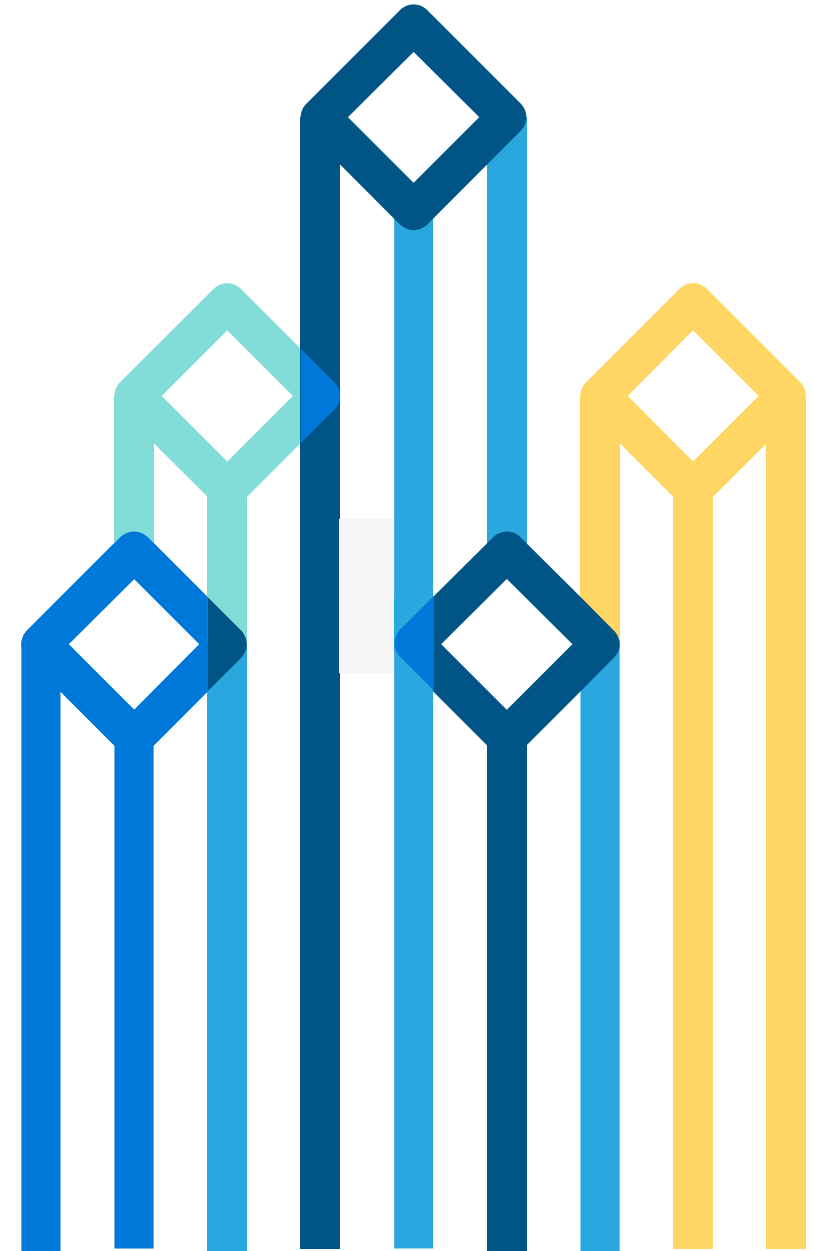




Cloudera Enterprise Introduction

-- 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

无所不在的数据



物联网及智能终端数据



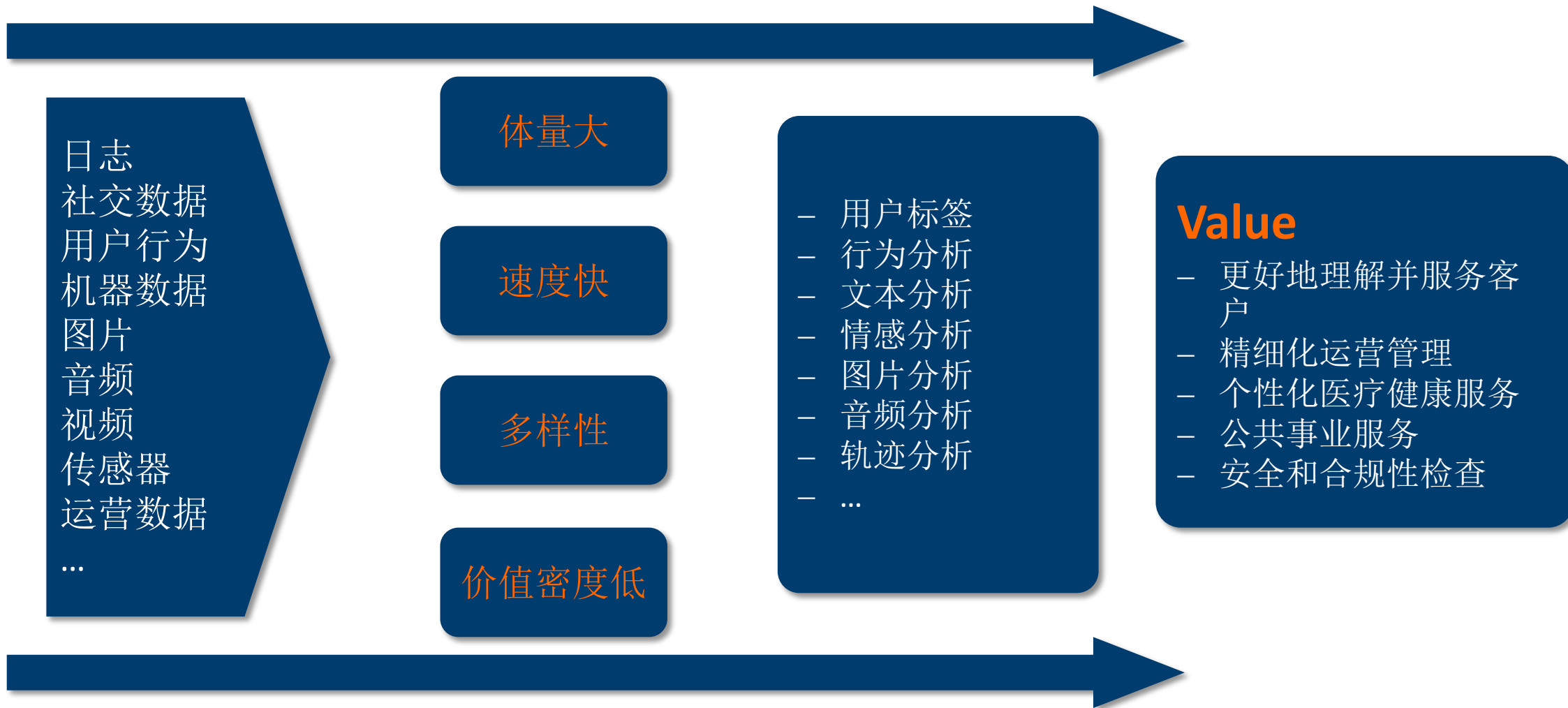
用户交互行为数据



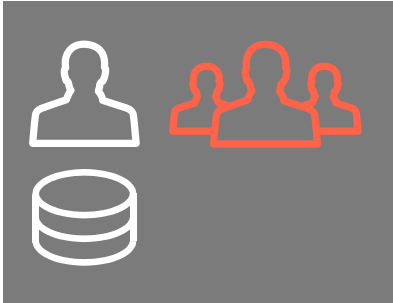
企业运营数据

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。

数据价值挖掘



传统架构的劣势



Limited Insights

Power users struggle with data.
Many users have **no data**.



