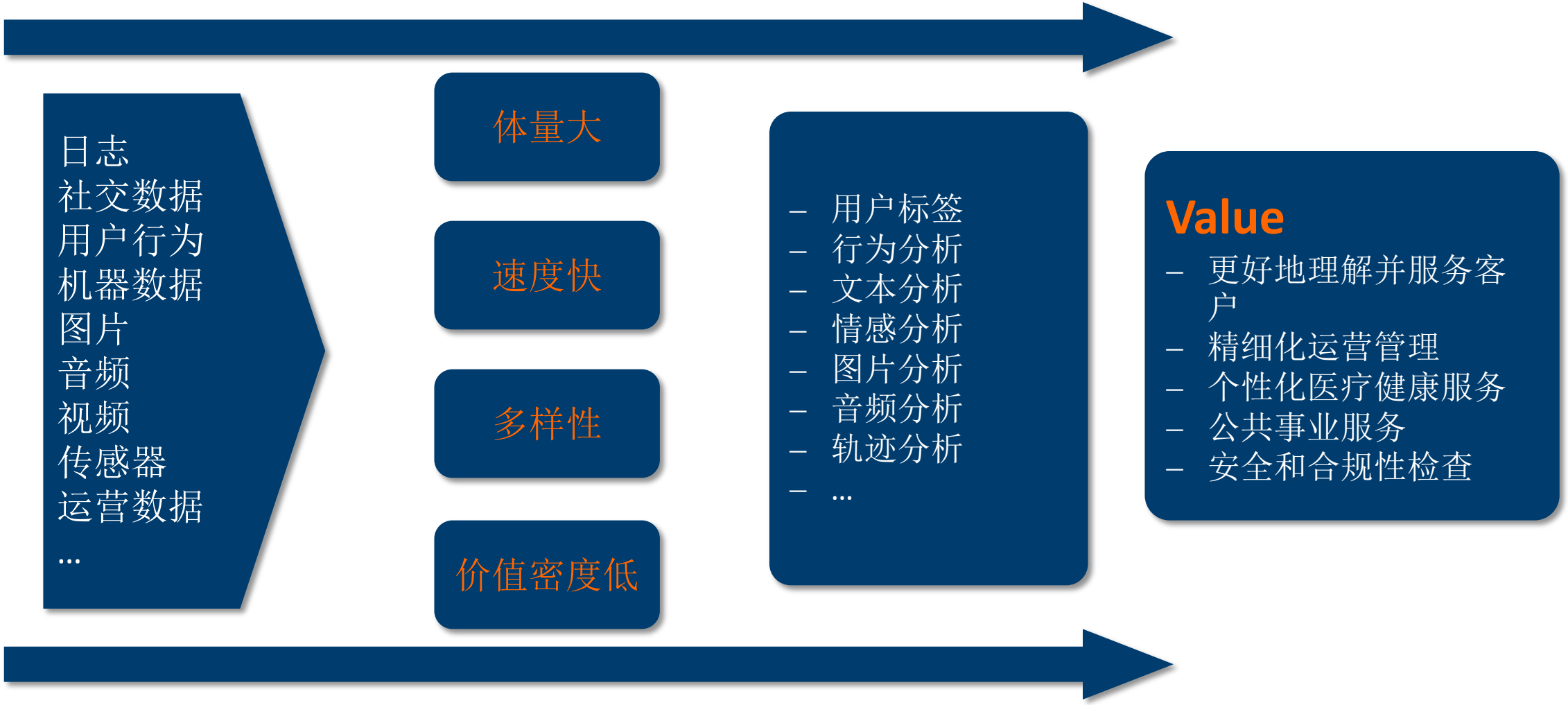
用户交互行为数据

企业运营数据

数据正成为企业的核心资产，数据可以帮助企业实现商业价值。

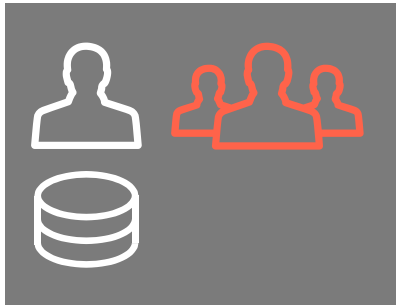
cloudera

数据价值挖掘



cloudera

传统架构的劣势



Limited Insights

Power users struggle with data.

Many users have no data.

Compliance and Privacy

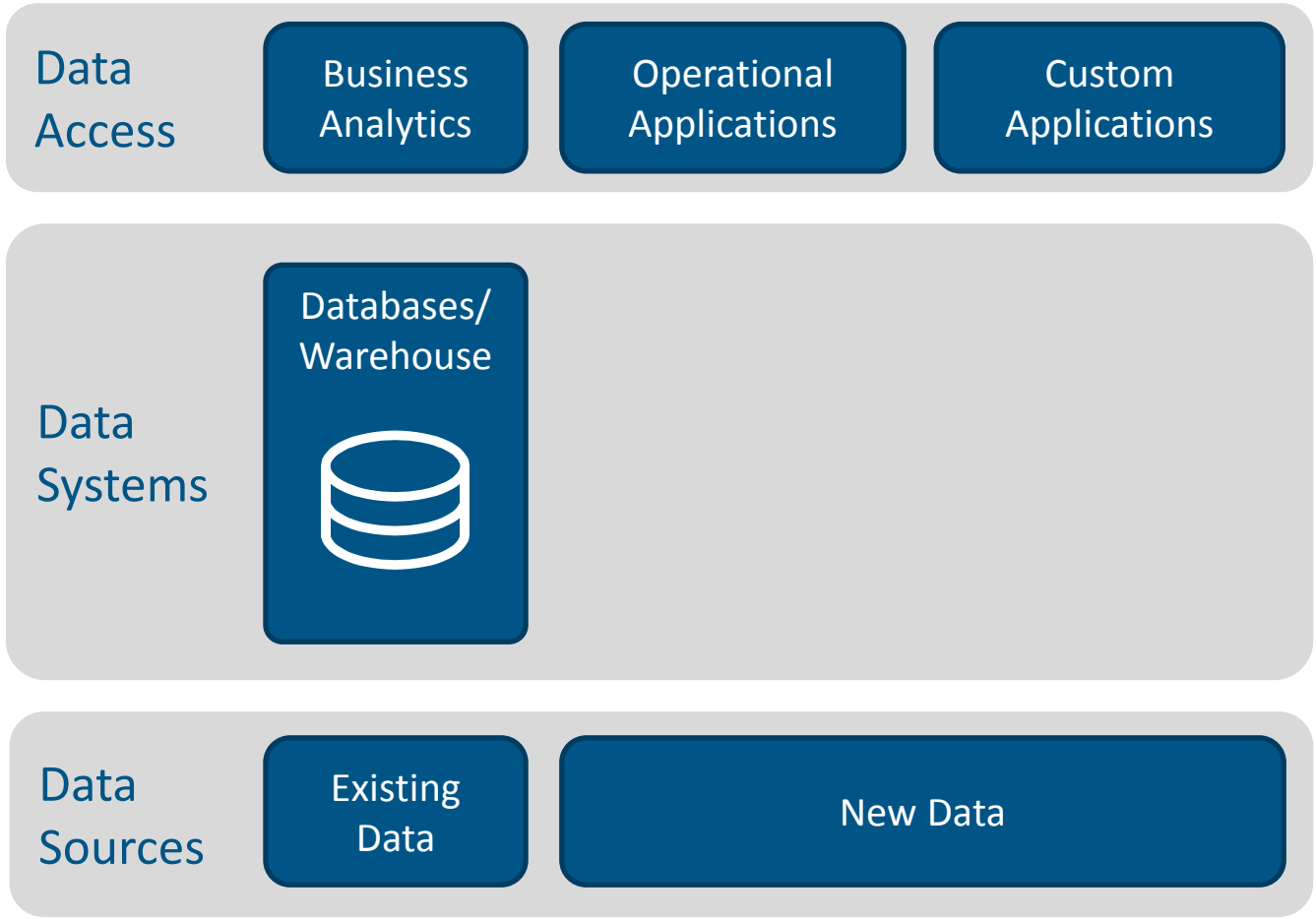
More data, more users, and more tools create complexity.

Need to balance business agility with security and governance.

Limited Data

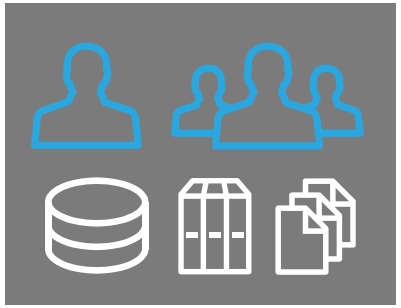
Not efficient to keep existing data, let alone handle new data sources.

Time consuming to transform data for analysis in existing systems.



cloudera

亟需新的数据平台架构



Unlock Value from Data

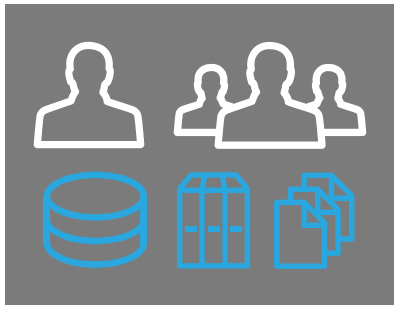
From analytics for some,

to insights for all.

Manage Compliance

From risk due to regulations and customer privacy concerns,

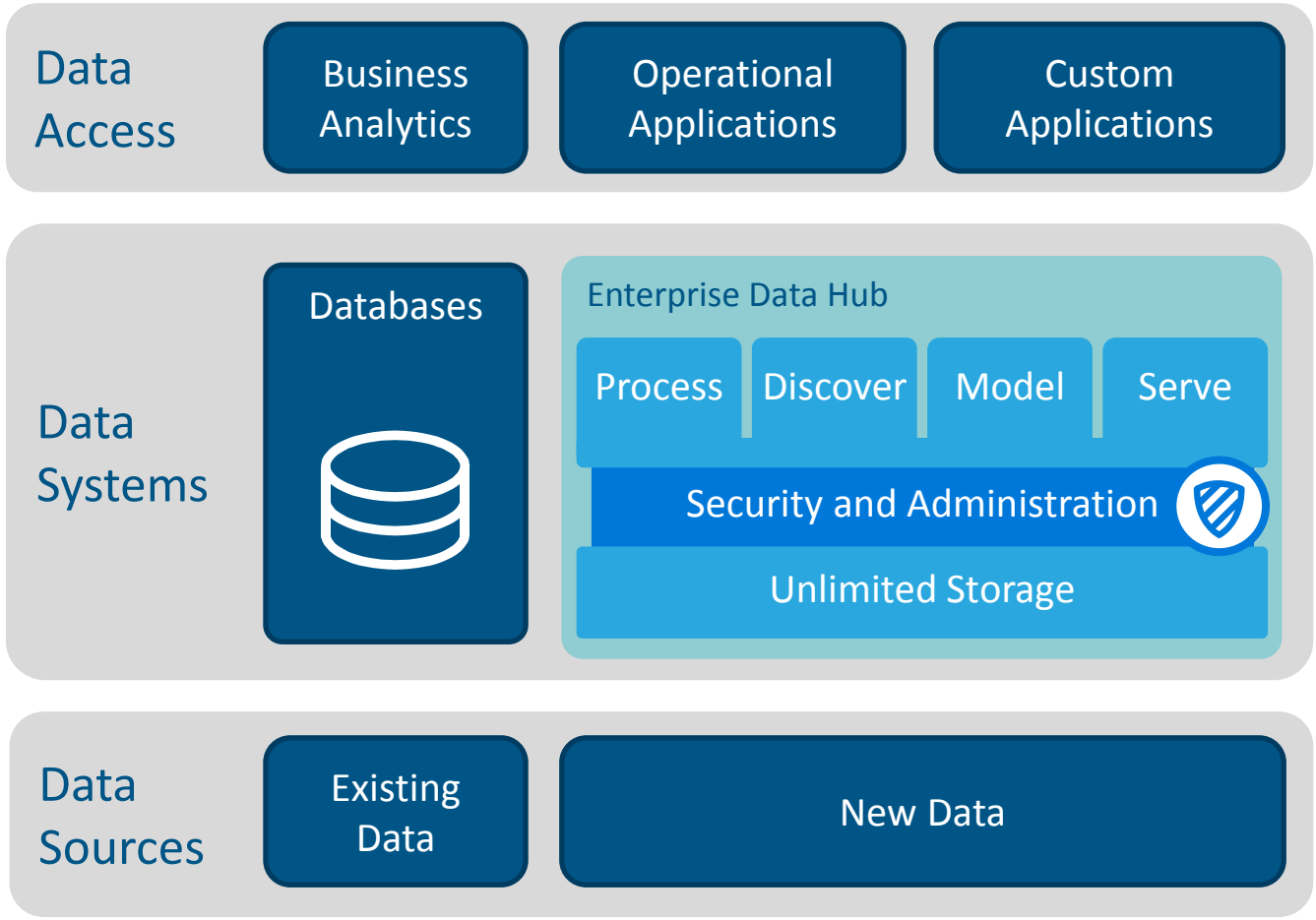
to trust in a secure and compliant platform.



Keep Unlimited Data

From disparate and limited views,

to unlimited information access.



cloudera

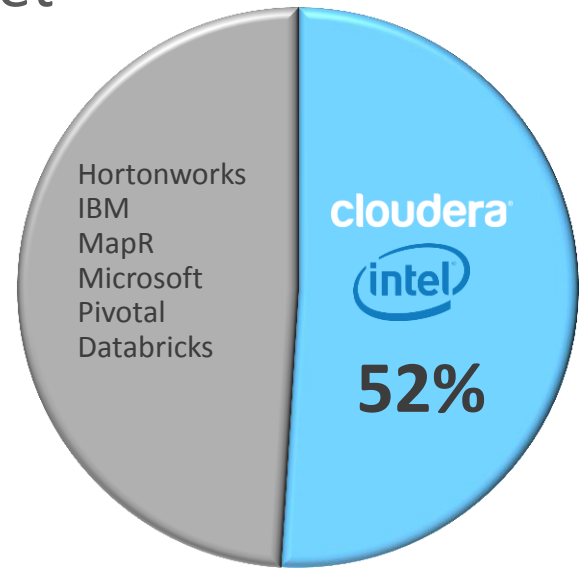
Cloudera

创立　　成立于2008，企业级Hadoop产品提供商
员工数量　　超过900名
全球支持　　24x7 全球支持
　　创新的主动支持和预测支持项目
客户群　　全行业客户(金融、电信、零售、能源、互联网、媒体等）
　　各行业的顶尖企业都有Cloudera Enterprise部署
强大的产业链　数百个生态链合作伙伴; Cloudera Connect Program (CCP)
培训和认证　超过80,000管理员、开发者等受训；最有价值的大数据证书
开源领导者　Hadoop及其相关生态项目的绝对领导者，和Intel合作加速
　　企业数据中心的革新
Cloudera中国　2014年9月成立，上海是大中华区总部，负责产品培训、
　　专业技术服务和产品支持，在北京和广州有本地支持

cloudera

Cloudera和Hadoop生态

- Cloudera是Hadoop项目的最大贡献者，同时也是No.1的Hadoop发行版提供商
- Hadoop平台标准化的领导者
  - 数据采集 – Flume, Sqoop
  - 数据存储 – HDFS, HBase, Avro, Parquet
  - 数据处理 – MapReduce, Spark, Hive
  - 数据分析 – Impala, Solr



JIRA Contributions, Source: ApacheJIRA
January 2012 – February 2015

Projects Included:

Accumulo
Avro
Bigtop
Crunch
Flume
Hadoop Core
HBase
Hive
Kafka
Mahout
Oozie
Pig
Solr
Spark
Sqoop
Tez
ZooKeeper

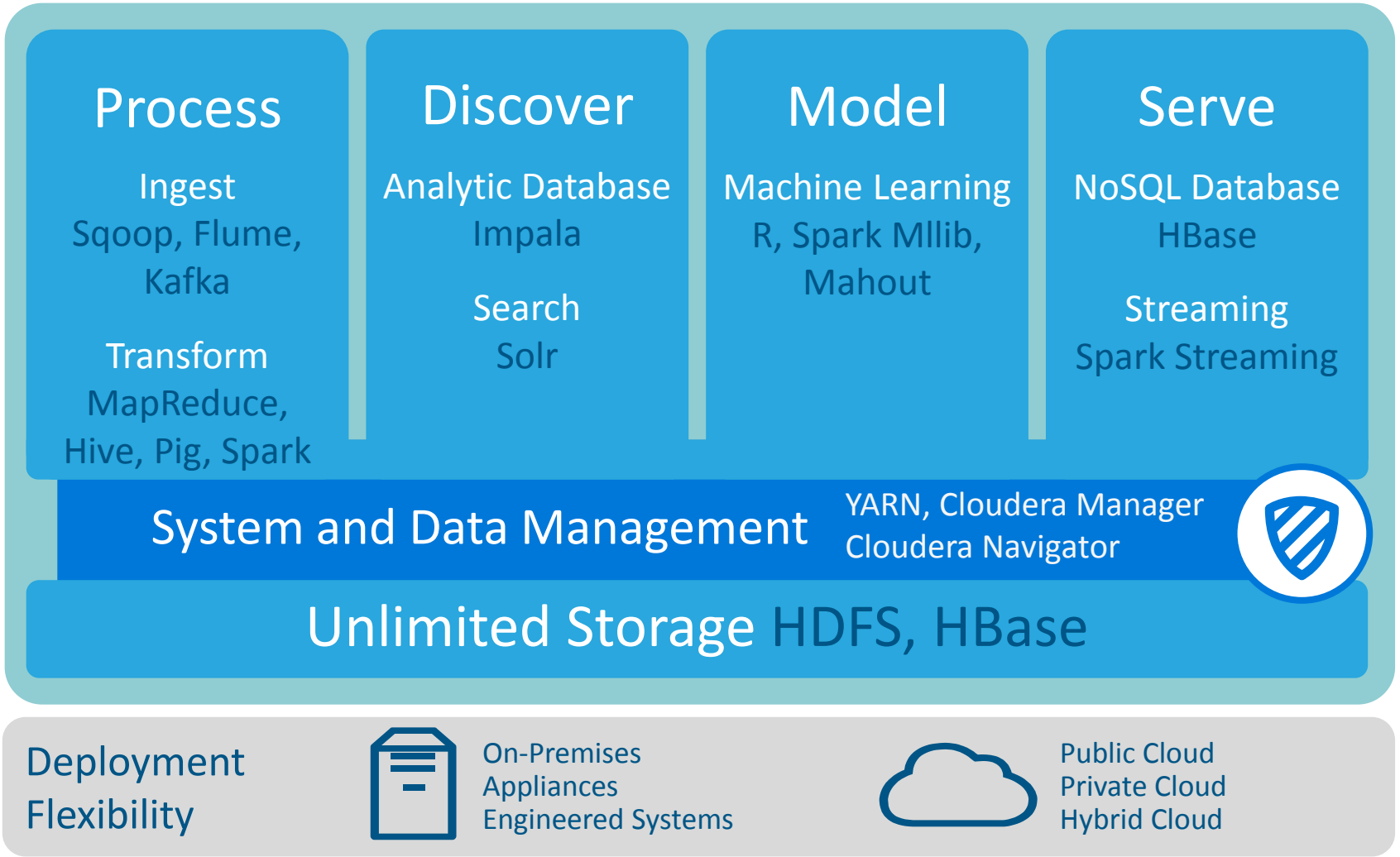
cloudera

Cloudera产品和服务

- Cloudera Enterprise
  - Cloudera提供了100%开源的，开放标准的Apache Hadoop发行版（CDH）
  - 让Hadoop真正进入企业级应用的Cloudera Manager和Cloudera Navigator
  - 提供虚拟化和云化大数据方案的Cloudera Director
- 业内最权威的Hadoop技能培训和认证
- 深耕于开源社区的专业技术支持团队和产品支持团队

cloudera

Cloudera Enterprise



完善的企业安全策略

- 身份认证，授权，审计，数据安全
- 数据可管理性

开放标准

- 100%开源Hadoop及其相关组件
- 3rd标准的软件集成
- 开放API
- 标准云服务集成

统一平台

- 数据导入导出
- 可扩展存储
- 多样化的处理引擎
- 安全
- 资源管理
- 元数据管理

cloudera

企业级数据平台

- 稳定性
  - 严格的测试
  - 被客户和开发者证明
  - 开源的模式
- 易用性
  - 标准的API（Java, SQL, Python, Rest）
  - 标准的工具集成（MS, Qlikview, Tableau, Teradata, Netezza, Quest…）
  - 一站式管理解决方案
- 安全性
  - 企业安全标准集成
  - 统一的访问安全控制
  - 全面的数据保护，密钥管理
- 可管理性
  - 部署、管理、监控、警告
- 可治理性
  - 数据溯源
  - 数据发现
  - 数据生命周期管理
- 灵活性
  - 不同的问题可以有不同的技术选择
- 性能
  - 高吞吐的NoSQL存储
  - 原生的大规模数据处理引擎
  - 内存计算
  - 为X86平台做的原生优化