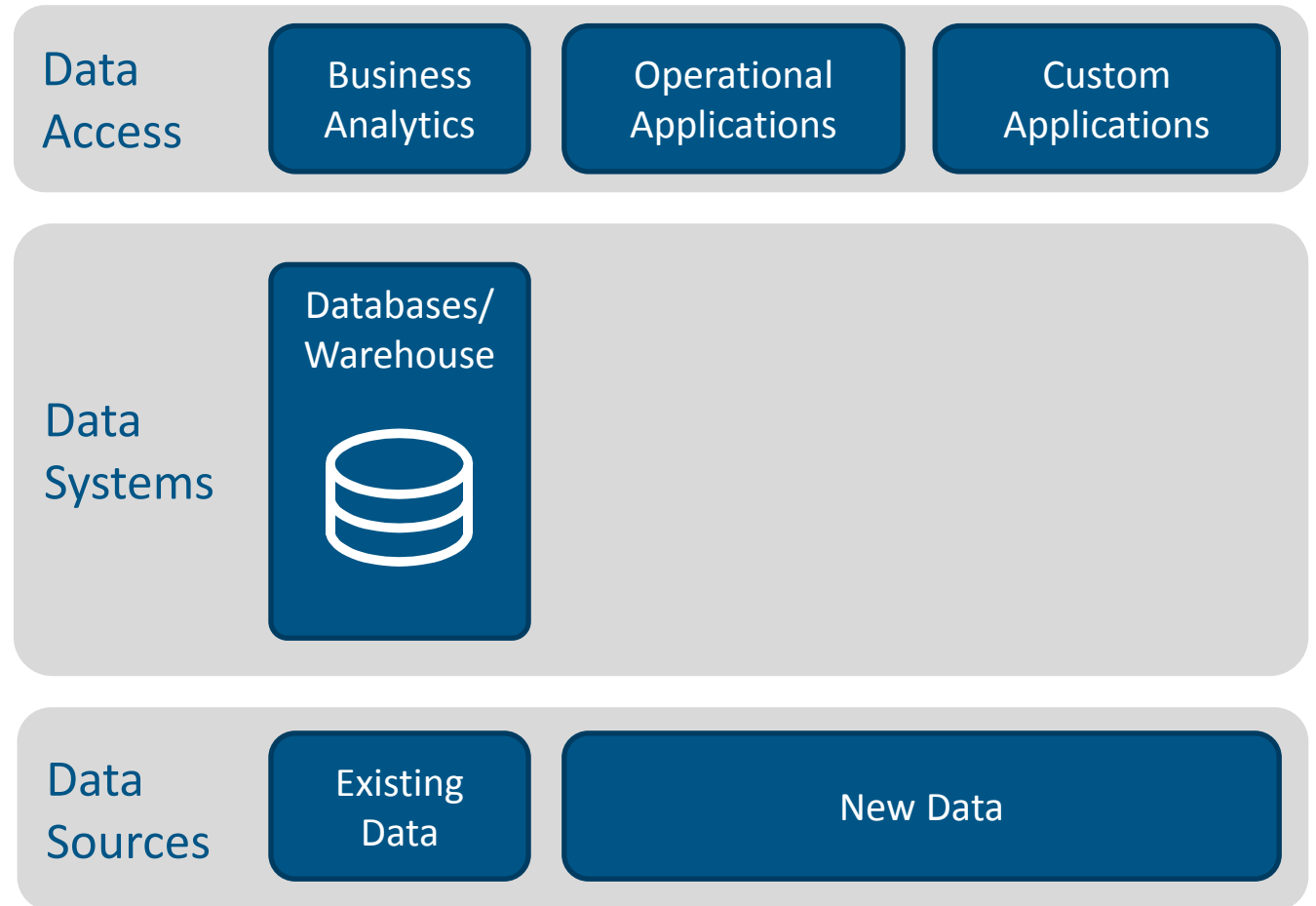
Compliance and Privacy

More data, more users, and more tools create **complexity**.
Need to balance business agility with **security** and **governance**.