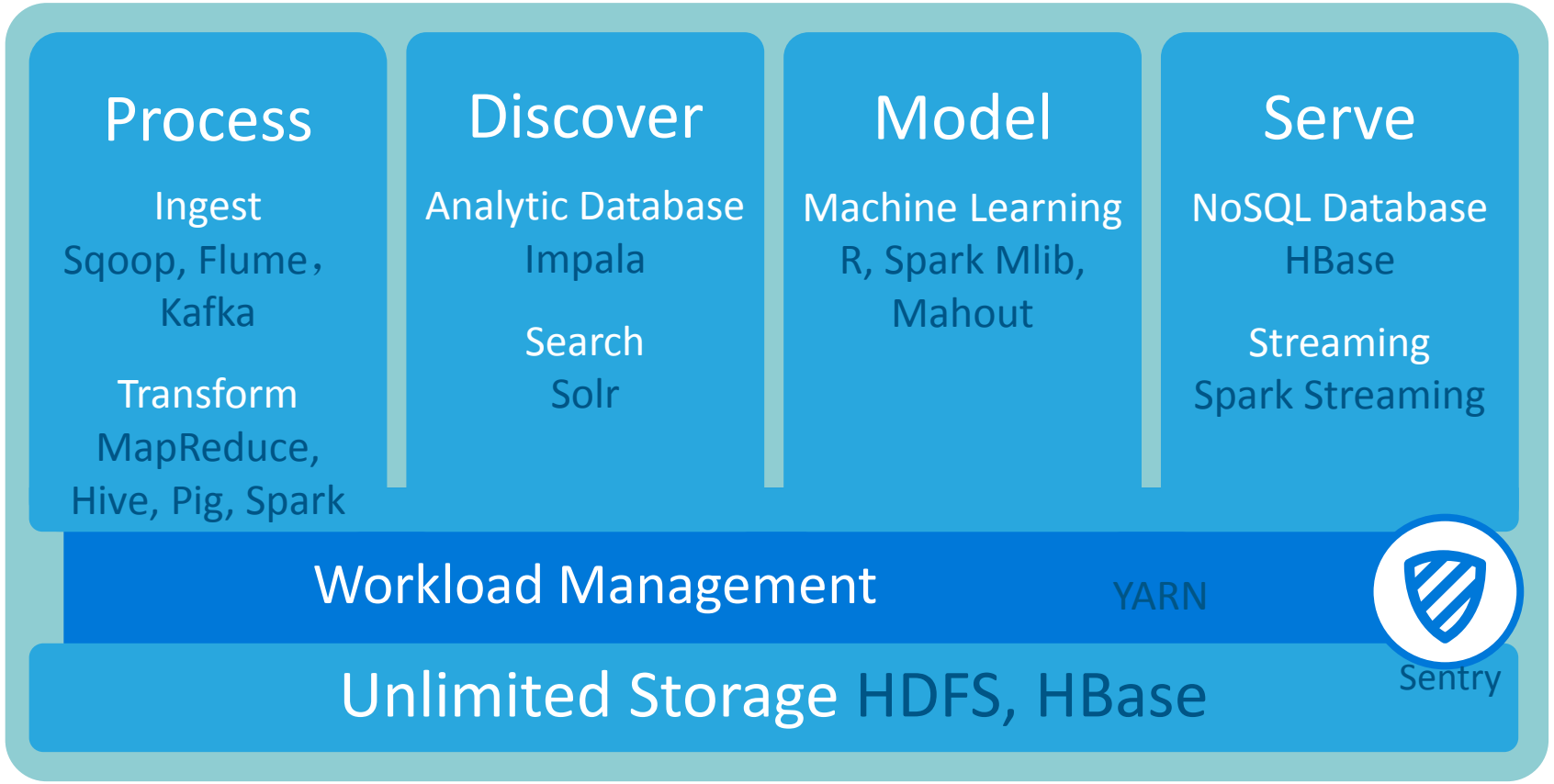
cloudera

The Open Source Platform - CDH

cloudera

最具创新的开源核心

CDH – Cloudera Distribution for Apache Hadoop



- 100%开源且开放标准的Hadoop核心
  - 数据采集
  - 多样化的可扩展存储
  - 资源（负载）管理框架
  - 灵活多样的处理引擎
  - 全面的安全技术体系
  - 易用的Hadoop交互界面

cloudera

CDH

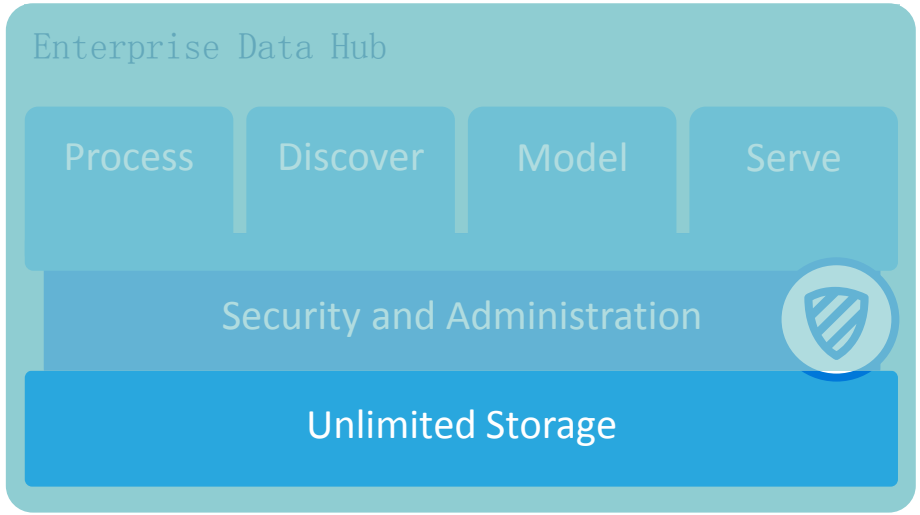
- CDH
  - 全球最流行的Hadoop发行版
  - 最完整且稳定的版本，经过严格的行业检验
  - 具有最快的更新，更多新的功能
  - 方便开发者和集成商使用Hadoop
- 和其他一些Hadoop发行版提供商对比
  - Cloudera做Hadoop开发的，其他厂商仅是做Hadoop集成或CDH集成
  - 和Hadoop trunk最快的同步，能保证业务的前向兼容性；其他厂商在Hadoop上做的定制优化或修复，无法保证兼容性
  - 所有组件的开发和专业支持能力，其他厂商也仅仅跟随Cloudera包含的版本进行集成，缺乏问题修复和专业支持能力

HDFS

分布式文件系统

灵活性	多样化数据的统一存储
可扩展性	良好的线性可扩展性
高容错性	设计之初就考虑了高容错性
开放性	存储的数据格式和内容完全可见

适合大文件的顺序读写，写一次读多次



cloudera

Apache HBase

构建在分布式存储上的NoSQL数据库

具有分布式存储的所有优点

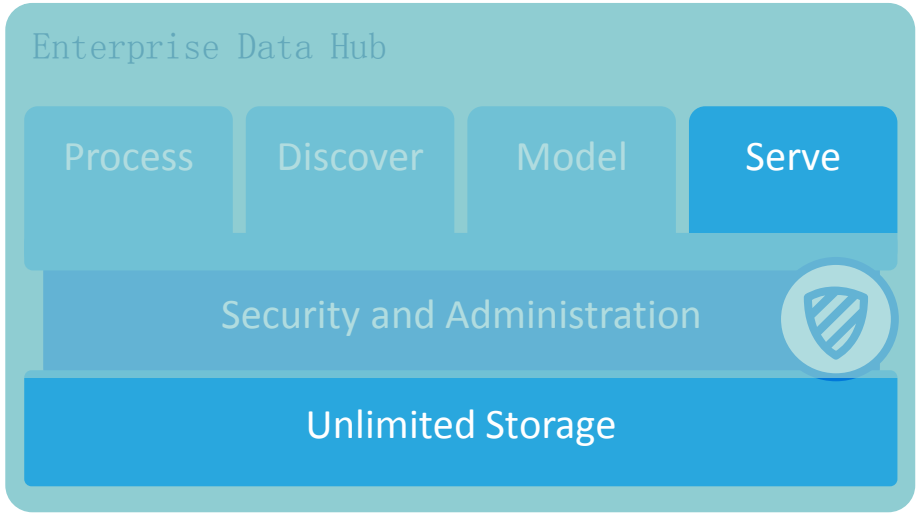
灵活性　　多样化数据的统一存储

可扩展性　　良好的线性可扩展性

开放性　　存储的数据格式和内容完全可见

+ 在线数据服务

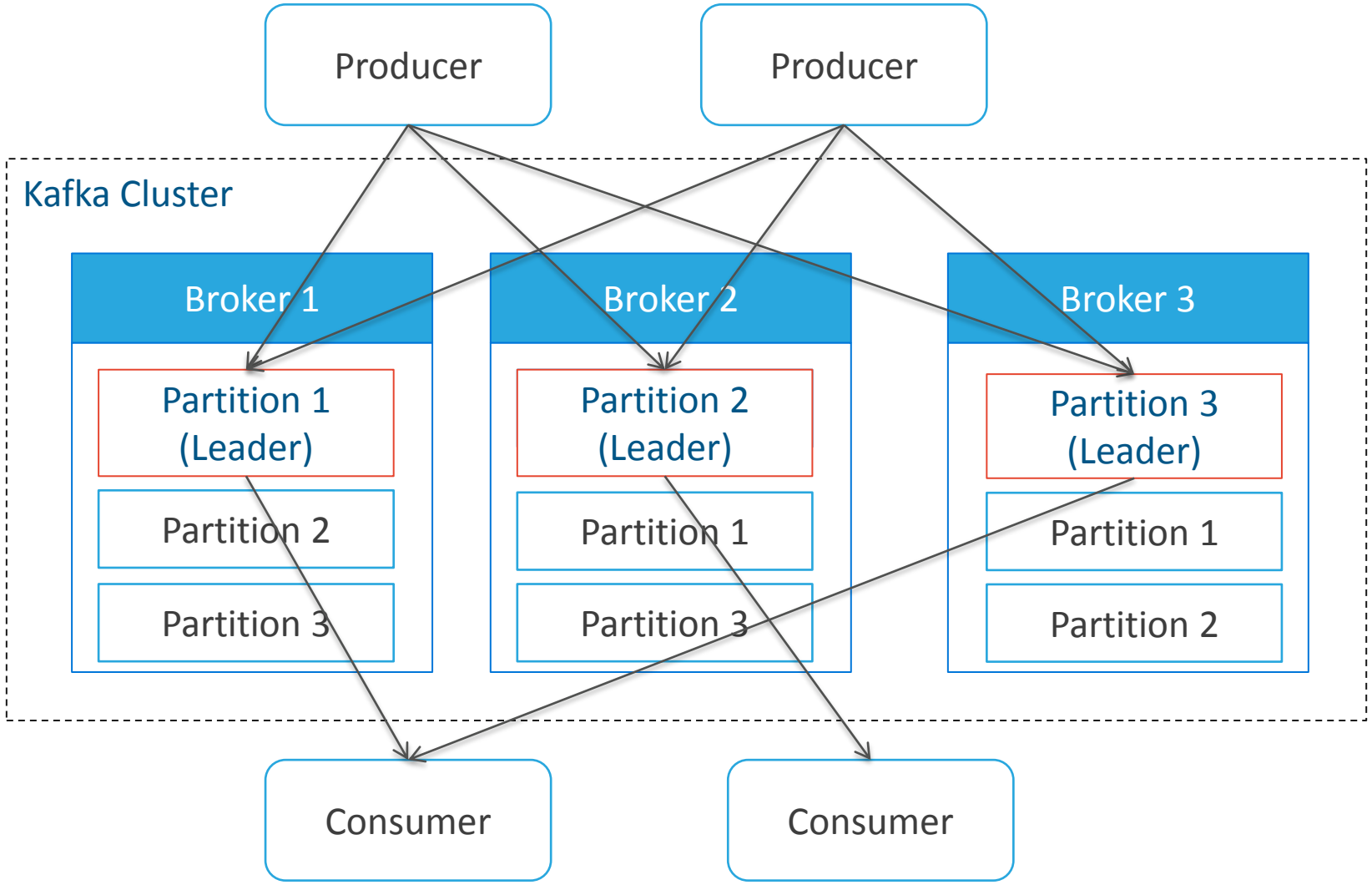
和HDFS紧密结合，适合高并发随机读写



cloudera

Apache Kafka

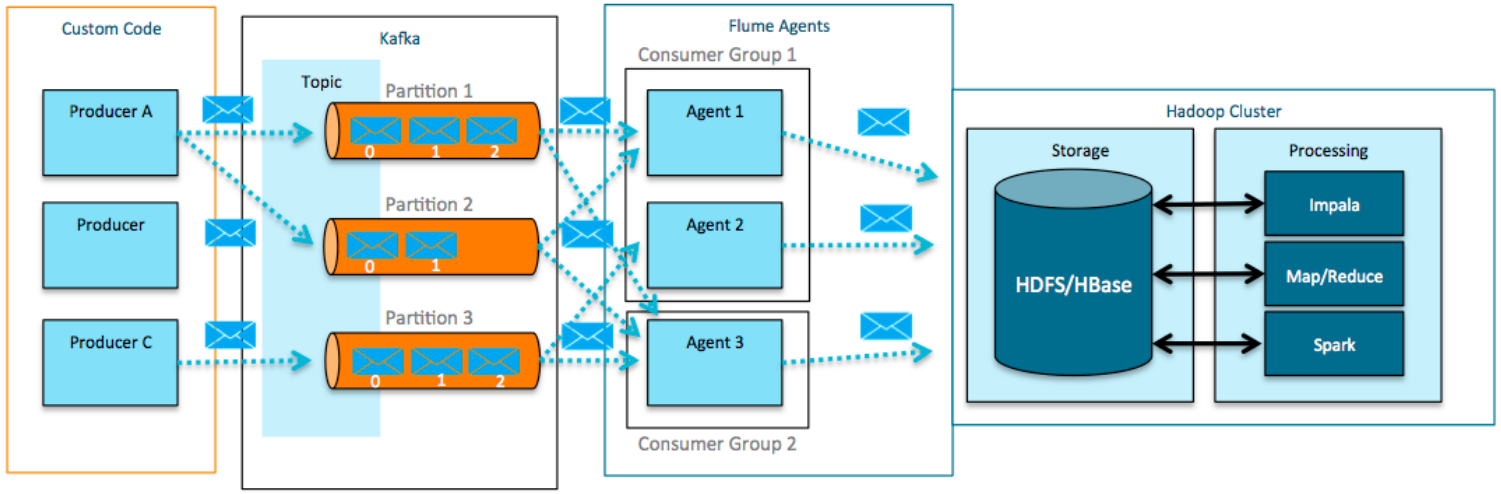
- 每个节点称为 Broker
- 数据以 Topics方式写入Kafka
- 每一个Topic都可以被分片
- 分片分布在Broker上
- 分片可以有多个副本，其中一个为Leader
- Producer, Consumer都与partition直接进行数据交换



cloudera

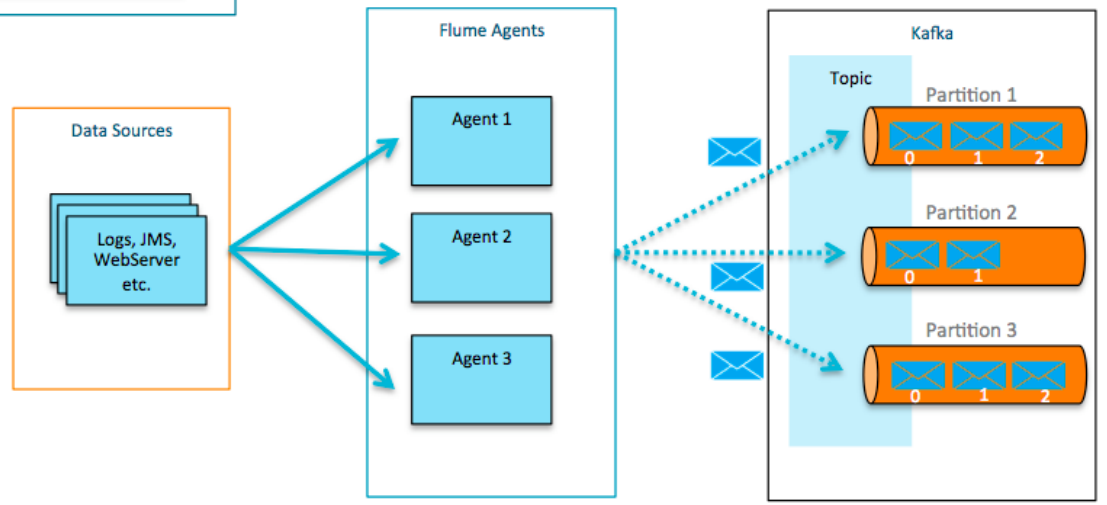
Kafka + Apache Flume

- Kafka 可以被配置为 Flume 的Channel
- Flume Sources 和 Sinks 可以配置成Kafka的Consumer和Producer



Flume Sources Consume from Kafka:
Write data to HDFS, HBase, or Search

Flume Sinks Write to Kafka:
Read from logs, files, jms, http, rpc, thrift, etc and write events to Kafka



cloudera

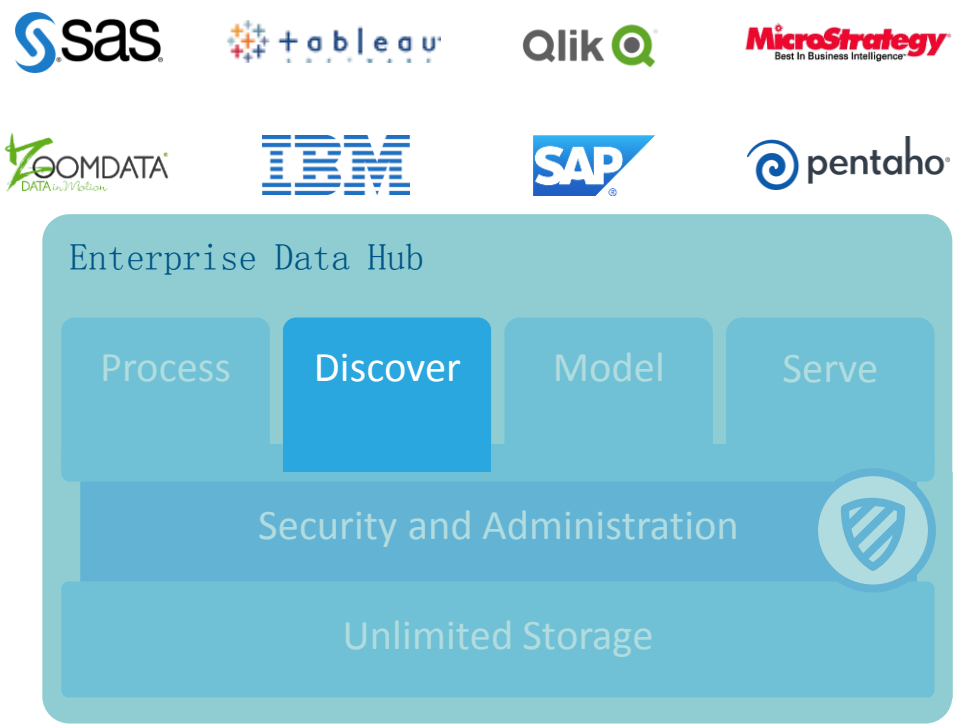
多样的工作引擎

- 批处理引擎（MapReduce, Hive, Spark） - 适合长时间的数据处理作业，高度成熟可靠
- 实时数据处理（Spark Streaming）- 实时的数据同时，异常检测，预测分析等
- 自助BI分析/交互式SQL （Impala）– 准实时的分析作业，高效的数据探索式分析，高并发的自助BI功能
- 搜索（Cloudera Search） - 快速的跨应用数据搜索能力
- 数据挖掘（Spark Mllib, R, Mahout）- 适合数据分析人员的快速模型创建，迭代
- 在线服务（HBase）- 提供实时的数据服务能力

交互式分析引擎Impala

构建于HDFS上的原生的分析型SQL

易用性	利用现有的SQL语法，和绝大多数BI工具完美集成
高并发	为高并发的随机分析而优化，用C++编写
交互性	提供交互式的体验
原生	和Hadoop栈深度融合



cloudera

Apache Spark

适合数据科学家的分布式内存计算引擎

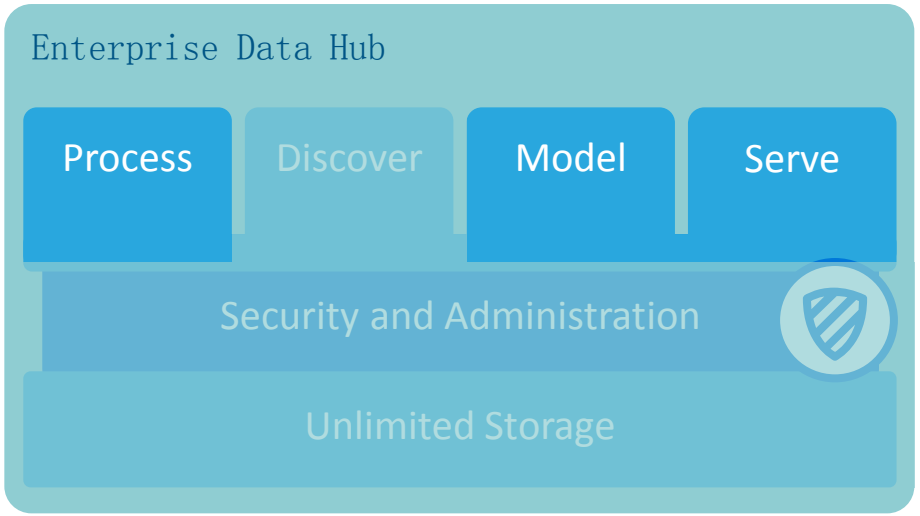
灵活　　多种接口，多种算法

高效　　内存计算，适合迭代是计算

易用　　好用且丰富的API

安全集成　　和数据平台的其他功能无缝集成

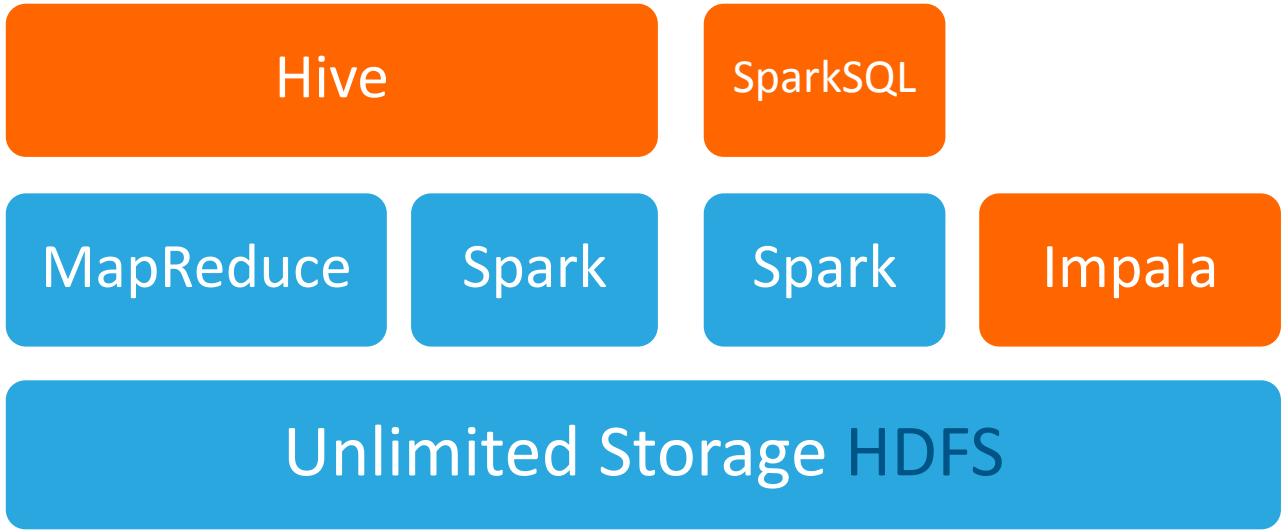
适合批处理、流计算以及迭代式计算



cloudera

SQL的实现途径

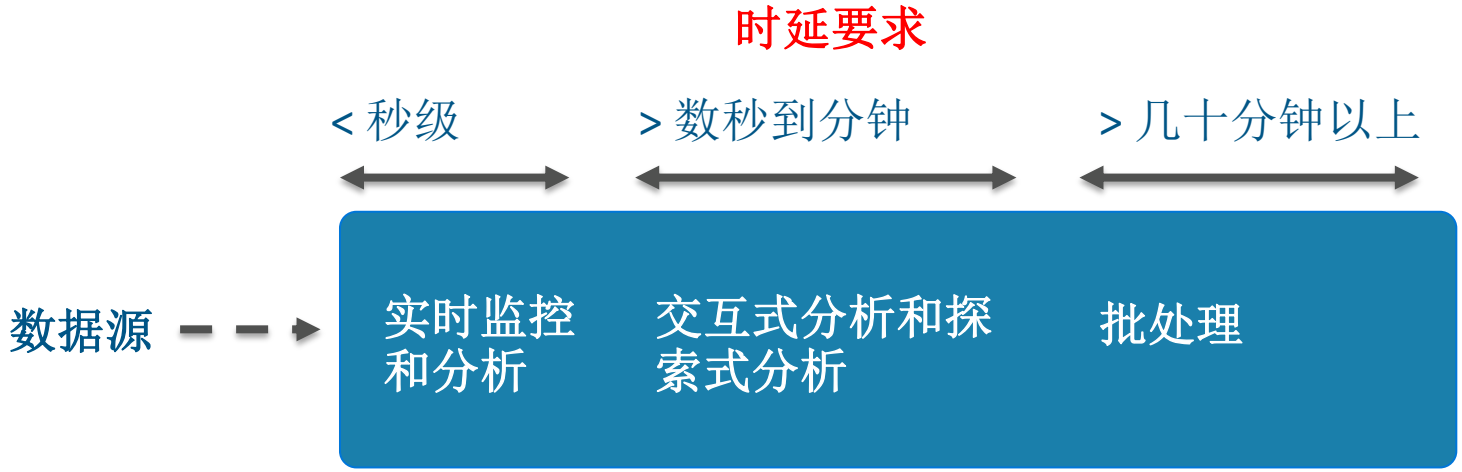
- Hive是一个SQL解析和优化层，底层引擎可以是MapReduce或是Spark
- SparkSQL是Spark生态系统的一个SQL解析和优化层，也需要借助于Spark引擎
- Impala就是一个原生的SQL解析、优化以及内存执行引擎，直接操纵HDFS



cloudera

数据处理和分析

- 多样化的SQL解决方案
  - 不同的需求需要不同的技术
  - 互补而不是替代



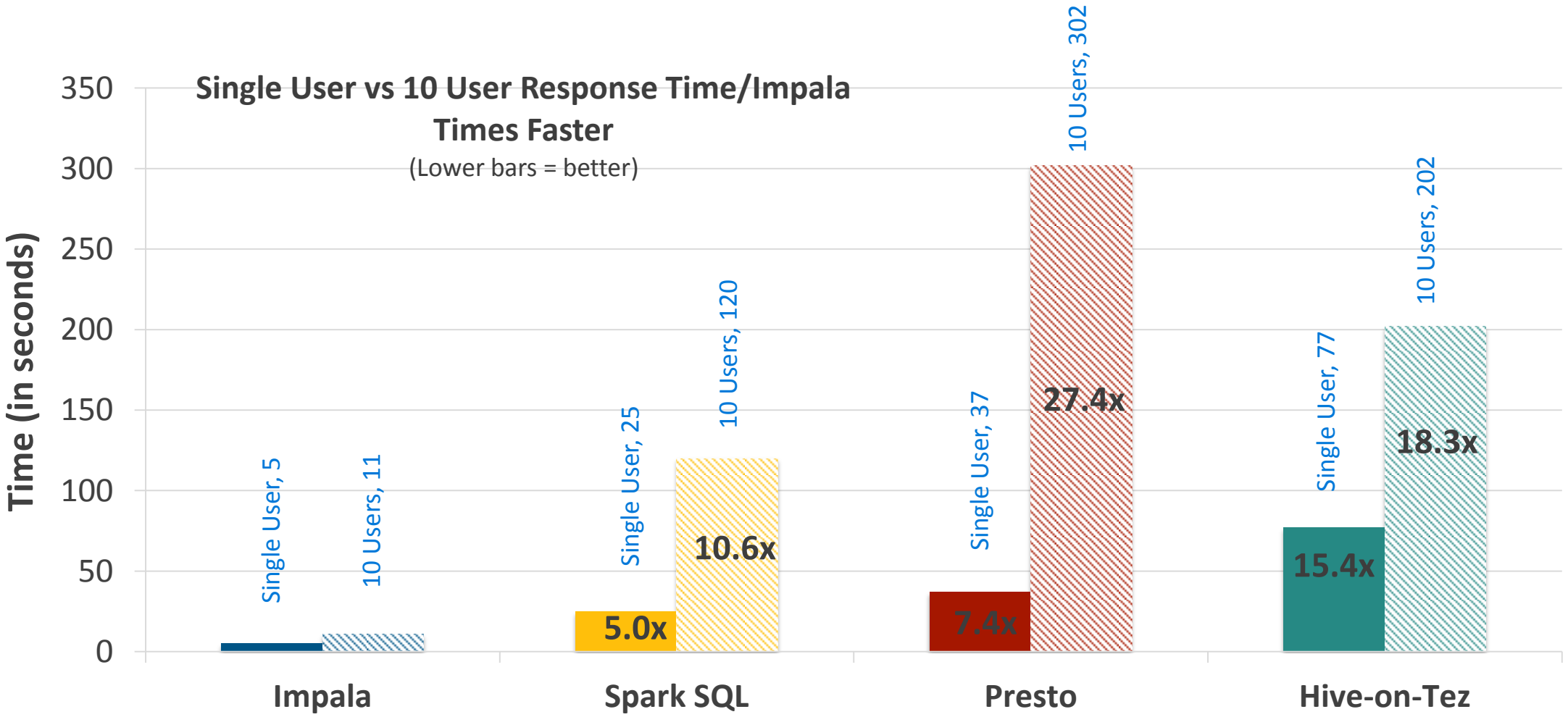
BI and SQL Analytics

Batch Processing

Spark developers

cloudera

交互式SQL性能



Independent validation by IBM Research SQL-on-Hadoop VLDB paper:
"Impala's database architecture provides significant performance gains"

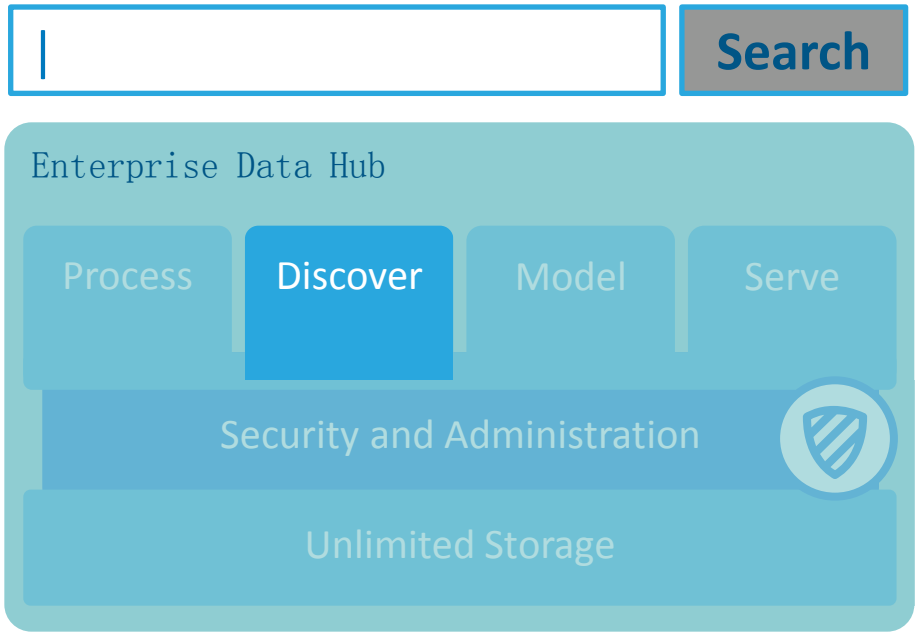
cloudera

Cloudera Search

大数据平台内的搜索引擎

易用性	实现了企业内数据平台的搜索引擎
标准化	基于Solr的标准搜索实现
灵活性	实现了多种索引的构建方式
安全和集成	和Cloudera企业级功能的紧密集成

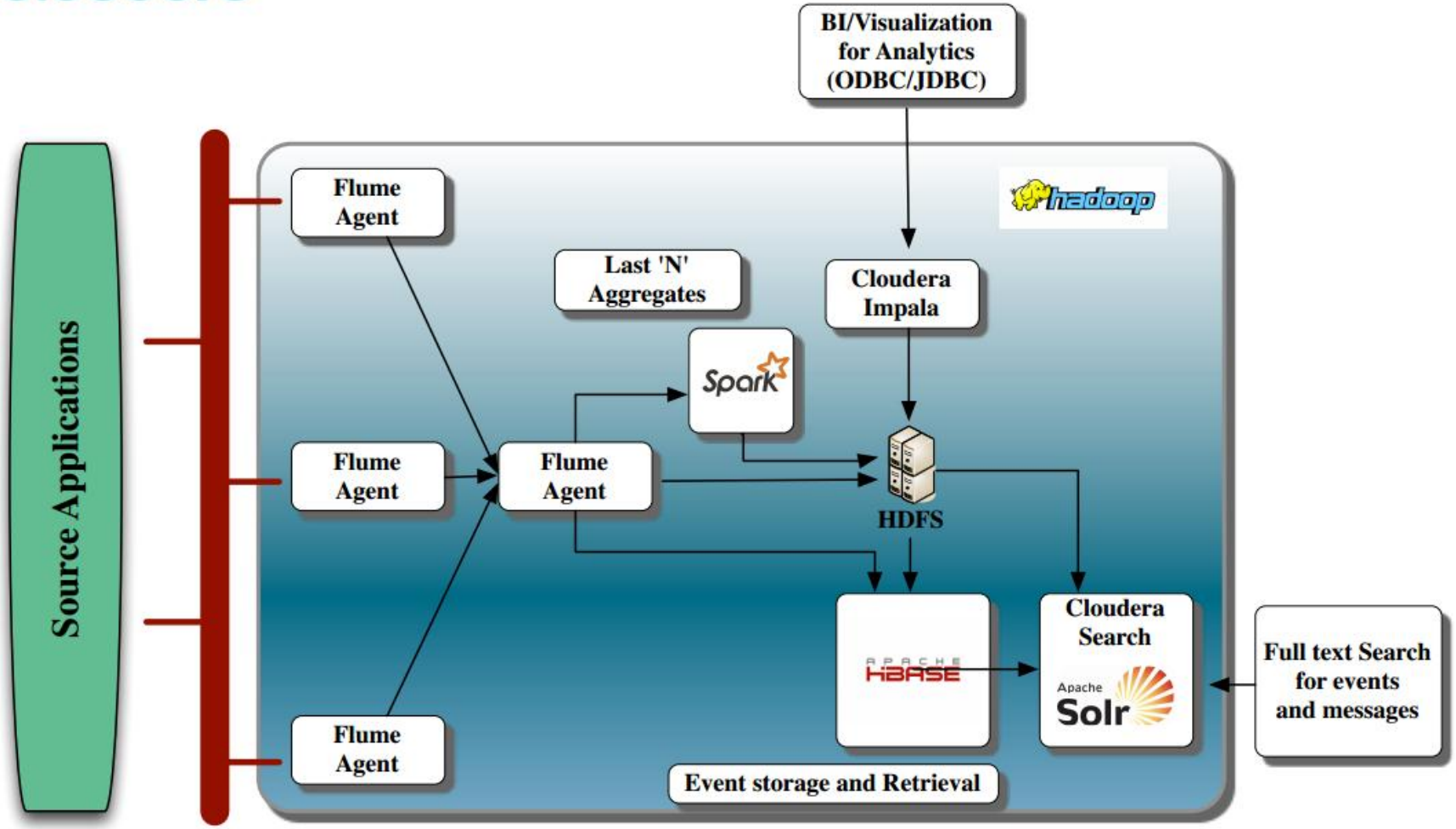
所有人都知道怎么搜索



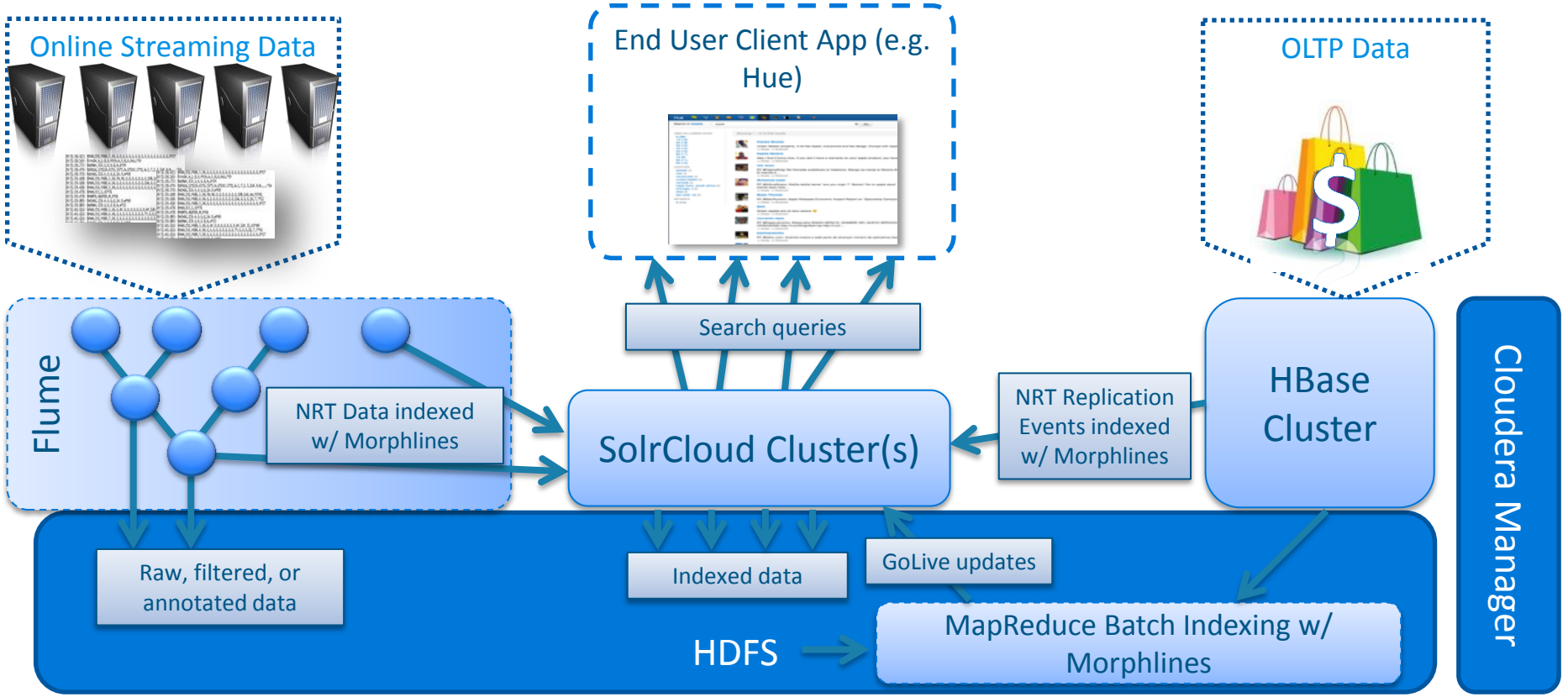
cloudera

实时数据处理

cloudera

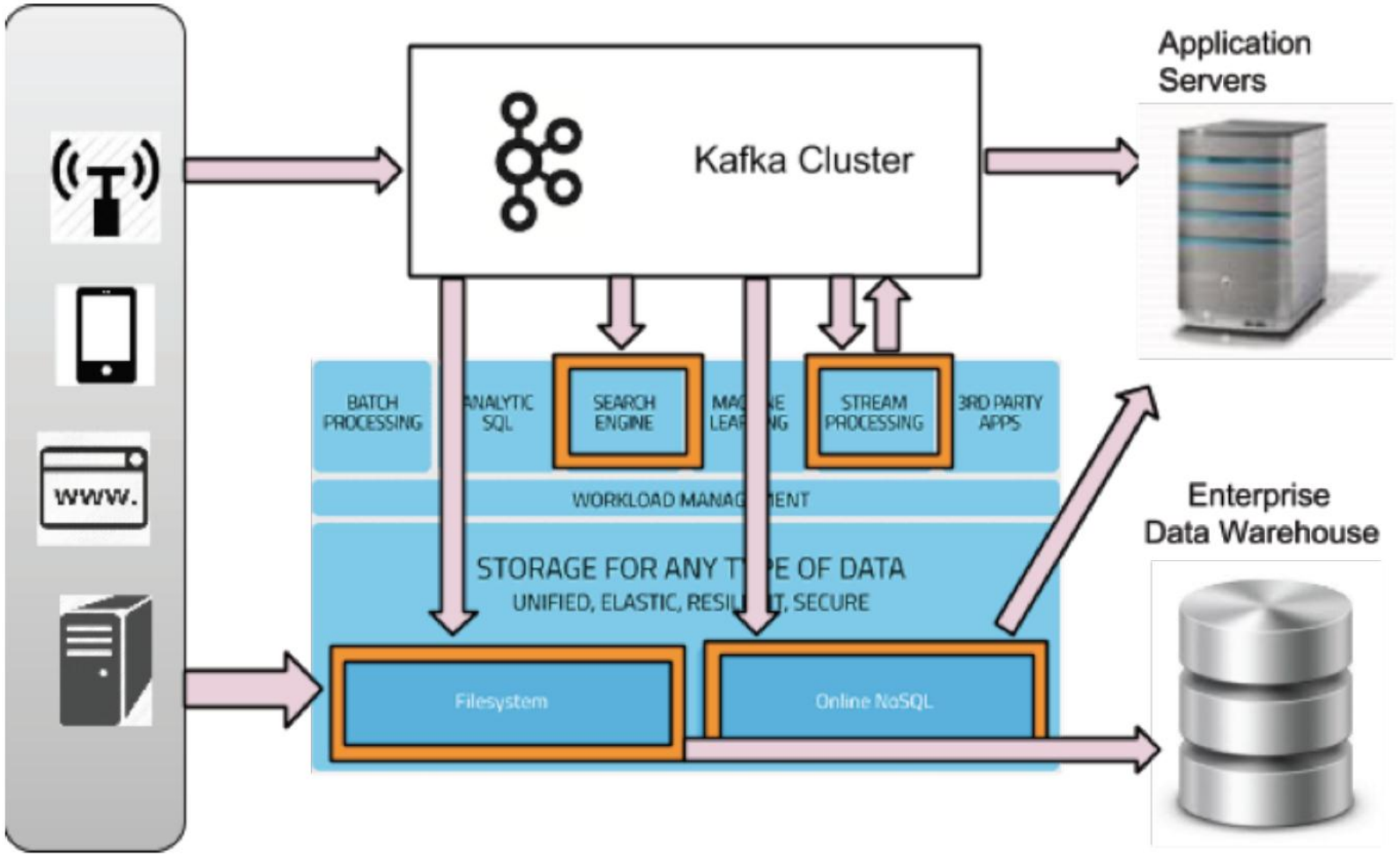


实时搜索



cloudera

企业消息总线



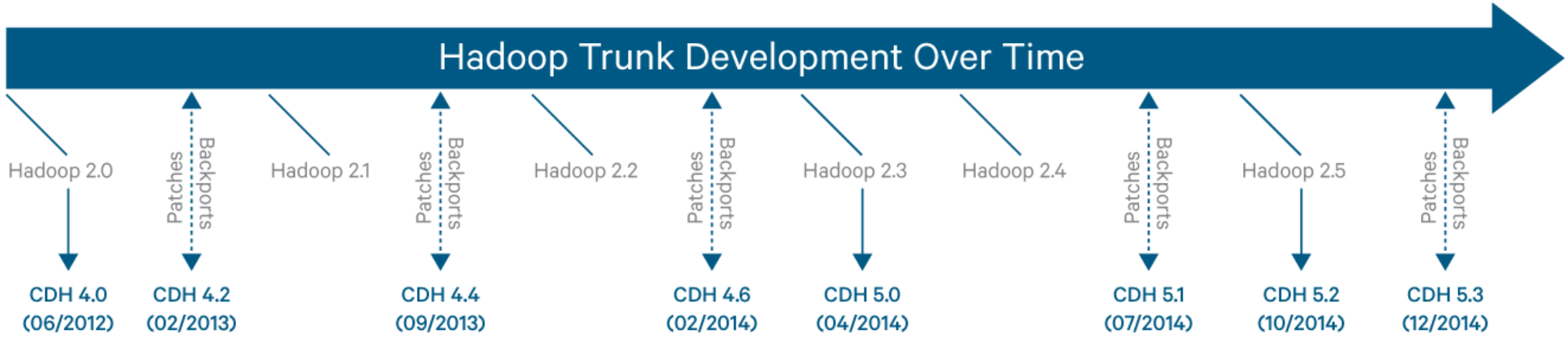
cloudera

Hue

- 专门为Hadoop打造的用户界面
  - HDFS的浏览以及管理
  - HBase的管理
  - 作业流设计，作业提交以及管理
  - SQL操作前端
  - 定制化的搜索前端
  - 访问权限配置界面

CDH发布模式

- 领先于开源的版本 –包含社区版本尚未发布的创新和稳定性功能
- 更快获取问题的修复 – 强大的Committer团队保证客户问题得到更快的修复
- 最广泛的测试 – 活跃的开源社区能让所有功能得到最全的测试



cloudera

CDH凝聚Cloudera在开源的贡献

- Cloudera有89位Hadoop以及相关生态的Committer，涵盖：
  - Hadoop, HBase, Hive, Spark, Lucene/Solr, Flume, Sqoop等项目
- Cloudera提供了最多的企业级Hadoop功能
  - HDFS/YARN HA, Hadoop Secure Communication, HDFS Short-Circuit, HDFS Caching, HDFS Transparent Encryption
  - HBase snapshots, HBase multi-tenancy
  - HiveServer 2, Hive-on-Spark
  - Spark Streaming exactly-once, Spark Shuffle Optimization
  - Solr + Hadoop Integration
  - ……