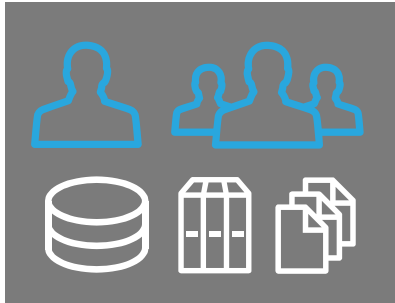


Limited Data

Not efficient to keep existing data, let alone handle **new data sources**.
Time consuming to transform data for analysis in existing systems.



亟需新的数据平台架构



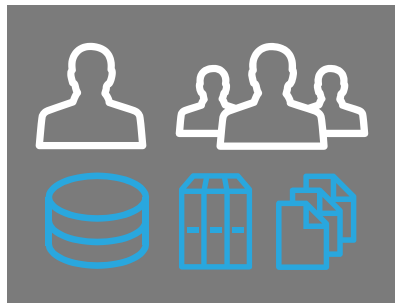
Unlock Value from Data

From analytics for some, to **insights** for all.



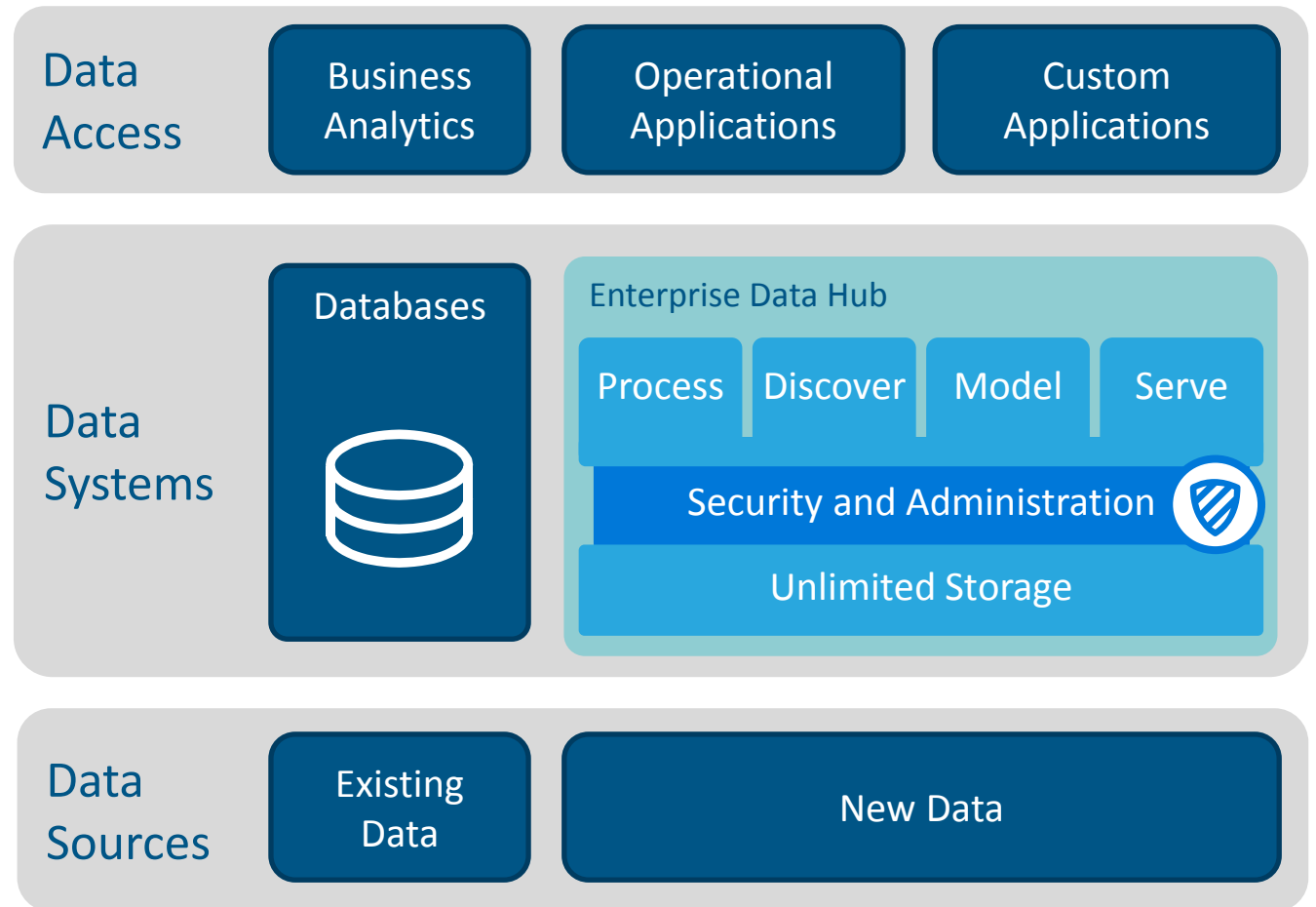
Manage Compliance

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

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

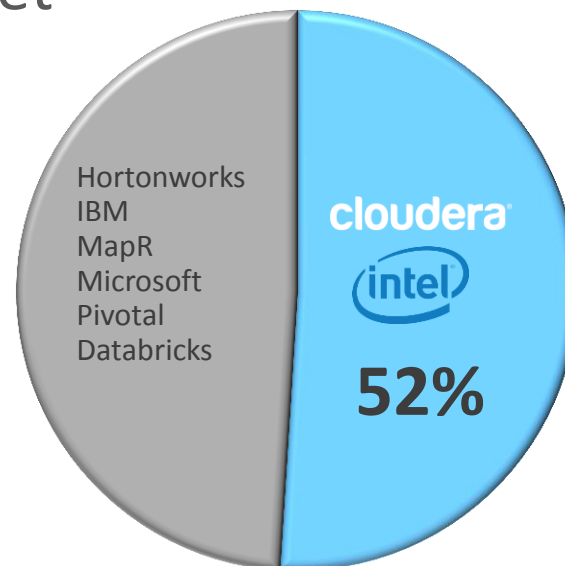
企业数据中心的革新

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



Projects Included:

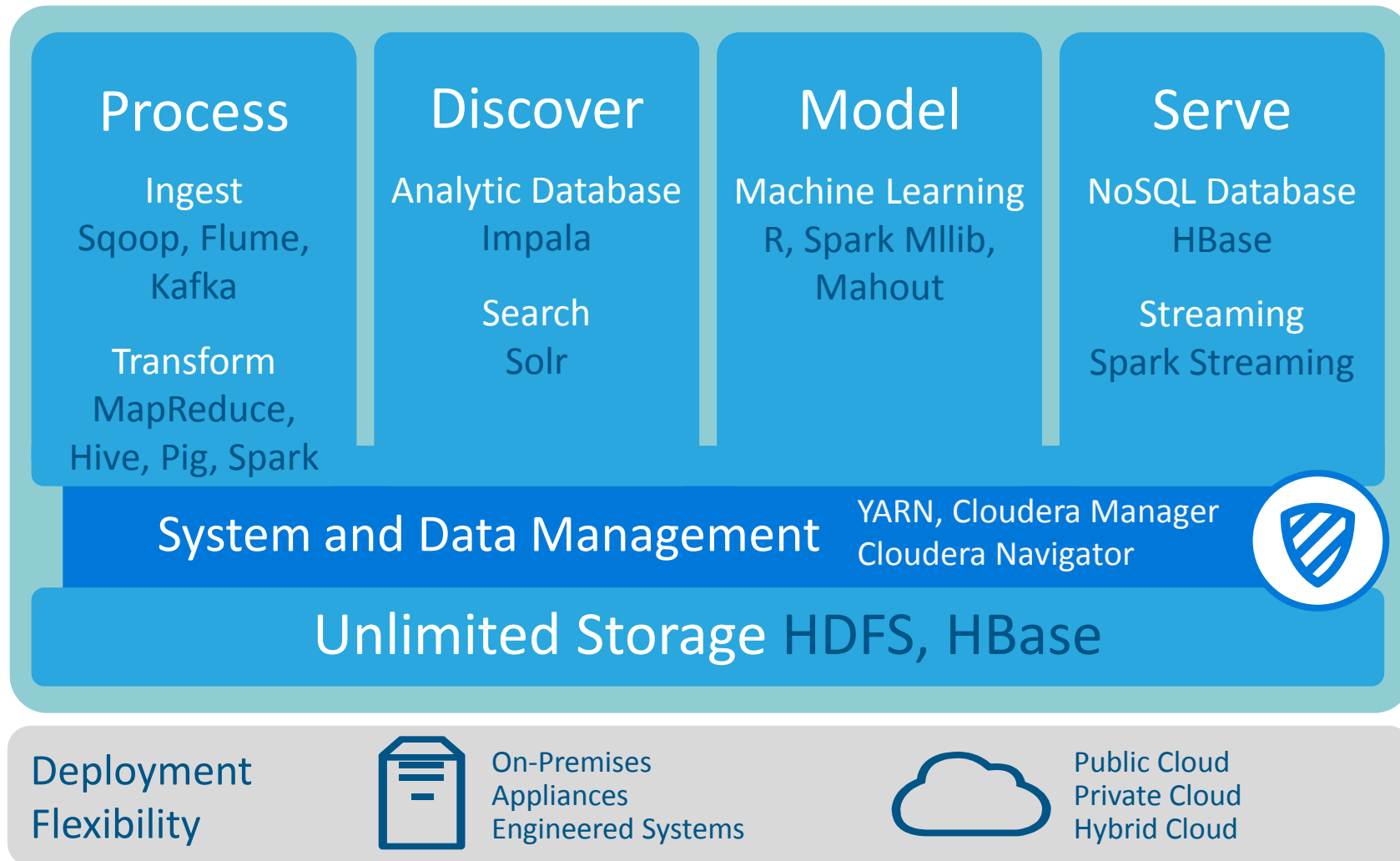
Accumulo	Mahout
Avro	Oozie
Bigtop	Pig
Crunch	Solr
Flume	Spark
Hadoop Core	Sqoop
HBase	Tez
Hive	ZooKeeper
Kafka	

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 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