Security Infrastructure in Cloudera Enterprise – Apache Sentry & Cloudera Navigator

cloudera

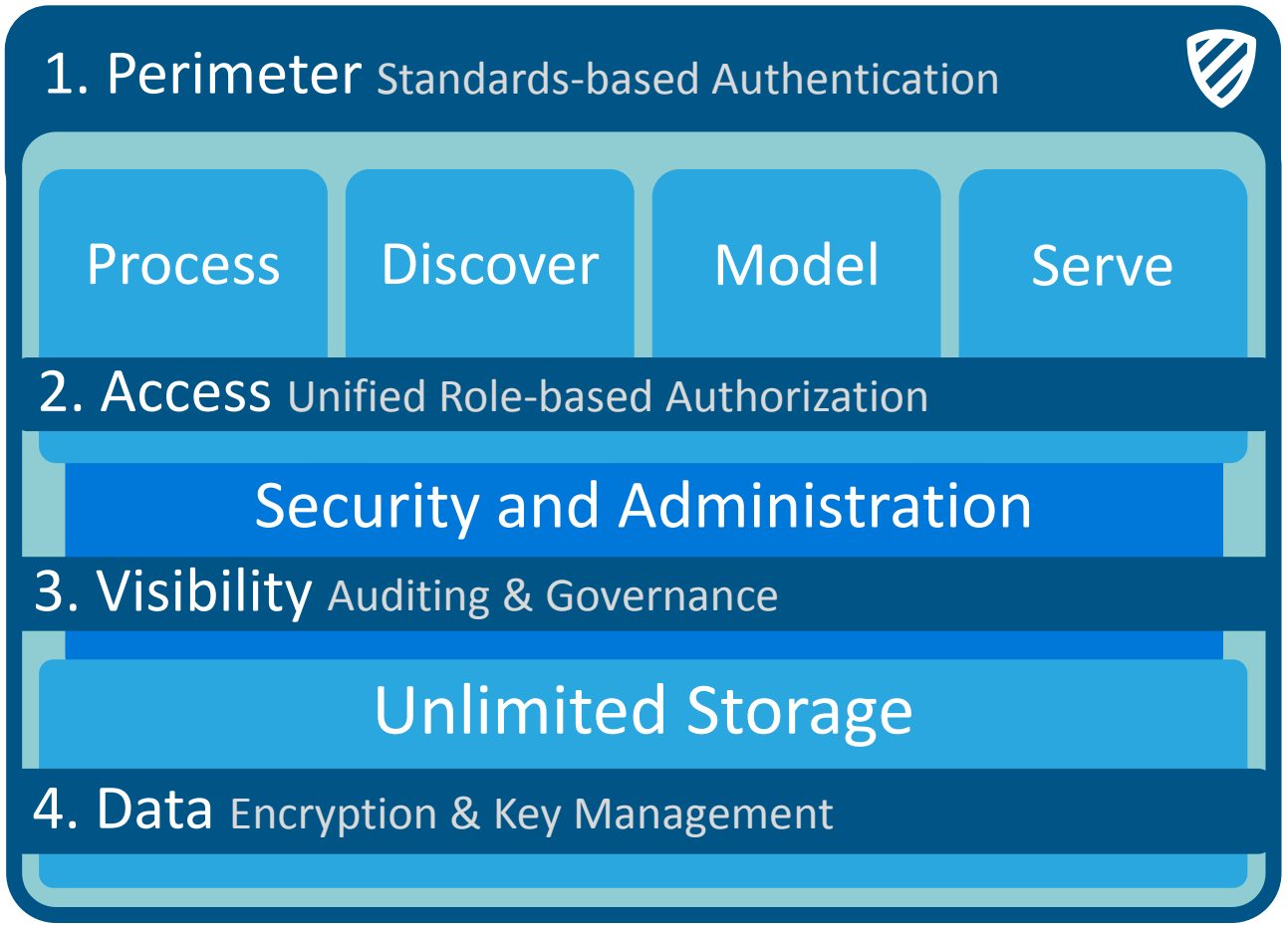
安全的挑战

- 越来越多的开发人员和业务人员会使用大数据平台
- 企业数据平台正成为黑客的主要目标
- Hadoop及衍生的众多项目缺乏统一的安全解决方案
- 传统的应用层安全方案难以胜任新平台
  - 平台有多种接口给用户使用
  - 传统方案中各应用系统相对独立
- 用户一旦突破应用层安全，数据平台就完全暴露
  - 数据没有任何保护
  - 访问没有任何限制

cloudera

全面的安全管控

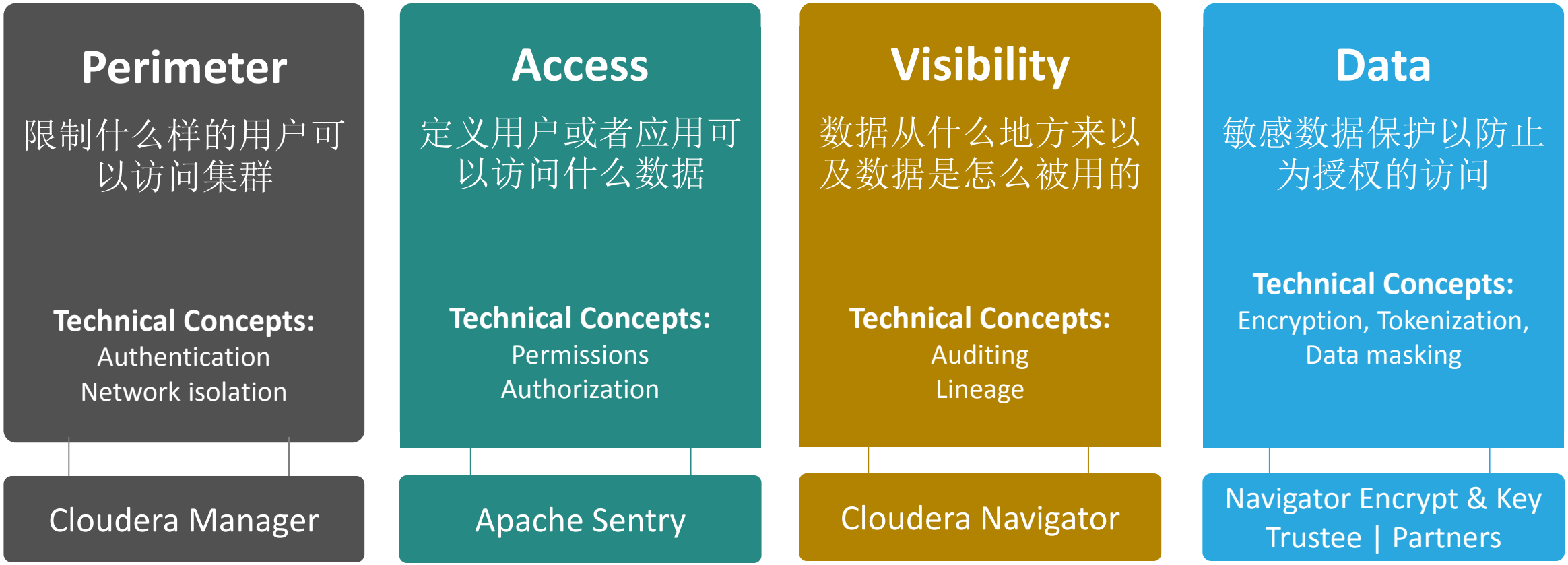
Apache Sentry, HDFS Encryption, Cloudera Navigator, Key Trustee



- 数据平台的安全不可或缺：
  - 多样化的数据导入方式
  - 多种引擎的协同工作
  - 多业务的并发
  - 多用户的访问
  - 和企业的基础设施集成
  - 符合行业的安全审查

安全技术架构

认证, 授权, 审计, 以及行业监管规范

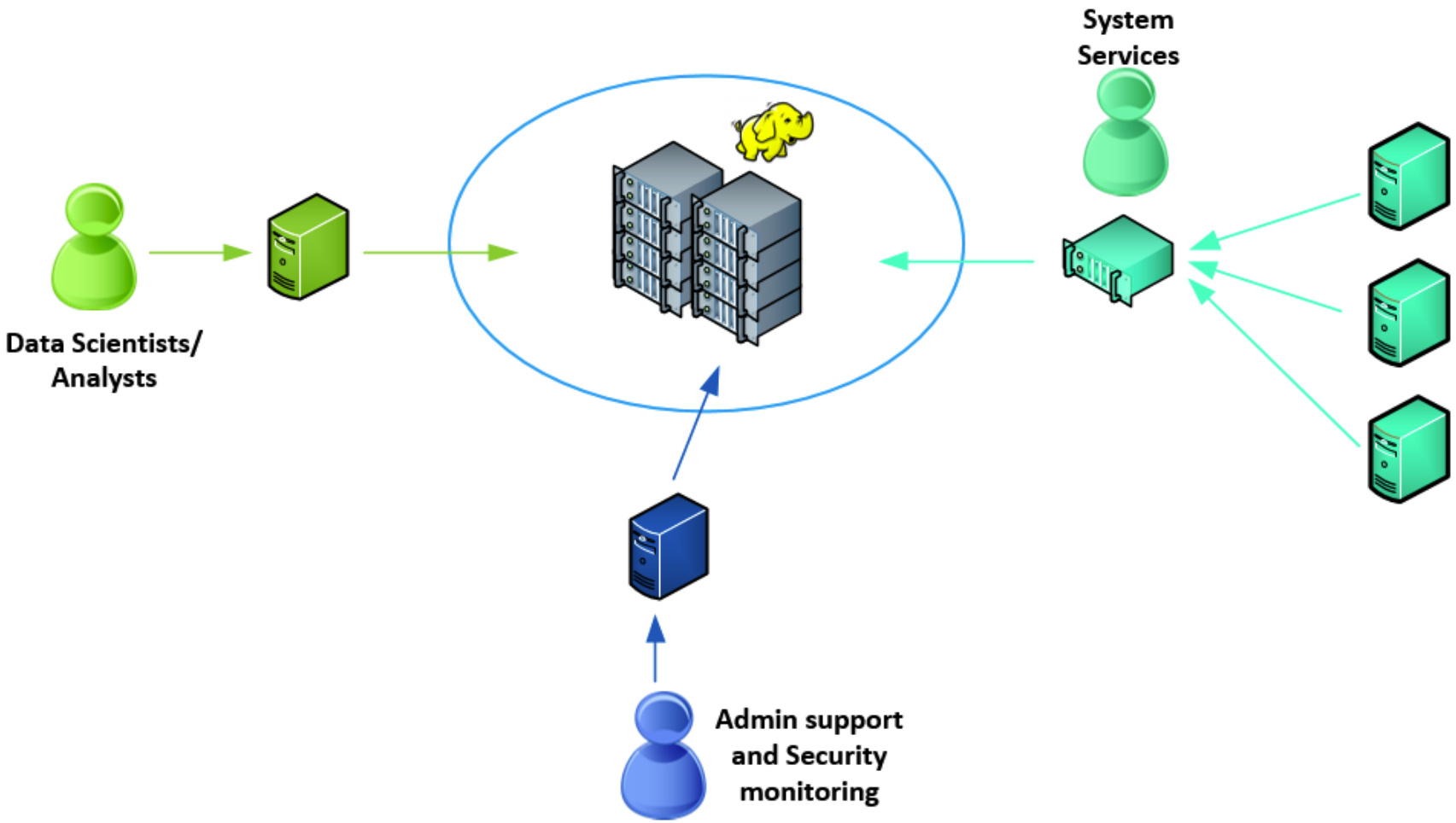


cloudera

多层次的安全

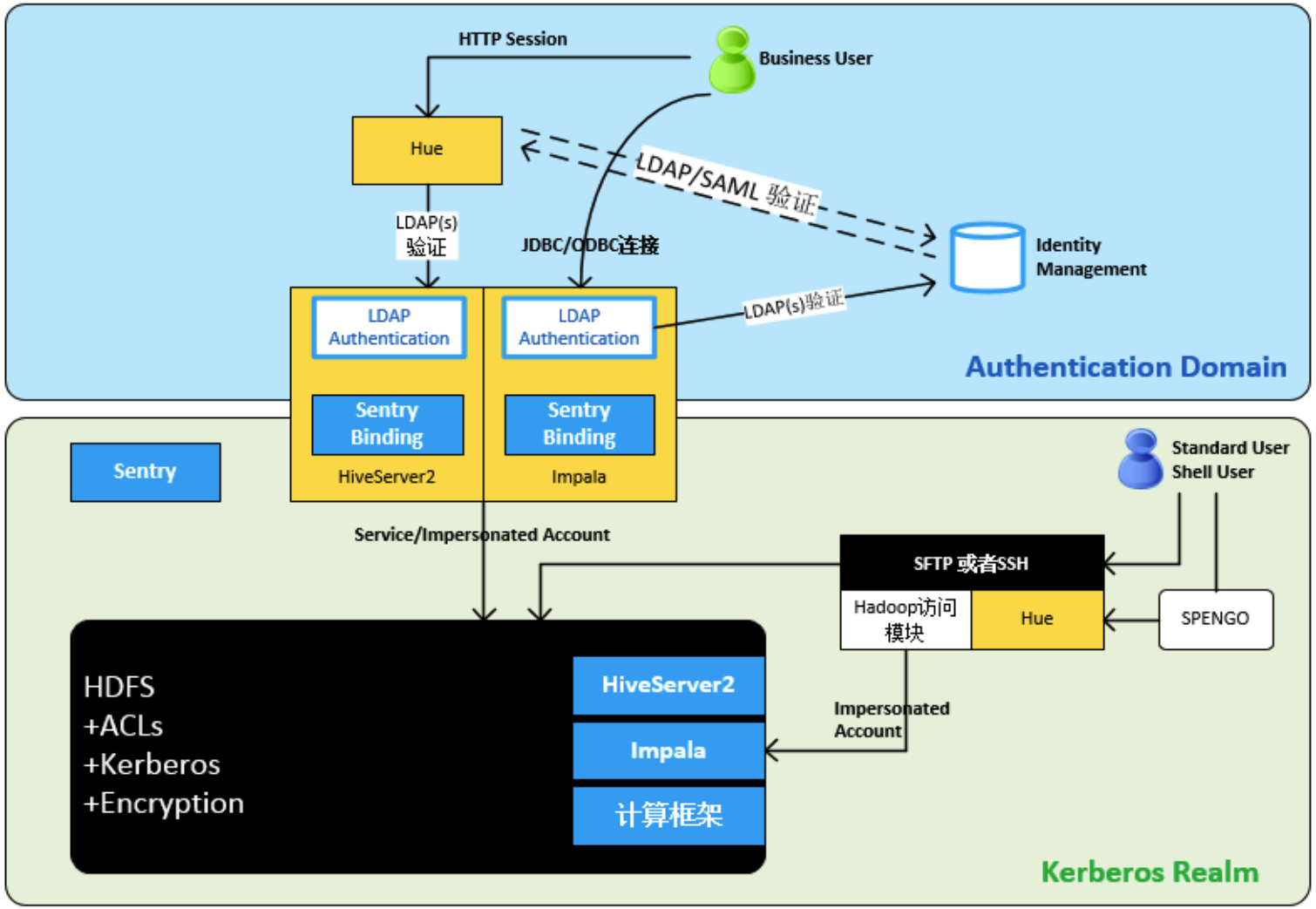
- 多层级的身份认证（Cloudera Manager, Kerberos, AD, Hue）
  - 管理平台，运维人员，客户端，BI工具
- 统一的授权访问控制（Apache Sentry）
  - 在平台上提供统一的访问安全控制策略
- 数据保护（HDFS At-Rest Encryption, Navigator Encrypt, Navigator KeyTrustee）
  - On-the-wire和at-rest数据保护，并内置有Key Management方案
- 全面的审计（Cloudera Navigator）
  - 不管以什么方式进行访问集群，都会得到审计

使用者视图



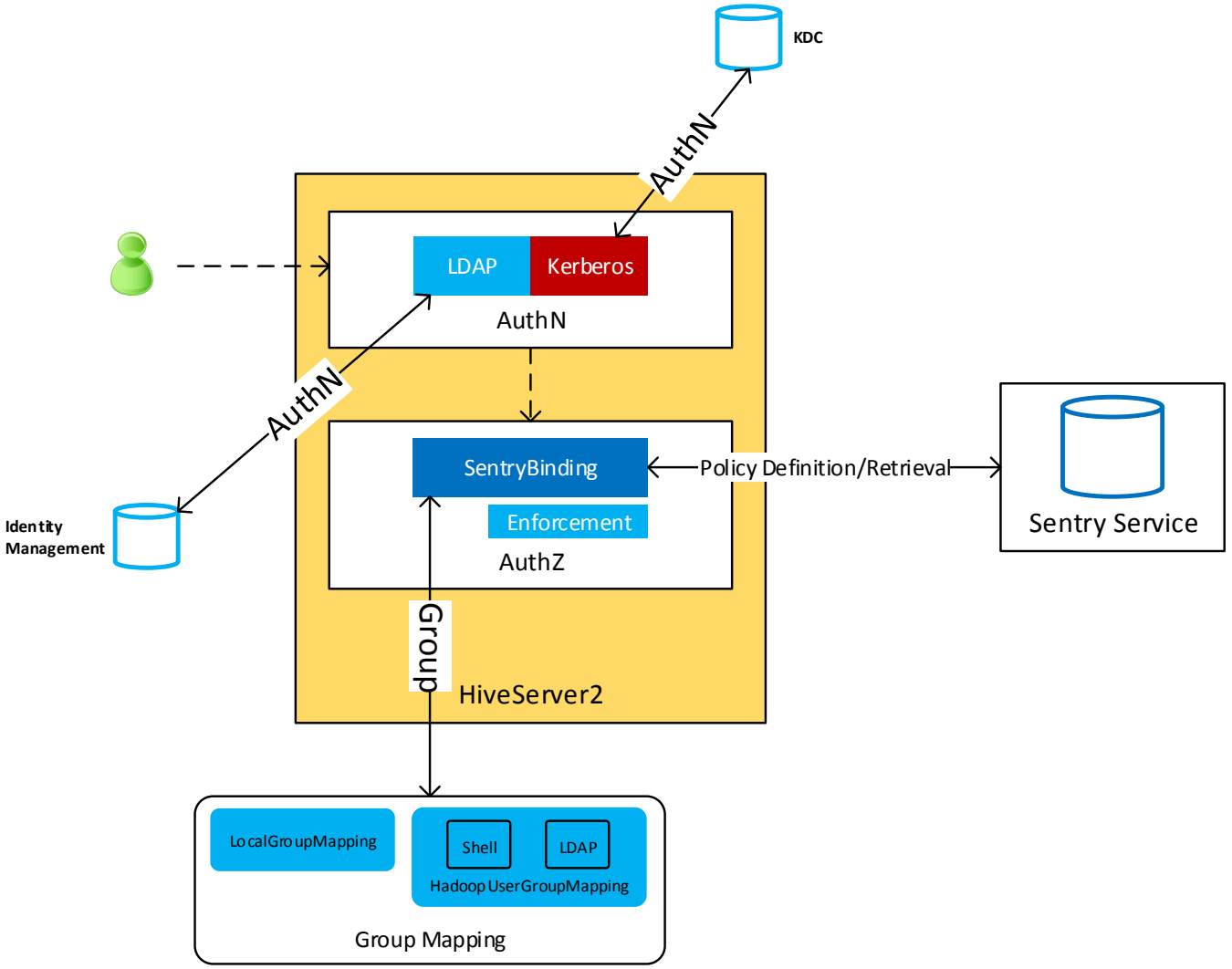
cloudera

认证和授权



cloudera

认证和授权



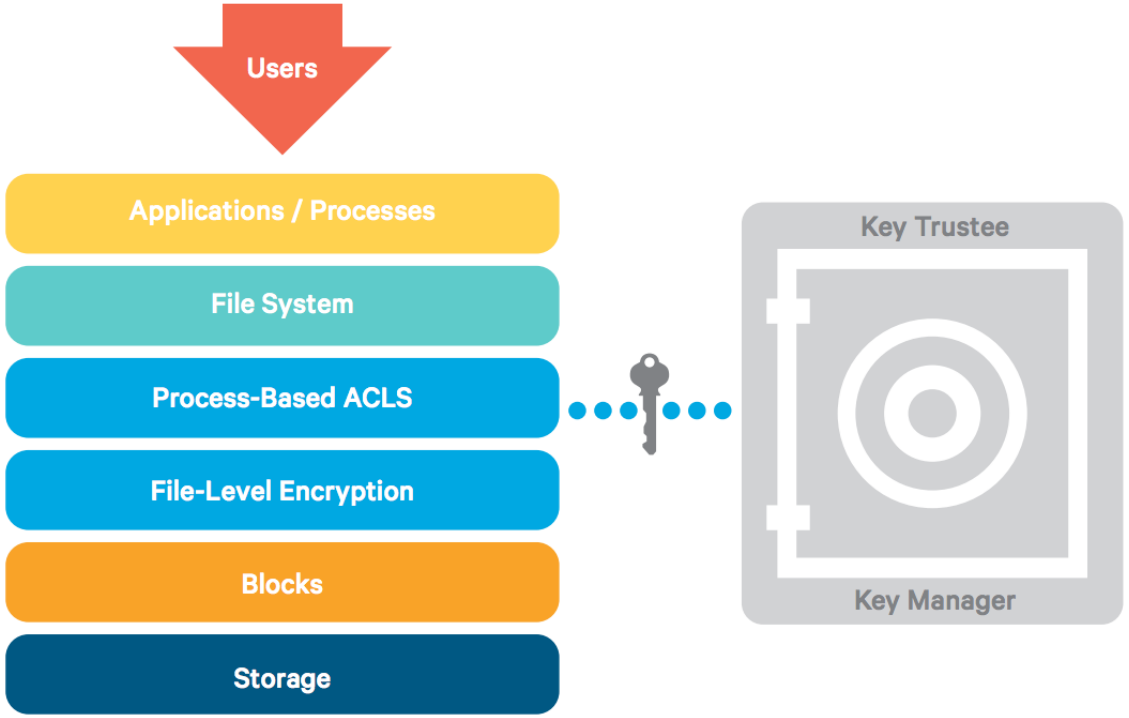
cloudera

数据保护

- 底层文件系统的数据保护 （Navigator Encrypt, Navigator KeyTrustee）
  - 临时文件，缓存到本地的中间计算结果，配置文件以及元数据文件
- HDFS文件的保护（HDFS Data-At-Rest Encryption, Navigator KeyTrustee）
  - 只能保护HDFS的文件或目录数据
- 网络传输的安全性（TCP over SSL）
  - 基于SSL的节点间网络通信

Navigator Encrypt/KeyTrustee (Gazzang)

- Navigator Encrypt
  - 全面高效的数据保护，Linux文件系统以下
  - 硬件指令加速（AES-NI）
  - 存储节点上的加解密方案
- Navigator KeyTrustee
  - 集中化的秘钥管理
  - 灵活的部署方式
    - on-premise或者SaaS



cloudera

Cloudera Navigator

- 全面的审计功能
  - 对HDFS、Impala、Hive、HBase和Sentry的审计追踪提供集中式的配置管理接口
  - 查看用户/用户组对HDFS、Impala、Hive和HBase的访问权限以保证对隐私及合规的正确配置
- 数据发现和探索
  - 快速检索相关数据，加速数据发现流程
  - 自动发现元数据并允许用户自定义可定制化标签与注释，便于数据追踪与归类
- 数据溯源
  - 帮助用户直观理解数据集的上下游血脉关系，验证数据源头与数据演变过程
  - 可以导出数据溯源信息到其他的溯源信息管理系统中
- 生命周期管理
  - 定义并自动化复杂的数据生命周期管理工作，包括分类，保留及加解密策略 – 一切都基于Navigator丰富的元数据管理能力

salesdata

This is the real estate sales history from March 2014 in New York. Obtained by monthly MLS feed.

tags: realestate, sales, mls

source type: HIVE

type: TABLE

path: hdfs://mdonsky-11.ent.cloudera.com:8020/user/hive/warehouse/salesdata

input format: org.apache.hadoop.mapred.TextInputFormat

output format: org.apache.hadoop.hive.ql.io.HiveIgnoreKeyTextOutputFormat

owner: training

created: Apr 3 2014 2:04 AM

source: HIVE-1

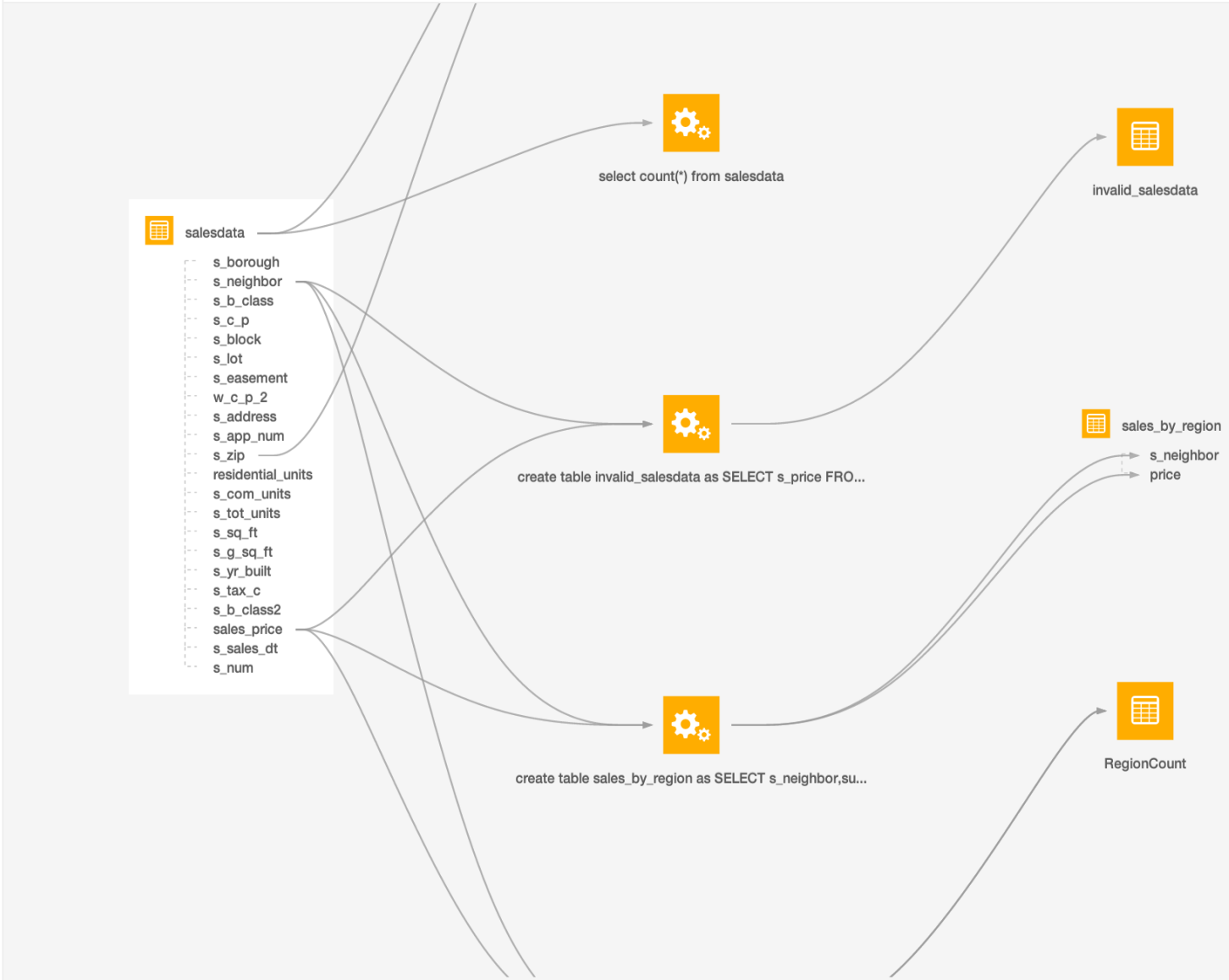
origin: mls

month: 03-2014

state: NY

retain-until: 5-15-2018

Lineage | Schema



sales_by_region

tags:

source type: HIVE

type: TABLE

owner: training

created: Apr 3 2014 2:55 AM

source: HIVE-1

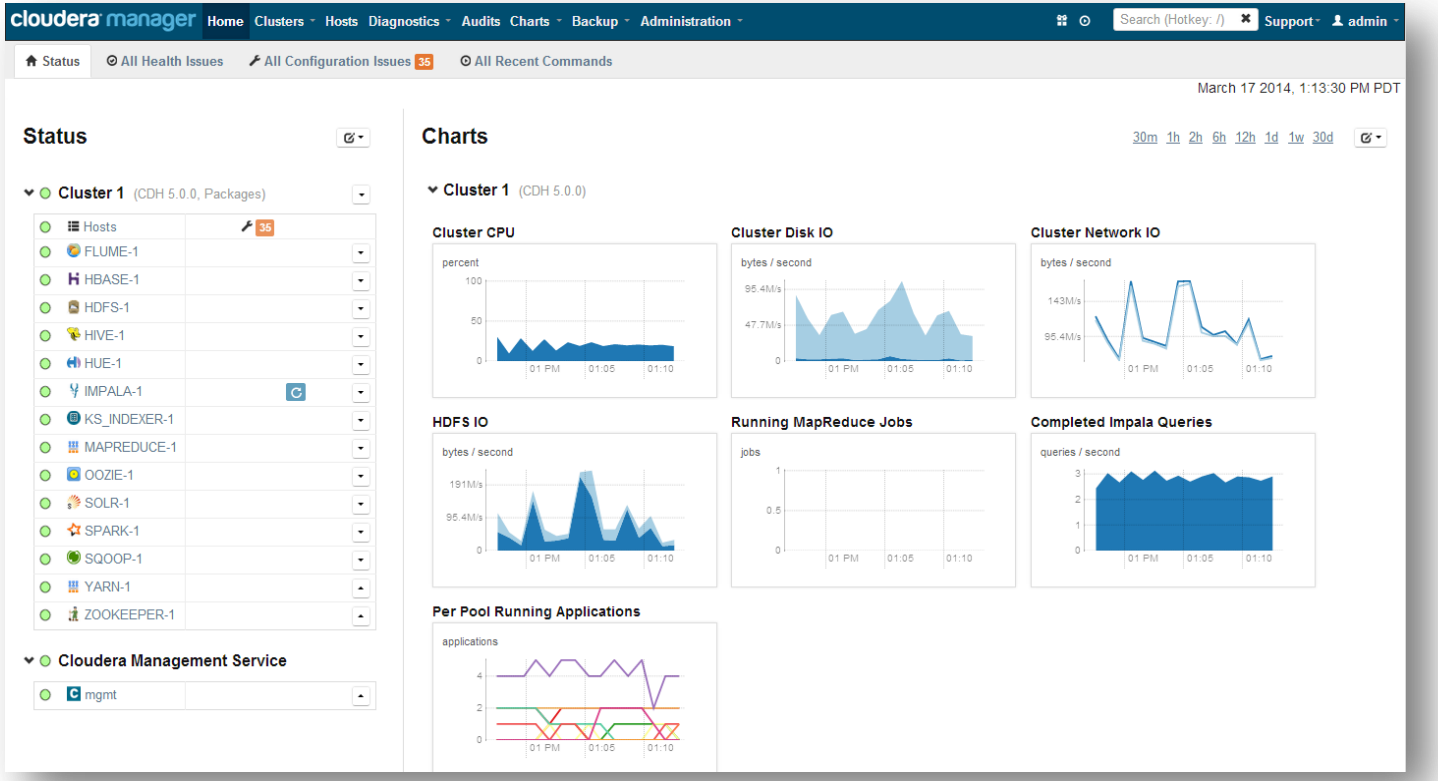
sensitivity: low (except for Brian)

expires: 2014

Most powerful Hadoop platform Management – Cloudera Manager

cloudera

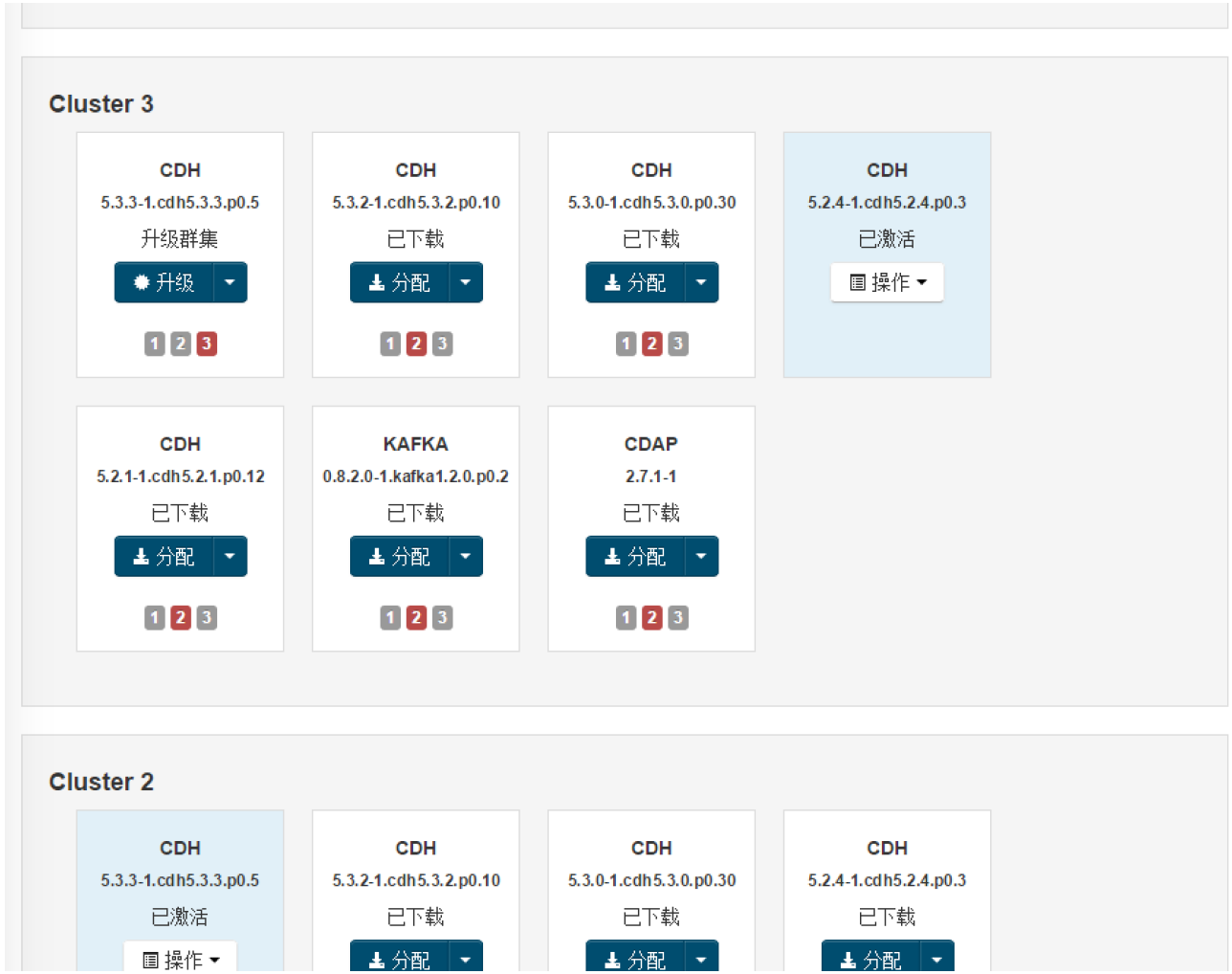
系统管理平台
Cloudera Manager

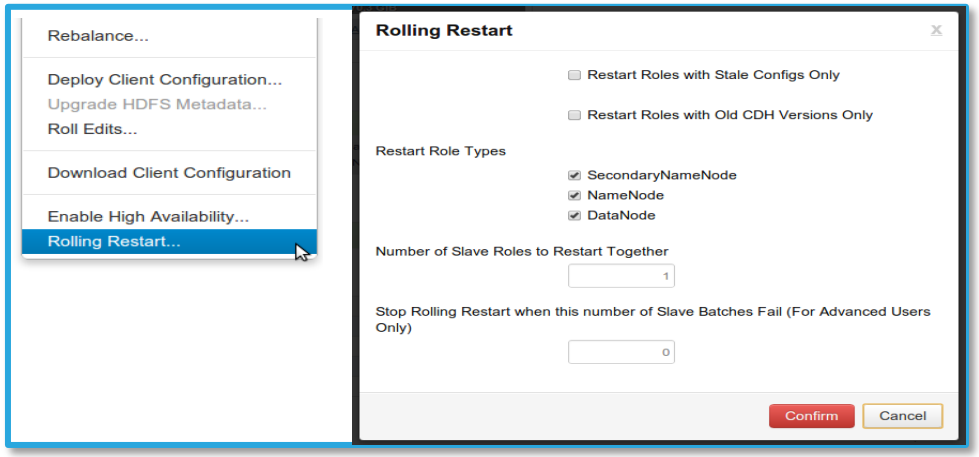


- Cloudera Manager – 专注于企业管理平台，而不只是一个集群管理工具
  - 基于角色的管理视图
  - 丰富且可定制化的监控图表展现
  - LDAP/Kerberos/SNMP/Rest API集成
  - 零宕机安装和升级
  - 复制和灾备
  - 多租户资源管理
  - 自动化的运营和诊断报告
  - 开放API可以集成第三方工具
  - ……

cloudera

滚动重启和升级

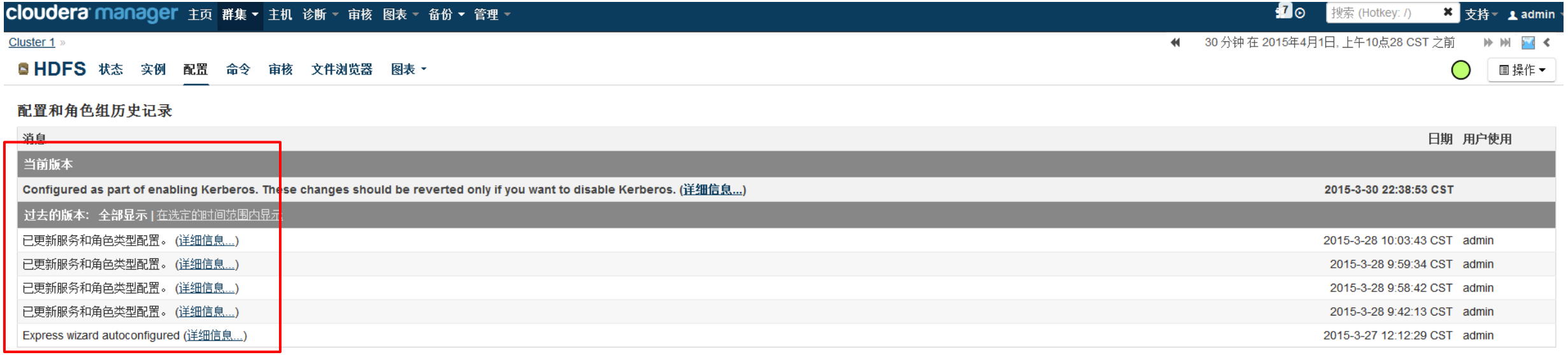




cloudera

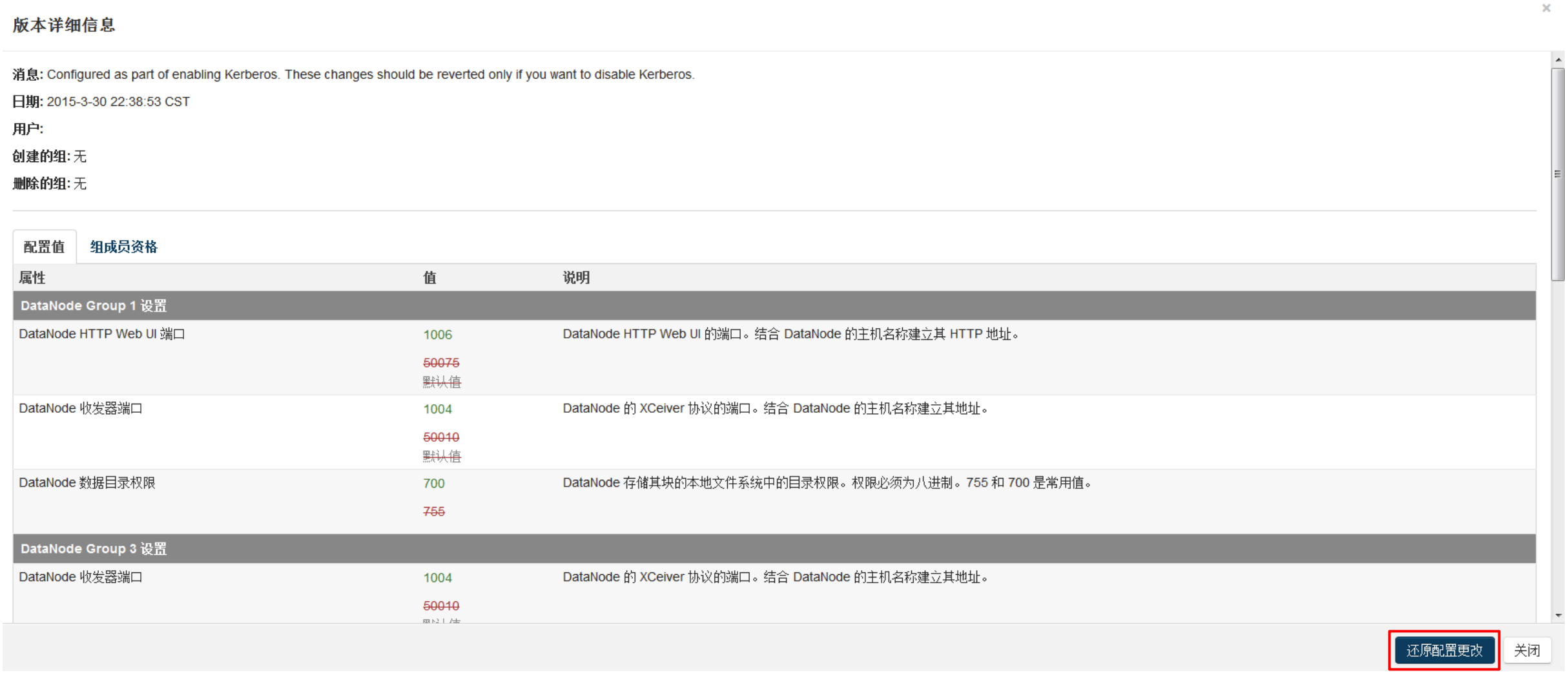
集群灾备

配置历史



cloudera

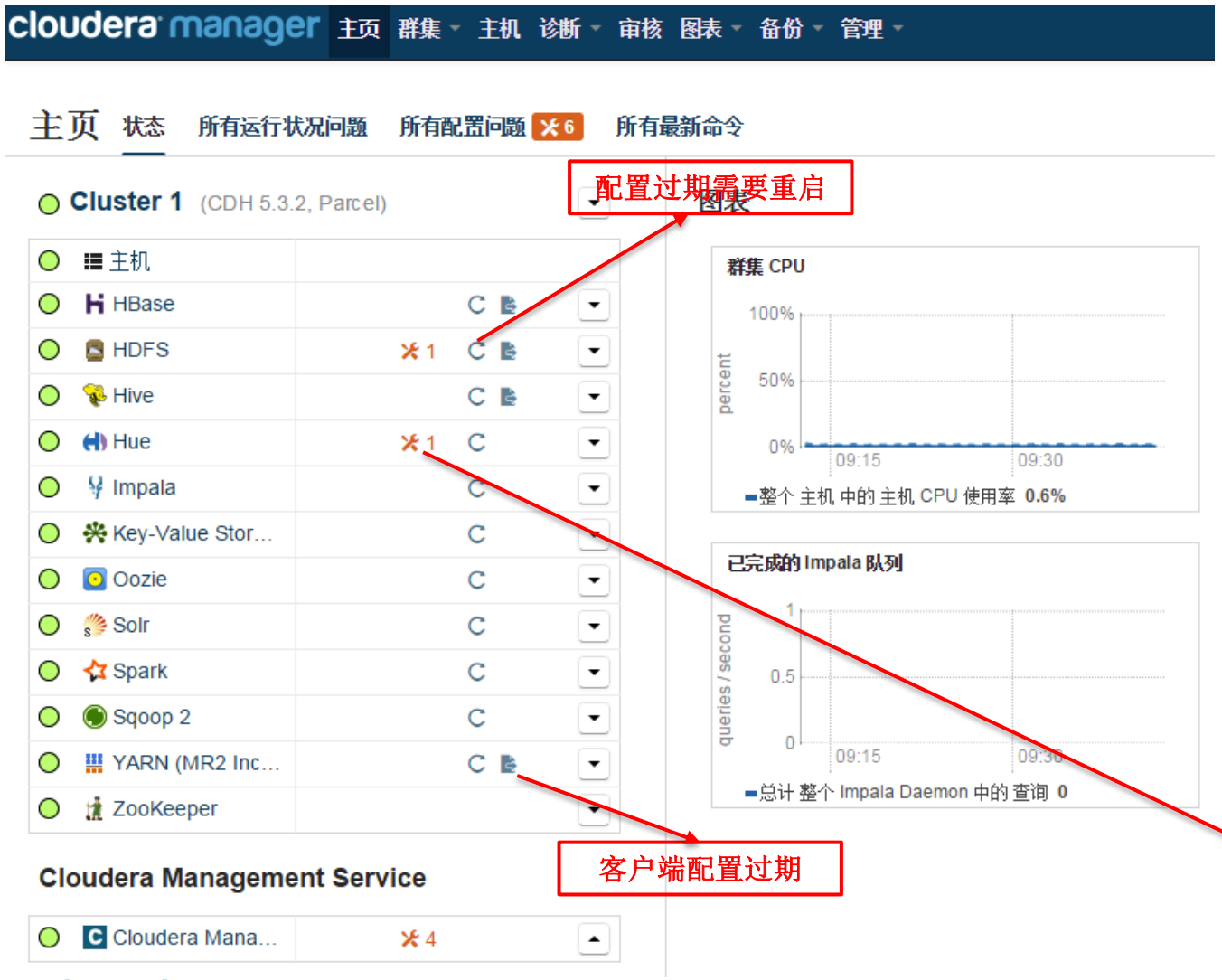
版本回滚

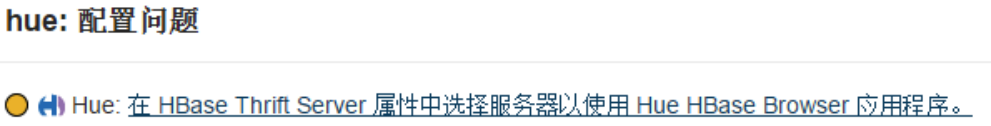


智能的配置警告

cloudera

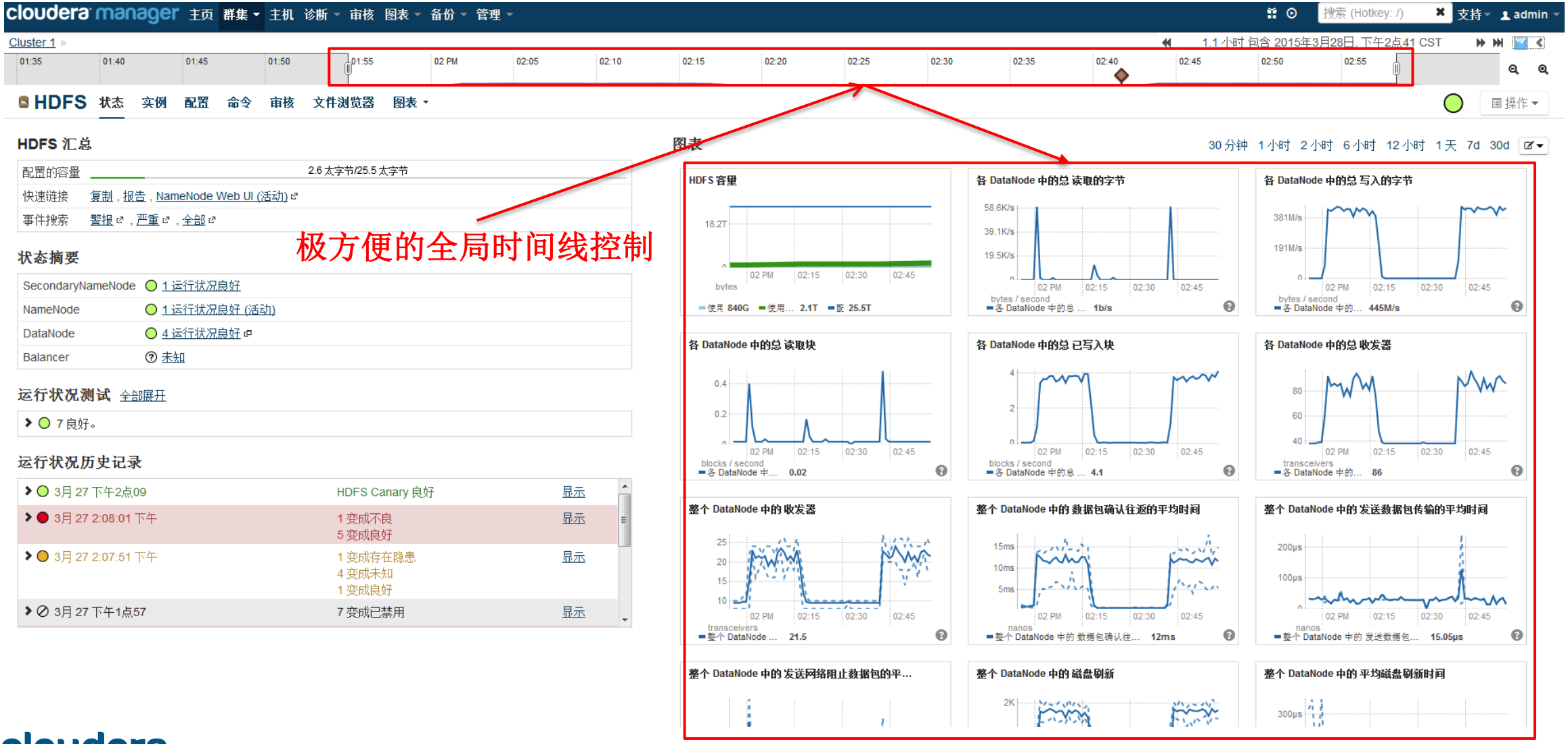
智能决策





cloudera

全局时间线控制方便诊断



cloudera

启用Kerberos

启用 Kerberos 用于 Cluster 1

欢迎

此向导将带领您完成配置 Cloudera Manager 和 CDH 以使用 Kerberos 进行身份验证的步骤。群集以及 Cloudera Management Service 中的所有服务将作为向导的一部分重启。在继续使用向导前，请阅读有关启用 Kerberos 的文档。

使用向导前，请确保已执行以下步骤：

设置正在运行的 KDC。Cloudera Manager 支持 MIT KDC 和 Active Directory。

☑ 是的，我已设置正在运行的 KDC。

KDC 应配置为拥有非零票证生存期和可更新的生存期。如果票证不可更新，则 CDH 不能正常工作。

☑ 是的，我已检查 KDC 允许可更新的票证。

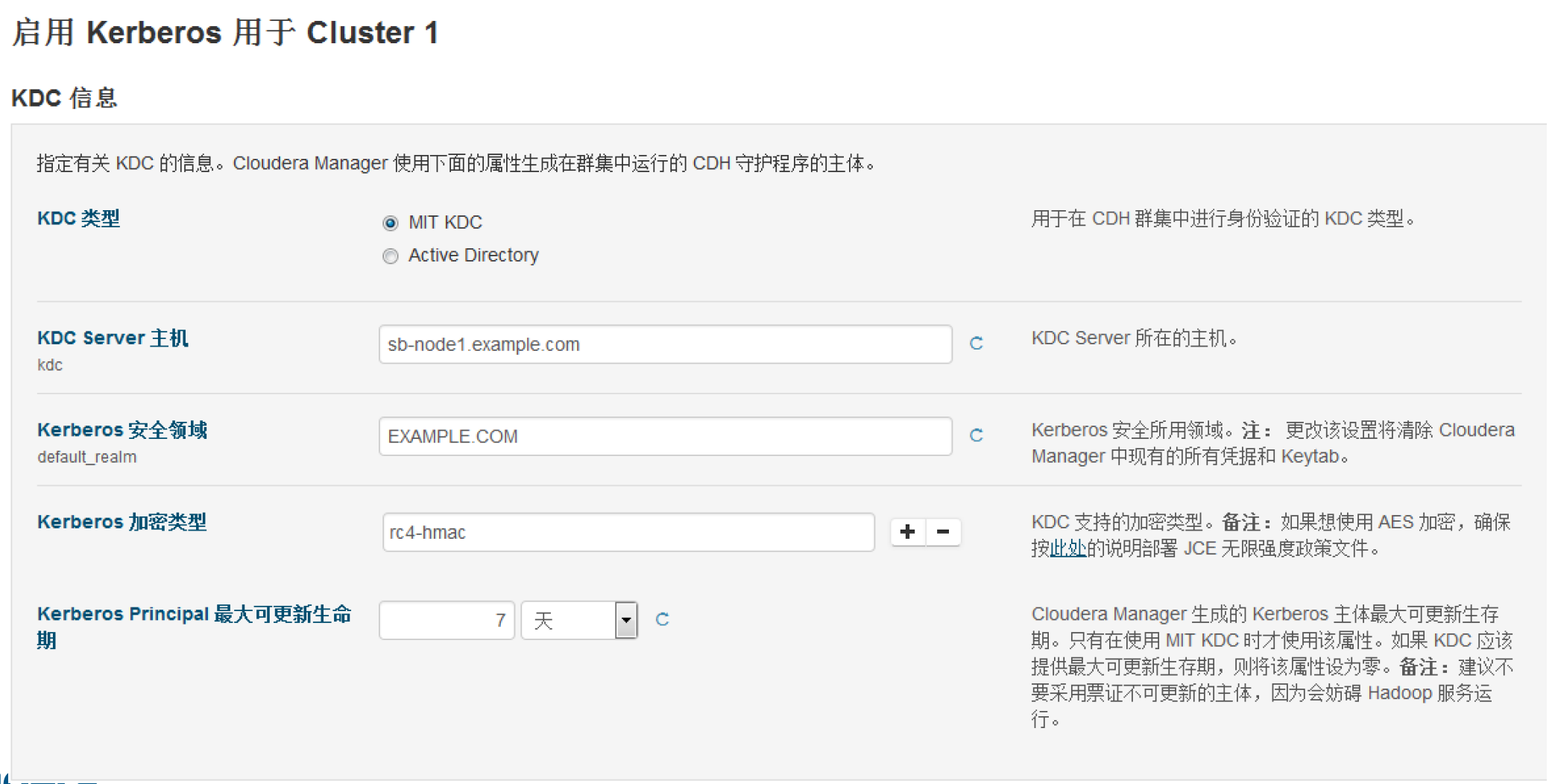
如果想使用 Active Directory，OpenLdap 客户端库应安装在 Cloudera Manager Server 主机中。另外，Kerberos 客户端库应安装在所有主机中。

☑ 是的，我已安装客户端库。

Cloudera Manager 需要有权限在 KDC 中创建其他帐户的帐户。

☑ 是的，我已为 Cloudera Manager 创建适当的帐户。

启用Kerberos



启用Kerberos

KRB5 配置

指定生成群集的 krb5.conf 所需的属性。可以使用安全阀字段指定高级 KDC 设置的配置，例如，使用跨领域身份验证。

通过 Cloudera Manager 管理 krb5.conf ☐

Cloudera Manager 是否应该在安全群集中配置和部署 krb5.conf。如果没有选中该属性，则必须确保在安全群集的主机中以及 Cloudera Manager Server 的主机中部署 krb5.conf。

cloudera

启用Kerberos

进度

命令	上下文	状态	开始日期	结束于
启用 Kerberos	Cluster 2	正在进行	2015-3-31 13:24:18 UTC	

命令进度

已完成 4 个步骤（共 7 个）。

- ✓ 停止群集
  `All services successfully stopped.`
  详细信息
- ✓ 停止 Cloudera Management Service
  `Command completed with 6/6 successful subcommands`
  详细信息
- ✓ 将所有服务配置为使用 Kerberos
  已成功完成 16 个步骤。
- ✓ 等待生成凭据
  已成功完成命令 (829)
- **部署客户端配置**
  详细信息
- 启动 Cloudera Management Service
- 启动群集

cloudera

通过Cloudera Manager管理用户自定义服务

GitHub, Inc. [US] https://github.com/cloudera/cm_ext/wiki

This repository Search Explore Gist Blog Help centiteo

cloudera / cm_ext Watch 29 Star 19 Fork 10

Home

sbodoff edited this page on 13 May 2014 · 31 revisions

Cloudera Manager Extensions

Cloudera Manager is a powerful tool for managing CDH services and the clusters they run on. Through the use of Cloudera Manager extensibility mechanisms, it is possible to use Cloudera Manager to manage non-Cloudera provided services alongside CDH services, and deploy plugins and extensions for any such managed services.

Before reading about Cloudera Manager extensibility, or developing your own extensions, you should be familiar with the basic operating principles of Cloudera Manager.

What is extensible about Cloudera Manager?

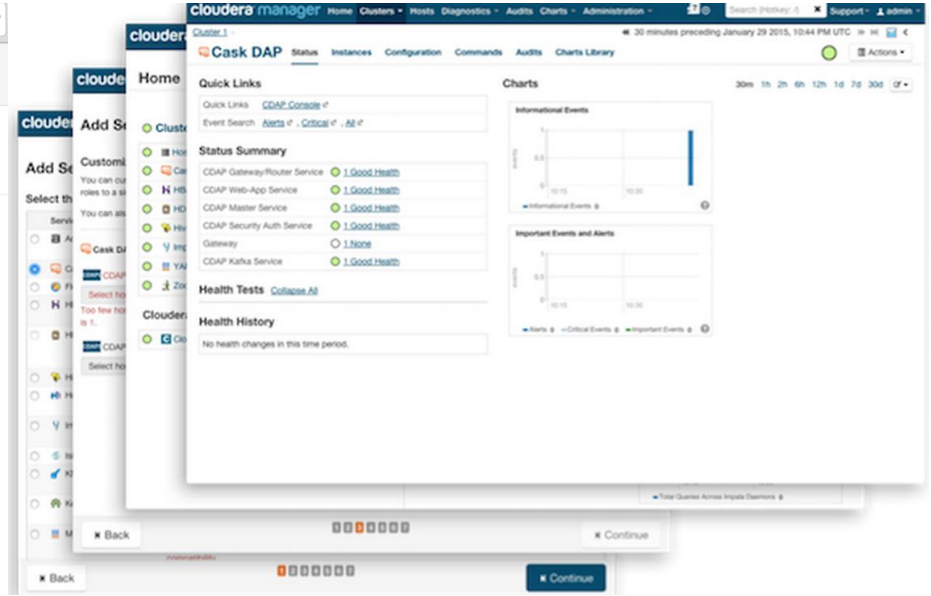
At the highest level, Cloudera Manager concerns itself with managing the lifecycle of various services running on a cluster. This can be broken down into two primary areas:

- Managing the deployment of the program/data files required for a service to run
- Managing and monitoring the configuration and running of those services

Pages 25

Find a Page...

- Home
- Administration of CSDs
- Building a parcel
- Control Scripts
- CSD Developer Tricks and Tools
- CSD Overview
- CSD Primer
- CSD Upgrade Process
- Monitoring Support for CSDs
- Parcel distro suffixes
- Parcels: What and Why?
- Plugin parcel environment variables
- Resource management support for csds



https://github.com/cloudera/cm_ext/wiki

Cloudera Manager Rest API

Cloudera Manager API v9

REST Data Model Files and Libraries

Home

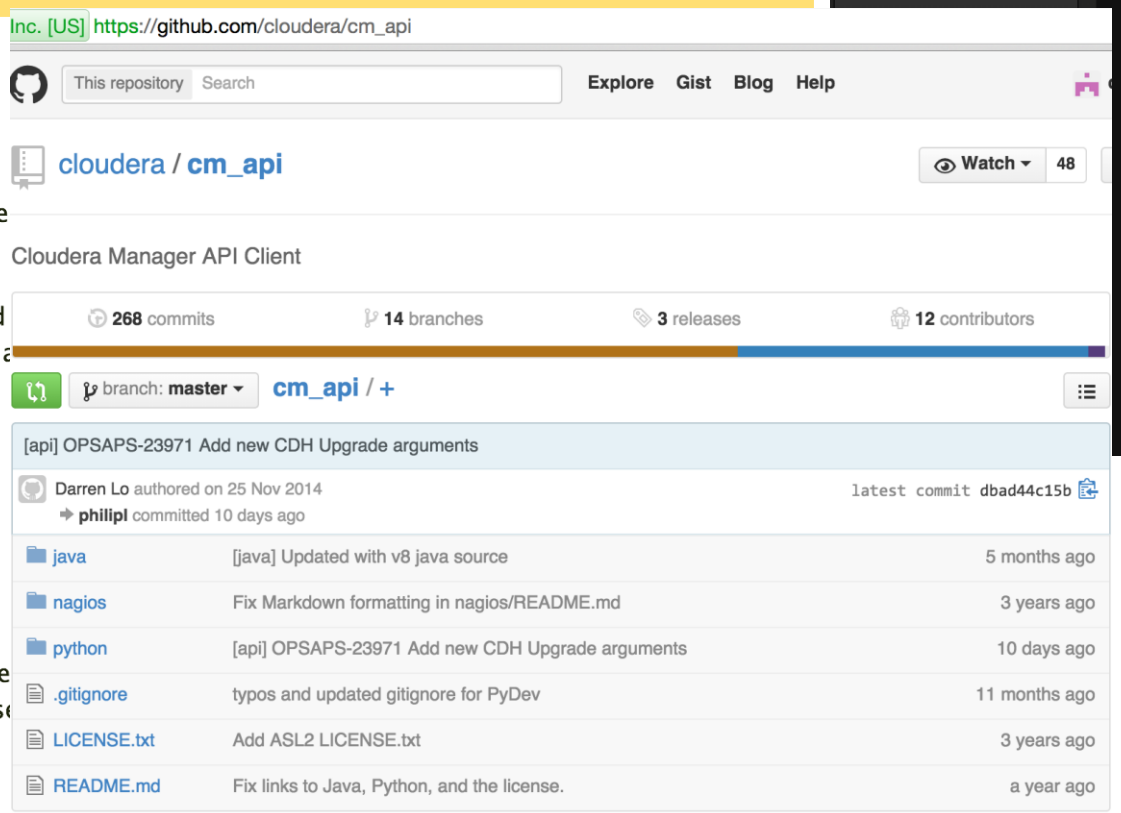
Introduction

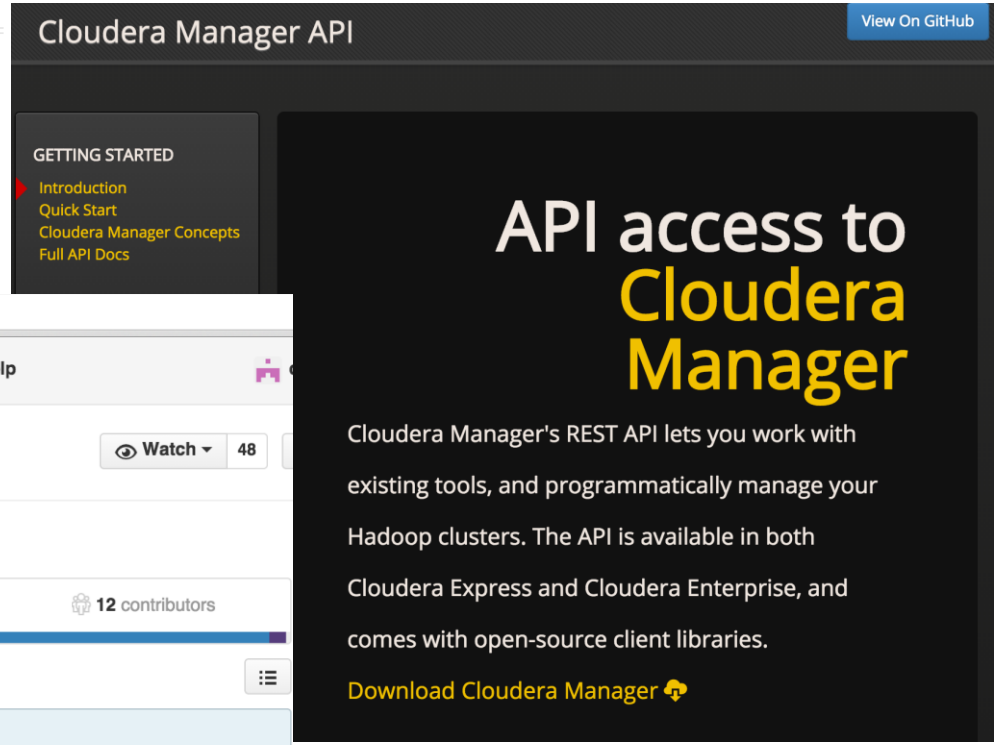
This document describes the Cloudera Manage
Object Notation (JSON).

The API resources listed below follow standard
request path defines the entity to be acted on a

HTTP Method	Operation
POST	Create entries
GET	Read entries
PUT	Update or edit entries
DELETE	Delete entries

All collections in the API use plural names, 'use
system, expand the URL path to include the us
'/users/bar' identifies user 'bar'.





cloudera

多租户管理理

- 在多用户的环境下共享相同的系统或程序组件，且仍可确保各用户间数据、配置甚至计算资源的隔离性。
  - 各租户的资源保障
  - 租户间的细粒度的安全隔离
  - 租户资源请求的快速响应
  - 租户资源使用的报告
- 多租户的优势
  - 数据共享
  - 方便运营
  - 提高资源使用率

cloudera

多租户的挑战

- 开源版本已经实现的
  - YARN的资源管理平台，可以实现对MapReduce、Spark的动态资源管理
  - 基于Queue的资源抽象描述
  - 基于Queue的用户权限控制
- 挑战
  - 只支持批处理的引擎
  - 对有时延要求租户的支持
  - 统一的权限控制模型
  - 没有对租户资源使用的详细报告

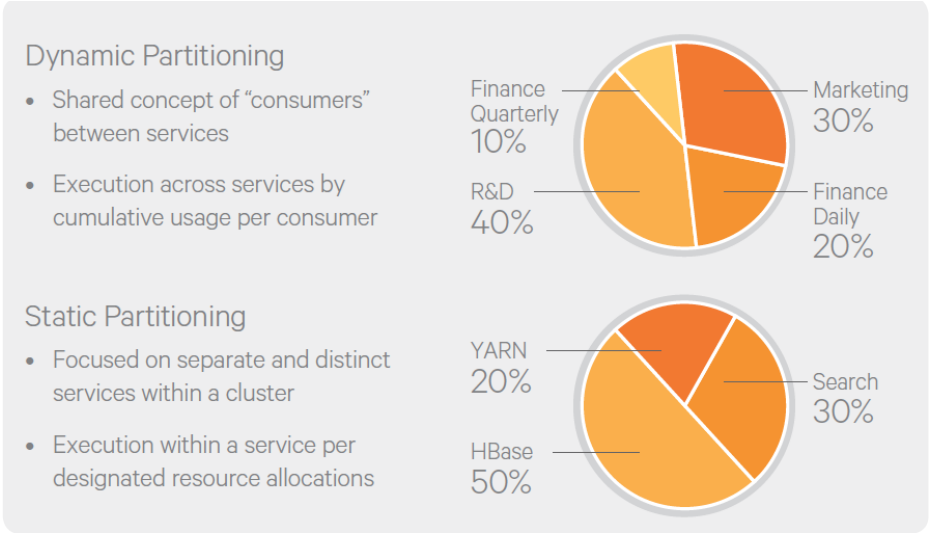
cloudera

Cloudera平台的多租户

- 资源隔离和管理
  - 保障租户对服务质量的要求，且有效利用集群的资源
- 安全和管治
  - Cloudera平台提供了从身份验证、授权、审计和数据安全的全面保护，确保租户之间的隔离性
- 资源使用报告
  - 统计租户对资源的使用要求，优化租户的资源分配

资源管理

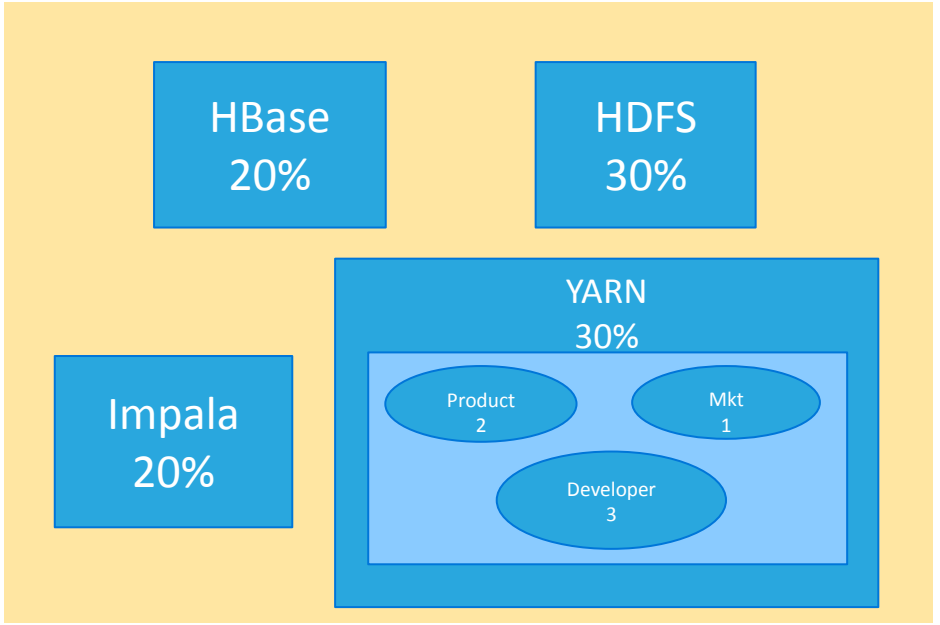
- 资源划分
  - 动态资源划分
    - 按需给租户提供满足服务质量的资源保障
    - 有效利用集群资源
  - 静态资源划分
    - 满足关键负载的作业保障
- 配额管理
  - 磁盘空间配额
  - 文件、目录数量配额，以优化文件系统元数据



cloudera

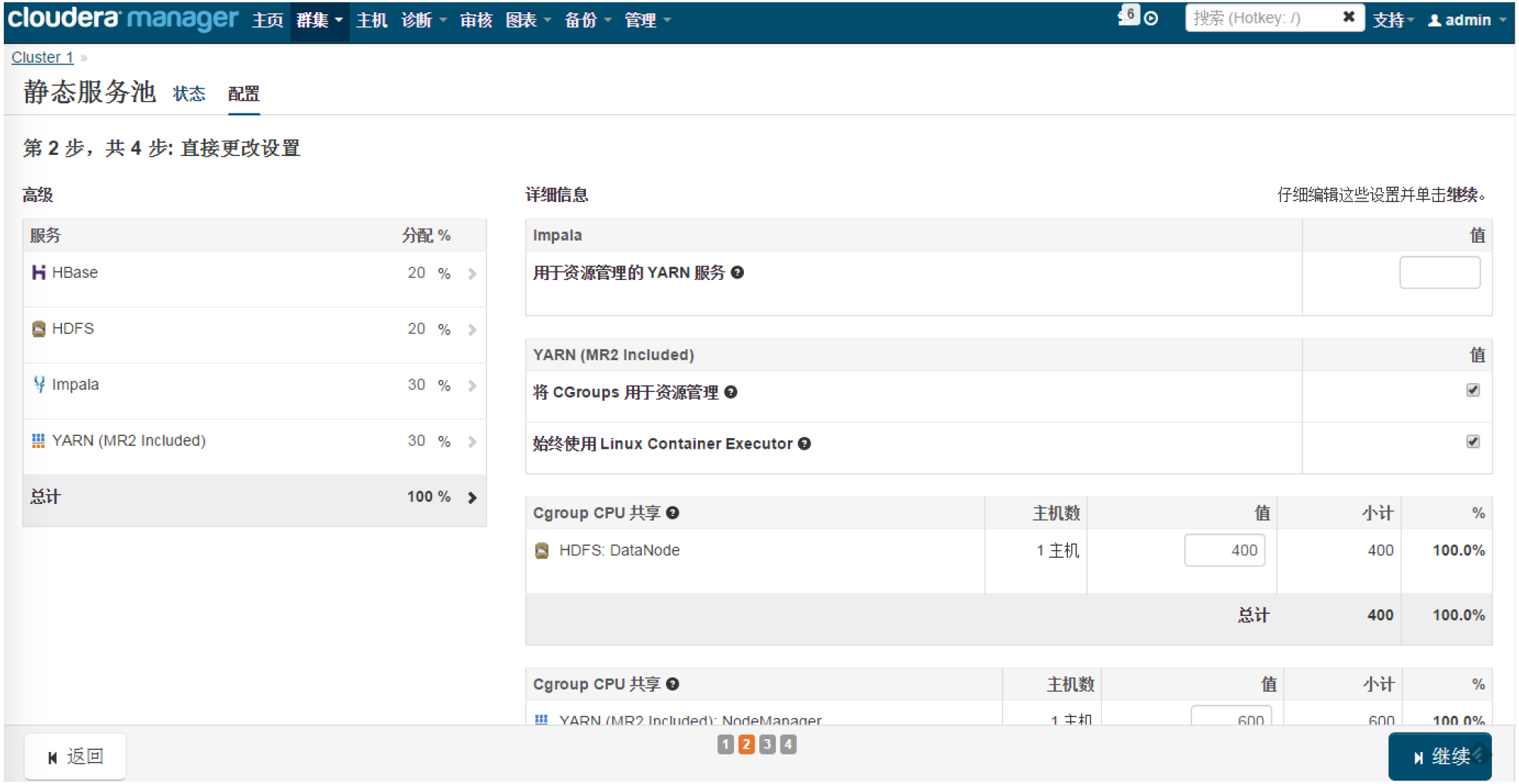
静态资源管理

- 通过Linux cgroup来静态划分各服务所占用的资源
  - 支持HBase, HDFS, Implala, YARN
- 保障关键作业的资源占用



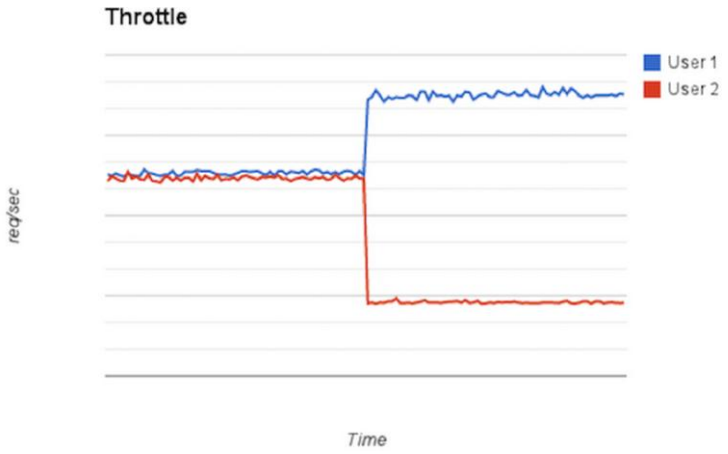
cloudera

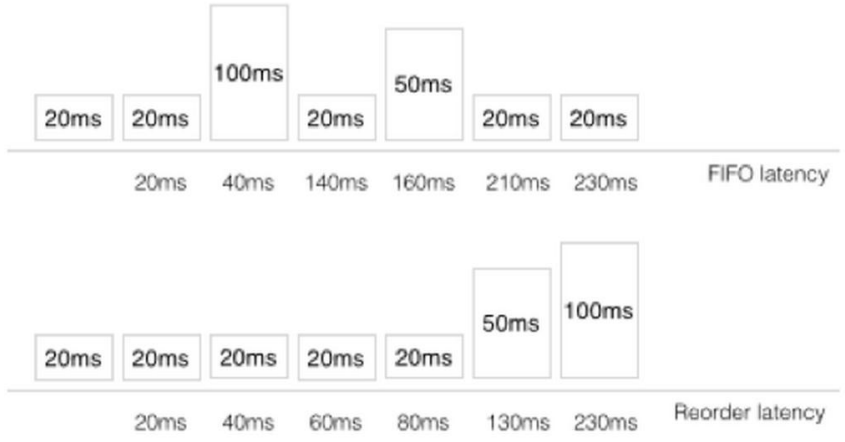
静态资源管理配置

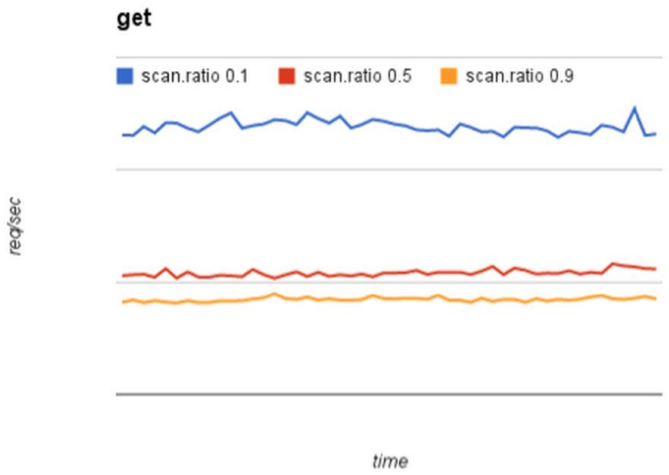


HBase内部的资源管理

- 对某个用户、某张表或某个表空间的访问进行限制（Throttling）
- 将HBase上的作业按类型进行调度
  - 分析或查询
  - 读或写



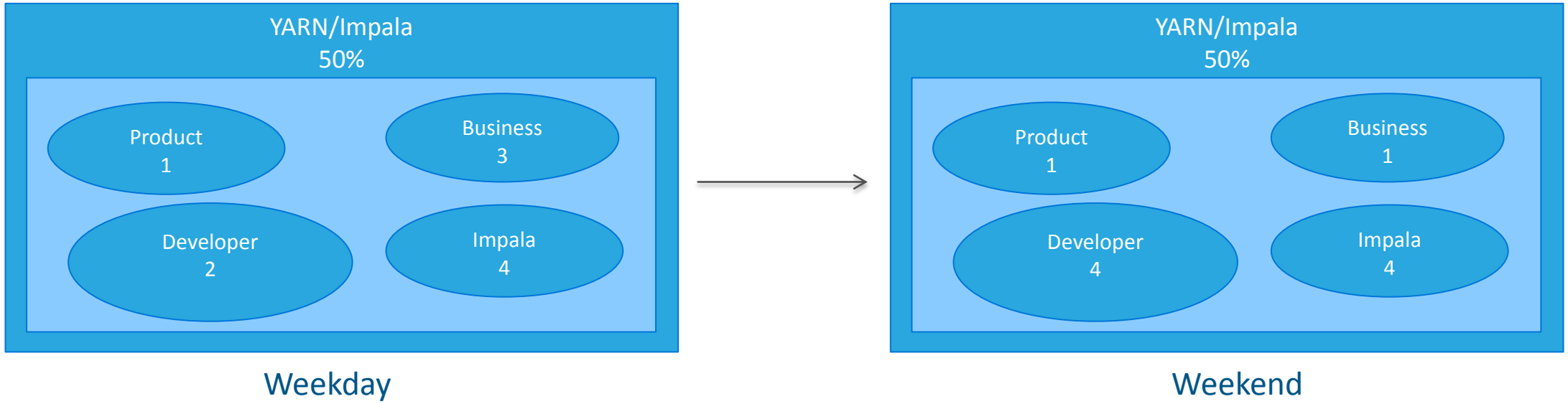




cloudera

动态资源管理

- 基于YARN的资源管理框架可以实现MapReduce, Spark以及Impala对资源的共享
  - 通过Llama实现Impala和YARN资源的集成
  - 按租户的资源使用状况定期调整资源分配策略



cloudera

资源使用状况统计

- 租户对于资源的历史使用统计和趋势，以更好满足企业内部的Showback和Chargeback模式

Reports

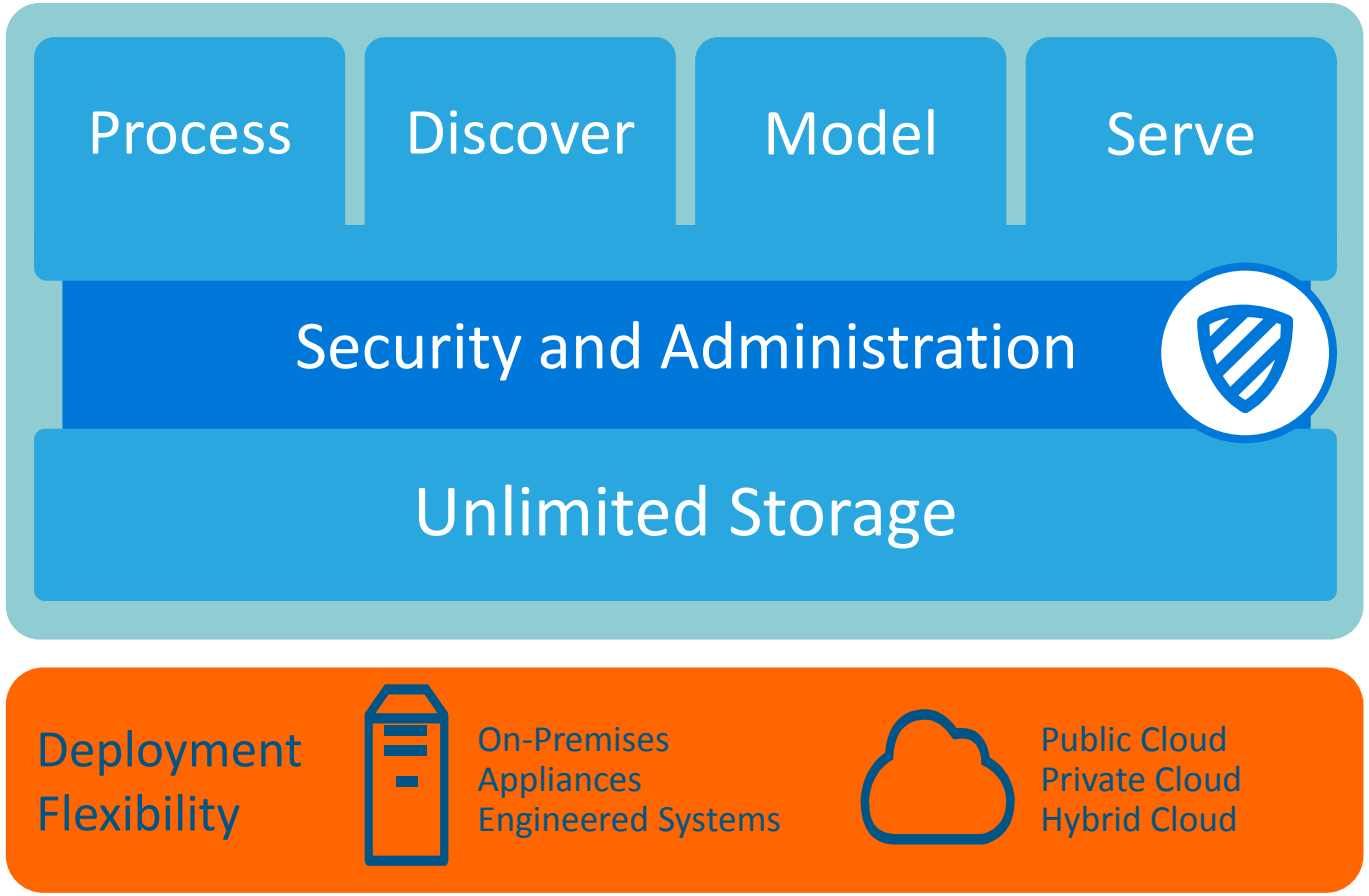
Cluster 1

Disk Usage (hdfs)	
Title	**Download**
Nameservice: nameservice1	
Current Disk Usage By User	CSV XLS
Current Disk Usage By Group	CSV XLS
Current Disk Usage By Directory	CSV XLS
Historical Disk Usage By User	CSV XLS
Historical Disk Usage By Group	CSV XLS
Historical Disk Usage By Directory	CSV XLS
Activities (mapreduce)	
Title	**Download**
MapReduce Usage by User	CSV XLS
User Access (hdfs)	
Title	**Download**
Nameservice: nameservice1	

Bring Cloudera Platform to Cloud – Cloudera Director

cloudera

Cloudera Director



Infrastructure Design:

- Cloud Strategy
- Reduce time to services

Low TCO, Time to Value

- Data in Cloud
- Workload in Cloud

Temporary Relief

- Ad-hoc/non-continuous services
- End-user self-service

cloudera

Cloudera Director

Portability: Multiple Deployment Options

Private Cloud Physical  

Public Cloud   Windows Azure

Flexibility: Pricing and Support

- Traditional licensing with Cloudera support
- Usage-based pricing with Cloudera + cloud vendor support

Choice: Growing Ecosystem

Rapidly expanding cloud provider and MSP ecosystem for choice in cloud-based services

cloudera

Ensure Customer Success – Industry-Leading Support

cloudera

Cloudera技术支持

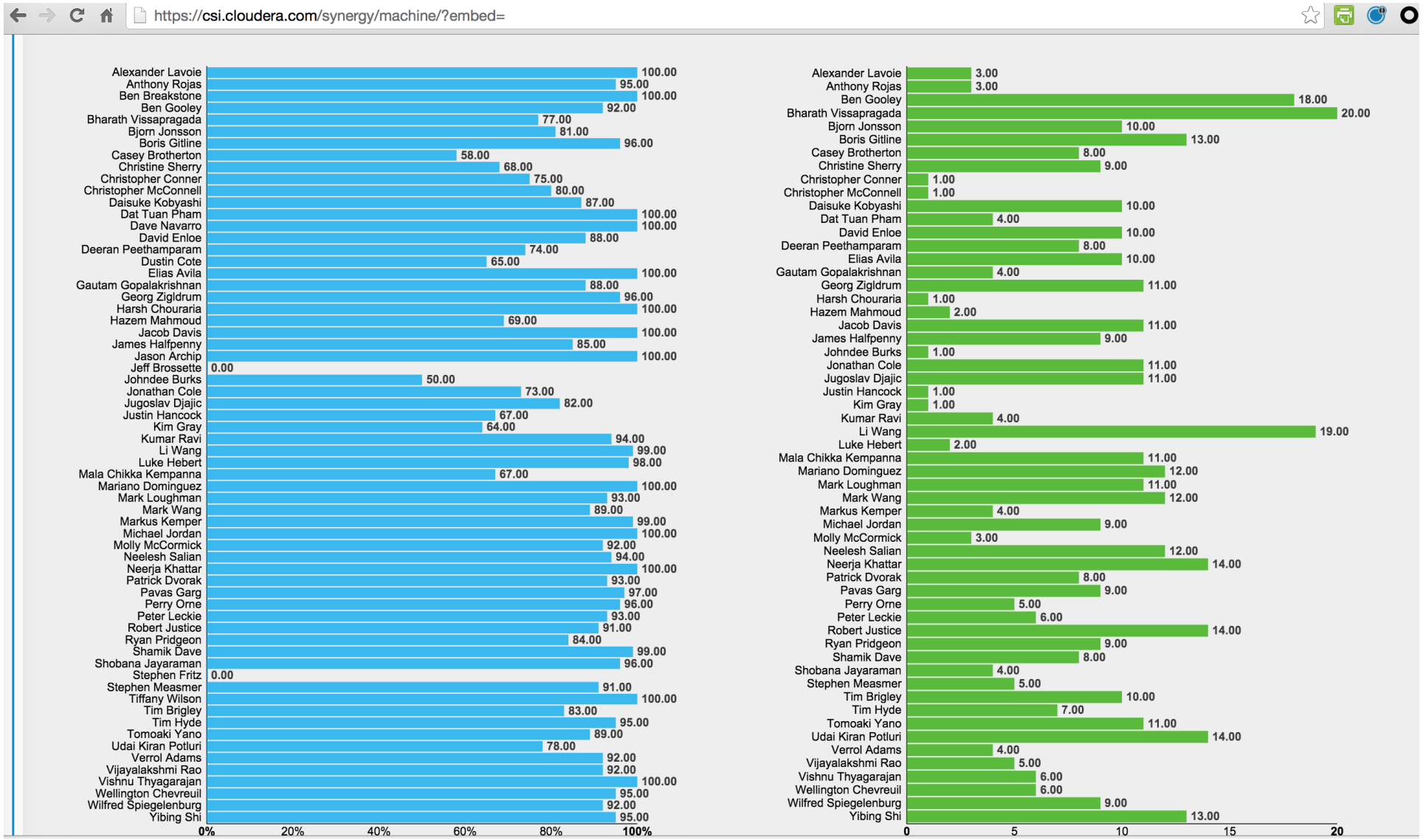
- 专业服务
- 近百人的专业技术支持团队
- 丰富的知识库
- 基于大数据技术的预测支持及主动支持
- 严格的问题修复流程

专业服务

- 预定义的企业服务内容
- 驻场架构师和专人技术支持

cloudera

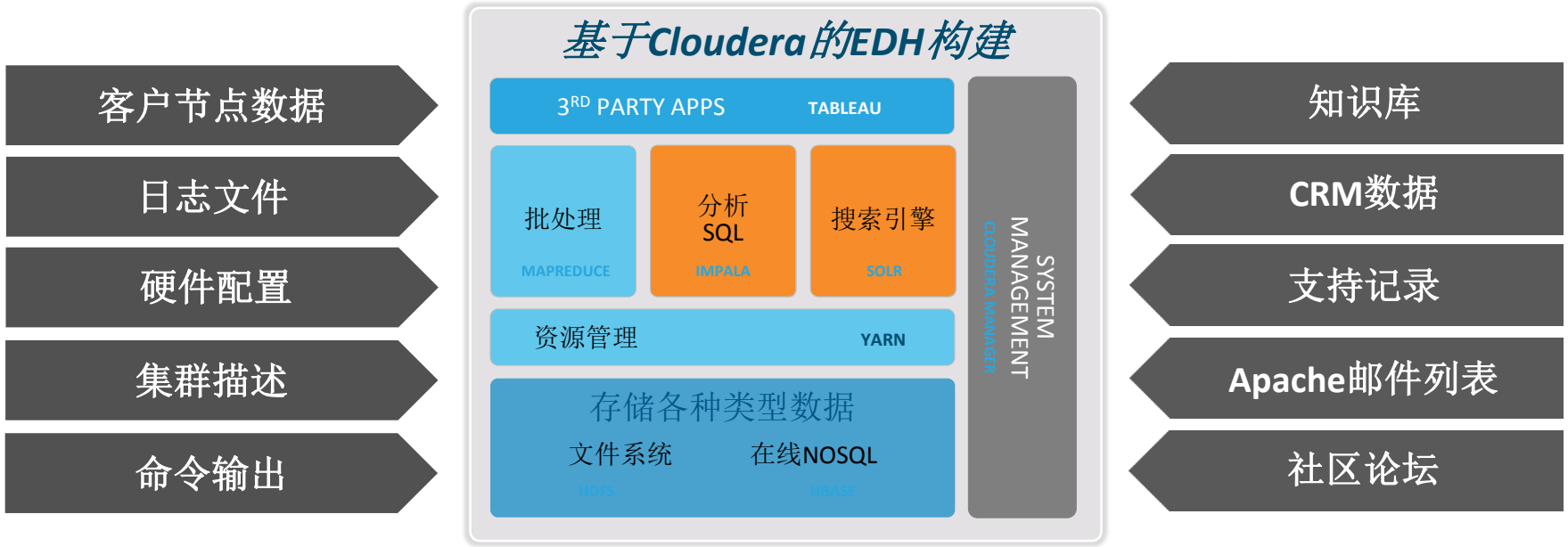
Cloudera客户支持中心（CSI）



cloudera

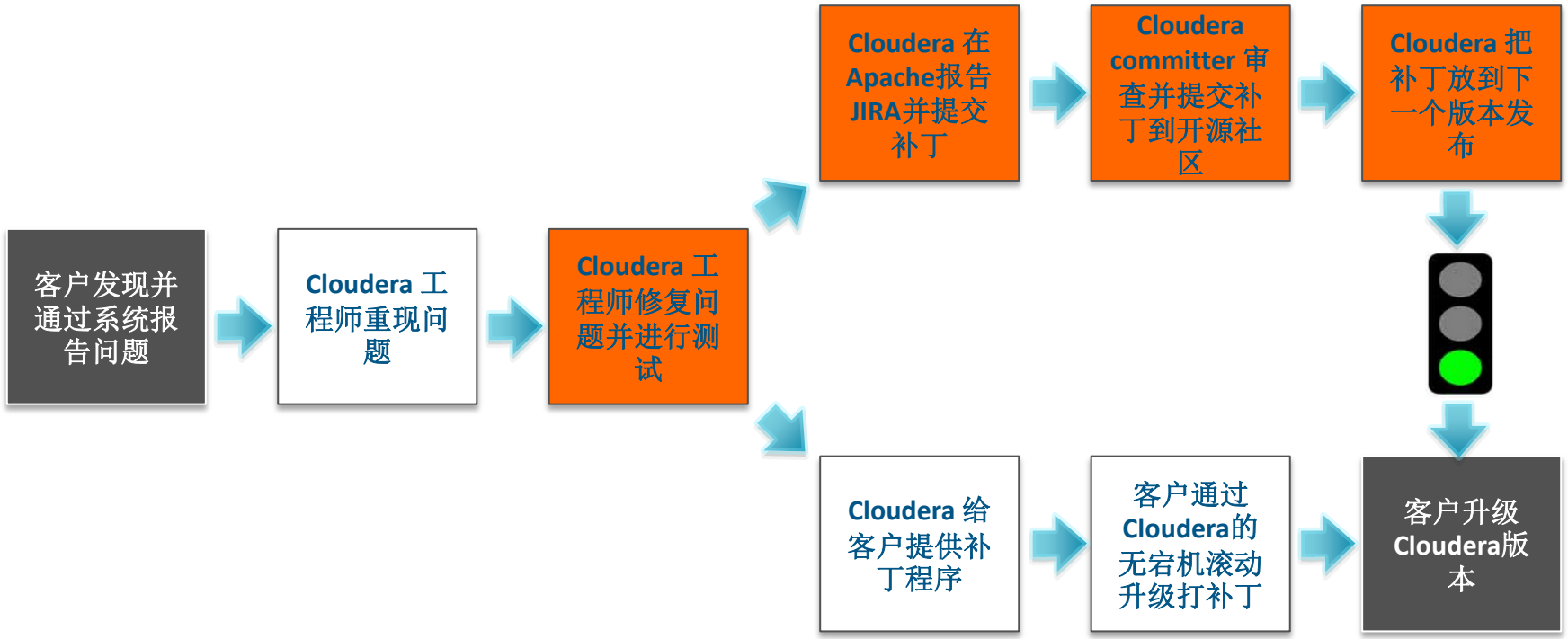
预测、主动技术支持

- 利用大数据平台技术，在客户集群还没发生问题之前就可以得到主动的预警
- 付费客户可以定期向Cloudera支持中心发送集群诊断包以获取主动支持



cloudera

严格的问题修复过程

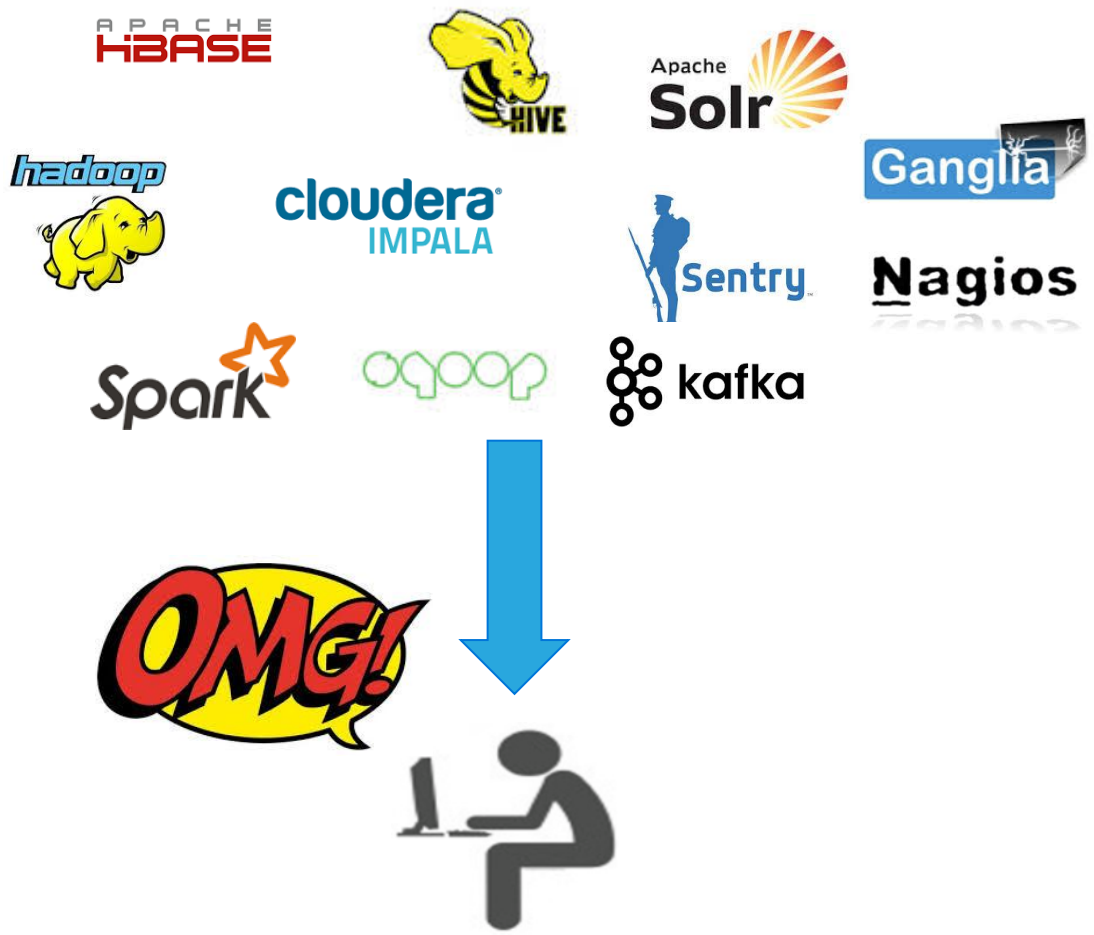


总结

- Cloudera Enterprise
  - 专注于开源Hadoop的开发，保护用户的投资（Open Standard）
  - 最具创新的Hadoop发行版 （Innovation）
  - 最好用的企业数据平台（Usability）
    - ✓活跃的Cloudera社区
    - ✓一站式的管理平台
  - 最完善的安全架构（Security）
  - 方便集成（Integration）
  - 全面可扩展（Extensibility）
  - 最专业的、可持续的技术支持

cloudera

与Apache开源项目比



1 集成性：包含了20多个开源项目，组件版本的兼容，解决了组件内部的配置和组件间的配置集成

2 管理性：自动化的安装部署；智能的配置优化；超级易用的监控诊断；企业级的管理能力；基础设施的集成

3 安全性：全面的安全技术架构；独有的主数据管理能力帮助用户快速发现数据并理解数据处理流程

4 技术支持：开源项目的发布周期不定，Cloudera会定期发布问题修复版，并提供快速的问题修复；同时开源项目有时会破坏一些兼容性

与社区版对比

	Cloudera Express	Cloudera Enterprise	
平台核心	CDH	CDH	包含数据采集、存储、处理和分析等组件
管理性	基本的安装、部署、监控、告警等管理功能	还包含一系列企业级功能： 配置历史修改和回退 平台运营历史报告 零宕机重启、升级 备份和复制 定期诊断等等	1. 不需要花大把的时间去查看由于配置修改导致的性能下降 2. 降低关键业务宕机的风险 3. 定期的诊断快照缩短解决问题的周期 4. 无意的数据损坏
安全性	有限的、松散的安全特性	自动化的Kerberos部署 统一访问权限控制 全面的审计 整体的数据保护解决方案	1. 发现恶意的访问 2. 防止系统管理员直接通过底层文件系统去读取敏感数据
数据治理	无	集群元数据的管理 数据溯源	1. 理解集群中有什么数据，快速发现数据 2. 数据的依赖关系，理解报表依赖的数据源
技术支持	无	主动的集群诊断、产品支持团队、 客户可以访问的知识库、专业技术服务 定期的平台缺陷通知、路线图	1. 需要花费大量的时间来优化集群来满足业务需求 2. 系统持续稳定运行的技术保障

cloudera

与闭源厂商对比

	闭源平台	Cloudera Enterprise	
平台核心	Unknown	CDH	闭源的组件或者功能缺乏和开源的持续兼容；闭源特性没有社区支持增加了用户使用代价
管理性	基本的安装、部署、监控、告警等管理功能	业界最好用，完全为Hadoop而开发的管理工具Cloudera Manager	
安全性	有限的、松散的安全特性	全面的安全解决方案，业界唯一一个符合PCI (Payment Card Industry)安全标准的平台	
数据治理	无	集群元数据的管理 数据溯源	
技术支持	有但不可持续	专业的产品支持团队，严格的问题修复流程，主动的集群诊断和预测支持	

版本和服务

- 免费版（Cloudera Express）和按年订阅的付费版（Cloudera Enterprise）
- 免费版包含CDH和功能受限的Cloudera Manager
- 付费版可以使用Cloudera Enterprise的所有功能，但根据可以享受的服务内容不一样
  - Basic Edition： 只提供Hadoop核心和Cloudera Director的服务
  - Flex Edition： HBase/Search/Impala/Spark/Navigator选择其一
  - Data Hub Edition： 所有组件都有服务提供
- Basic Edition只有5x8或7x24的标准支持
- Flex Edition和Data Hub Edition有5x8或7x24 Premium支持可选

cloudera

许可证模式

- Cloudera不提供永久的许可证
- Cloudera产品采取的是按年订阅许可证模式，假设用户订阅了三年的，则具体的付费方式根据客户要求：
  - 一次性付费
  - 按三年平均，分三次付
  - 第一年可以付大部分费用，后两年以维保的名义付费
- 订阅期结束之后，如果用户不再续订，则原有的功能都可以继续使用（包括付费版才有的功能）
- 订阅期结束之后，如果用户需要续订，则视为一次新的订阅期，此次订阅的价格会视前次订阅的周期和本次订阅的周期酌情考虑

迅速体验

- Cloudera Express – 完全免费
  - 全功能数据平台（CDH），无存储容量和节点数限制
  - 一站式的管理工具（Cloudera Manager）
  - 获取社区支持Cloudera Community
- Cloudera Enterprise Trial
  - 企业版60天的试用
  - 获取试用版许可证，得到专业的技术支持
- Cloudera Live
  - 在线的数据分析体验（Hue，Tableau，Zoomdata，Trifacta）

cloudera

资源

- Cloudera Product - http://www.cloudera.com/content/cloudera/en/downloads.html
- Cloudera Live - http://www.cloudera.com/content/cloudera/en/products-and-services/cloudera-live.html
- Cloudera Community - http://community.cloudera.com/
- Cloudera Documentation - http://www.cloudera.com/content/cloudera/en/documentation.html

cloudera

cloudera

Thank you

@Contact Info

Backup

cloudera

灵活的版本选择（1）

	CLOUDERA EXPRESS	CLOUDERA ENTERPRISE		
		Basic Edition	Flex Edition	Data Hub Edition
许可证	免费	按年订阅		
100% 开源的数据存储及处理平台（CDH）				
Hadoop, Flume, HBase, Hcatalog, Hive, Hue, Impala, Mahout, Oozie, Pig, Cloudera Search, Sentry, Spark, Sqoop, Whirr, Zookeeper	✔	✔	✔	✔
系统管理平台（Cloudera Manager）				
集群部署和配置	✔	✔	✔	✔
服务管理	✔	✔	✔	✔
服务和主机监控	✔	✔	✔	✔
安全管理	✔	✔	✔	✔
诊断（日志搜索、事件）	✔	✔	✔	✔
扩展和Rest API	✔	✔	✔	✔
滚动升级和重启		✔	✔	✔

cloudera

灵活的版本选择（2）

	CLOUDERA EXPRESS	CLOUDERA ENTERPRISE		
		Basic Edition	Flex Edition	Data Hub Edition
AD/Kerberos集成		✔	✔	✔
SNMP支持		✔	✔	✔
LDAP集成		✔	✔	✔
参数配置历史和回滚		✔	✔	✔
运营报告生成		✔	✔	✔
定期诊断		✔	✔	✔
自动化复制和灾备		✔	✔	✔

cloudera

灵活的版本选择（3）

	CLOUDERA EXPRESS	CLOUDERA ENTERPRISE		
		Basic Edition	Flex Edition	Data Hub Edition
产品售后服务覆盖				
Hadoop核心		✔	✔	✔
Cloudera Director		✔	✔	✔
Online NoSQL RDBMS (HBase)			只能选择一种组件提供支持	✔
交互式SQL (Impala)				✔
交互式数据分析 (Apache Spark)				✔
搜索引擎 (Cloudera Search)				✔
审计、数据发现、溯源、加解密、密钥管理 (Cloudera Navigator)				✔
敏捷部署模块				
Cloudera Director	✔	✔	✔	✔

cloudera

灵活的版本选择（4）

	CLOUDERA EXPRESS	CLOUDERA ENTERPRISE		
		Basic Edition	Flex Edition	Data Hub Edition
服务内容				
专职支持团队		✔	✔	✔
主动技术指导		✔	✔	✔
预测性问题分析		✔	✔	✔
全面的知识库		✔	✔	✔
产品解决方案和指南		✔	✔	✔
客户需求纳入新产品路线图		✔	✔	✔
5 x 8 或 7 x 24小时标准服务		✔	✔	✔
增强服务*			✔	✔

* 5x8或7x24服务时间内，对于严重的产品问题，15分钟内有响应

cloudera

Hadoop和Cloudera

- Cloudera创建了Hadoop生态
  - Doug Cutting是公司的首席架构师
- Cloudera开源了诸多Hadoop工具，现已形成了Hadoop生态链中的标准
  - 采集： Apache Flume, Apache Sqoop
  - 存储： HDFS, HBase, Parquet
  - 处理： MapReduce
  - 分析： Hive, Impala, Solr
  - 服务： Avro, Zookeeper, Sentry

开源模式

- 可以防止被某一个提供商绑定，在后期可能需要付出高昂的维护和技术支持费用
- 产品的稳定性更好，有更多的用户参与产品的使用和测试，使得产品存在的问题更少
- 安全性更好，有更多的人可以审查代码，任何代码的安全漏洞可以被很快地发现和修复
- 汇聚全球智慧，加速产品创新；没有任何一个提供商能够提供比社区更快、更全的产品更新
- 开源比闭源能更好地遵守开放标准，不受专有的数据存储和处理引擎限制，方便业务部门、企业间的互操作
- 企业可以通过多种渠道快速解决问题，培养团队的自我技能

企业需要开源Hadoop平台

- Hadoop及其生态的项目属性决定核心平台要开源，任何定制化开发最终损坏的是客户利益
- 开源是帮助客户解决问题的手段，不是目的
  - 对于任何开源项目的问题都能够以开源的方式解决，否则长此以往会和社区主流差异化越来越大，或者只能等下一个开源版本的发布
  - 有足够的解决开源问题的能力，这样可以更快的满足客户的需求
- 领导Hadoop作为企业级应用的缺陷功能定义和开发

  - HDFS HA, Short-circuit read, Network Encryption, HBase snapshots, Hive authentication, HDFS Caching, At-rest HDFS Encryption,…

Cloudera Committers by Apache Project

89 total seats, 67 PMC* seats (Page 1 of 2)

Project	Founder(s) Employed By:	Committers	Names (PMC Members in blue)
Accumulo	NSA	3	Mike Drob, Sean Busbey, Bill Havanki
Avro	Cloudera	5	Doug Cutting (Founder), Tom White, Jeff Hammerbacher, Philip Zeyliger, Ryan Blue
Bigtop	Cloudera -> Pivotal	9	Andrew Bayer, Eli Collins, Patrick Hunt, Tom White, Stephen Chu, Sean Mackrory, Michael Stack, Anatoli Fomenko, Mark Grover
Crunch	Cloudera	3	Josh Wills (VP/PMC Chair/Founder), Brock Noland, Tom White
Flume	Cloudera	10	Andrew Bayer, Hari Shreedharan, Brock Noland, Jarek Jarcec Cecho, Henry Robinson, Jon Hsieh (Project Founder), Mike Percy, Patrick Hunt, Prasad Mujumdar, Wolfgang Hoschek
Hadoop Core	Independent/Yahoo! -> Cloudera	14	Doug Cutting (Project Founder), Tom White, Todd Lipcon, Patrick Hunt, Eli Collins, Aaron Myers, Michael Stack, Colin McCabe, Andrew Wang, Karthik Kambatla, Harsh Chouraria, Sandy Ryza, Robert Kanter, Yongjun Zhang
HBase	Powerset -> Cloudera	10	Michael Stack (Project Co-founder/VP/PMC Chair), Todd Lipcon, Jon Hsieh, Lars George, Jean-Daniel Cryans, Jimmy Xiang, Matteo Bertozzi, Gregory Chanan, Misty Stanley-Jones, Sean Busbey
Hive	Facebook -> Cloudera/Qubole	5	Xuefu Zhang, Brock Noland, Prasad Mujumdar, Szehen Ho, Chao Sun

* PMC = Project Management Committee; guides project roadmap and direction

Cloudera Committers by Apache Project

89 total seats, 67 PMC* seats (Page 2 of 2)

Project	Founder(s) Employed By:	Committers	Names (PMC Members are in blue)
Lucene/Solr	Independent -> Cloudera	6	Doug Cutting (Founder), Mark Miller (VP/PMC Chair), Yonick Seeley, Erick Erickson, Wolfgang Hoschek, Greg Chanan
Mahout	Independent	1	Sean Owen
Oozie	Yahoo!	2	Harsh Chouraria, Robert Kanter
Pig	Yahoo! -> Hortonworks	2	Santhosh Srinivasan, Xuefu Zhang
Spark	Quantifind -> Cloudera	2	Imran Rashid, Sean Owen
Sqoop	Cloudera -> Independent	9	Andrew Bayer, Jarek Jarcec Cecho, Jon Hsieh, Kathleen Ting, Patrick Hunt, Tom White, Hari Shreedharan, Abe Elmahrek, Gwen Shapira
Whirr	Cloudera	6	Tom White (Founder), Lars George, Patrick Hunt, Andrew Bayer (VP/PMC Chair), Andrei Savu, Graham Gear
ZooKeeper	Yahoo! -> Cloudera	2	Patrick Hunt (Founder), Henry Robinson

* PMC = Project Management Committee; guides project roadmap and direction

How Customers Benefit from the CDH Life Cycle

- They can confidently access new Apache releases that are certified after extensive testing and integration work.
- They can count on their issues being fixed permanently upstream.
- They can access the most critical new upstream bug fixes and innovations at a regular cadence, between Apache releases.
- Cross-compatibility and stability are ensured across releases, as well as with upstream project trunks (which ensures application portability).
- Upgrades are much easier.

cloudera

Hadoop 1.x vs. Hadoop 2.x

- HDFS
  - Hive Availability – NameNode HA
  - Further Scalability – HDFS Federation
  - Usability – HDFS Rolling Upgrade/NFSv3 Access to HDFS
  - Performance – DataNode Caching/Heterogeneous Storage Hierarchy
  - Security – Fine-grained Access Control/HDFS Snapshots/HDFS At-Rest Encryption
- YARN
  - YARN HA, Rolling Upgrade
  - A new workload management
  - Support not only MapReduce engine, but also other computation engines, like spark

Vender Differentiation

- Comparison Dimensions
  - Product
  - Business Model
  - Team
  - Support
  - History Tracking

Overall Competition with Other Vendors

- Product (HortonWorks, IBM BigInsights, Huawei, Transwarp, Open Source)
  - **Hadoop Distribution** - CDH is the most widely adopted Hadoop platform, which features as an open, scalable, integrated, flexible, compatible secure and high available
  - **System Management** - Cloudera Manager is the most advanced system management software built natively for Hadoop, and has a lot of great features to support customers' business continuity
  - **Data Management** - Cloudera Navigator is the end-to-end data management software for Data Lineage, Audit, Data Lifecycle and Data Discovery
  - **Cloud Deployment** - Cloudera Director can implement customers' Big Data strategy with your existing or future Cloud platform

cloudera

Overall Competition with Other Vendors

- Business Model (IBM, Huawei, Transwarp)
  - Development in open source ensures the fast innovation and quality
  - Hadoop Customization leads to lock-in and incompatibility
- Team (HortonWorks, IBM, Huawei, Transwarp)
  - Cloudera employs the most open source core project committers (89) to support the its open strategy
- Support (HortonWorks, IBM, Huawei, Transwarp)
  - Dedicated support team (Customer Operation Engineering) & Customer Centric Engineering) working on open source project to ensure customer success
- History Tracking (HortonWorks, IBM, Huawei, Transwarp)
  - CDH Core prevalence and the customer installation base

cloudera

Hadoop 1.x vs. Hadoop 2.x

- HDFS
  - High Availability – NameNode HA
  - Further Scalability – HDFS Federation
  - Usability – HDFS Rolling Upgrade/NFSv3 Access to HDFS
  - Performance – DataNode Caching/Heterogeneous Storage Hierarchy
  - Security – Fine-grained Access Control/HDFS Snapshots/HDFS At-Rest Encryption
- YARN
  - YARN HA, Rolling Upgrade
  - A new workload management
  - Support not only MapReduce engine, but also other computation engines, like spark

Vender Differentiation

- Comparison Dimensions
  - Product
  - Business Model
  - Team
  - Support
  - History Tracking

cloudera

Overall Competition with Other Vendors

- Product (HortonWorks, IBM BigInsights, Huawei, Transwarp, Open Source)
  - **Hadoop Distribution** - CDH is the most widely adopted Hadoop platform, which features as an open, scalable, integrated, flexible, compatible secure and high available
  - **System Management** - Cloudera Manager is the most advanced system management software built natively for Hadoop, and has a lot of great features to support customers' business continuity
  - **Data Management** - Cloudera Navigator is the end-to-end data management software for Data Lineage, Audit, Data Lifecycle and Data Discovery
  - **Cloud Deployment** - Cloudera Director can implement customers' Big Data strategy with your existing or future Cloud platform

Overall Competition with Other Vendors

- Business Model (IBM, Huawei, Transwarp)
  - Development in open source ensures the fast innovation and quality
  - Hadoop Customization leads to lock-in and incompatibility
- Team (HortonWorks, IBM, Huawei, Transwarp)
  - Cloudera employs the most open source core project committers (89) to support the its open strategy
- Support (HortonWorks, IBM, Huawei, Transwarp)
  - Dedicated support team (Customer Operation Engineering) & Customer Centric Engineering) working on open source project to ensure customer success
- History Tracking (HortonWorks, IBM, Huawei, Transwarp)
  - CDH Core prevalence and the customer installation base

cloudera

cloudera

Thank you