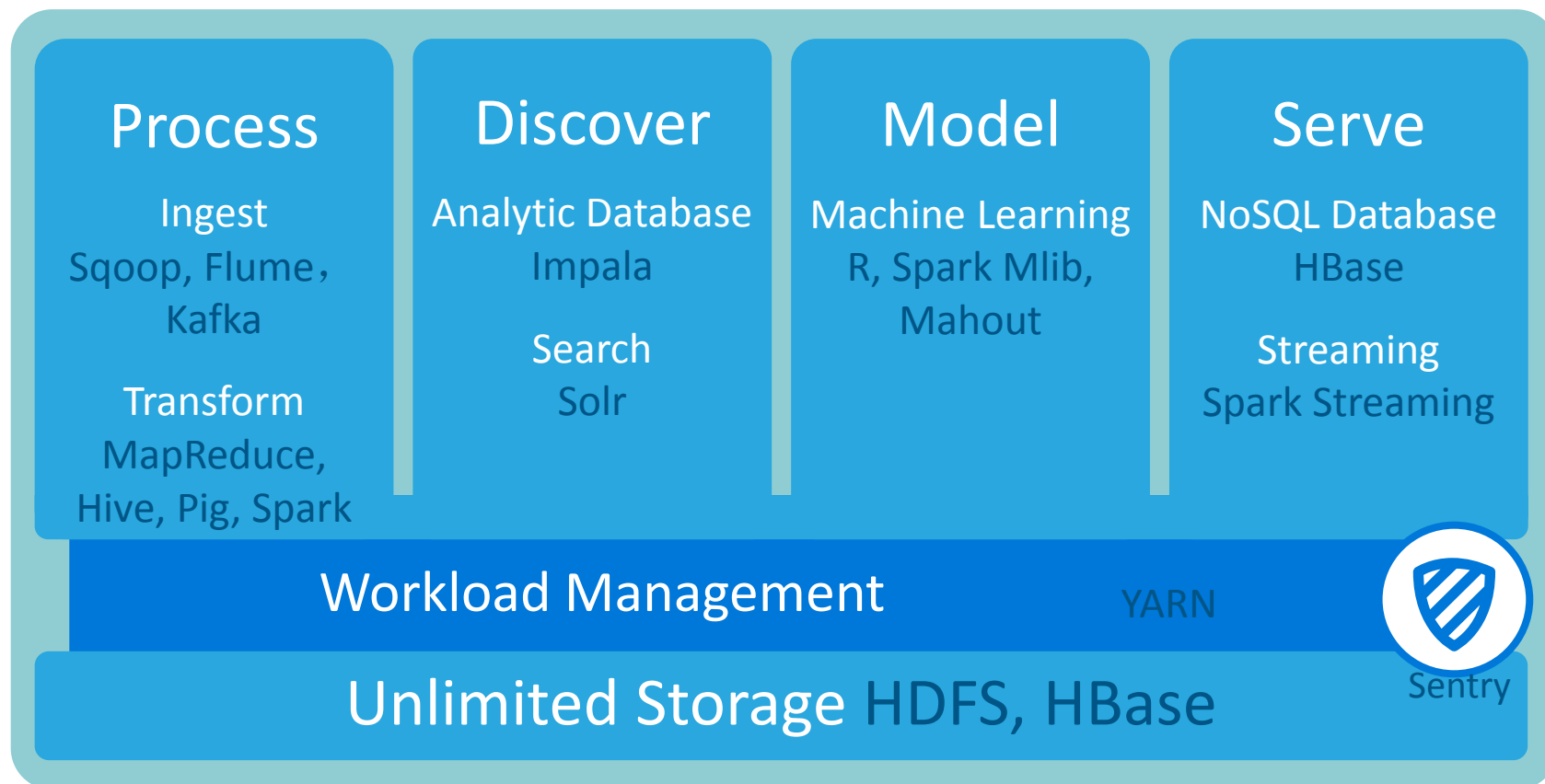
企业级数据平台

- 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

The Open Source Platform - CDH

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

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

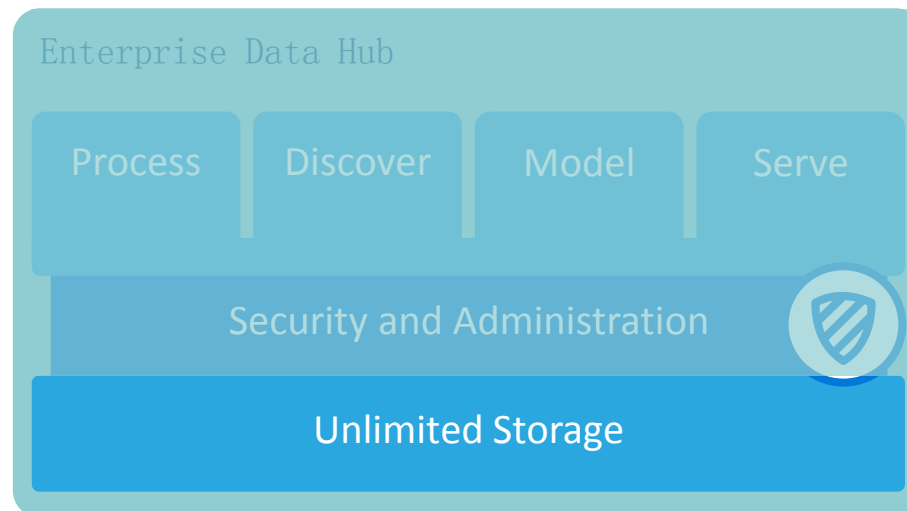
高容错性

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



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

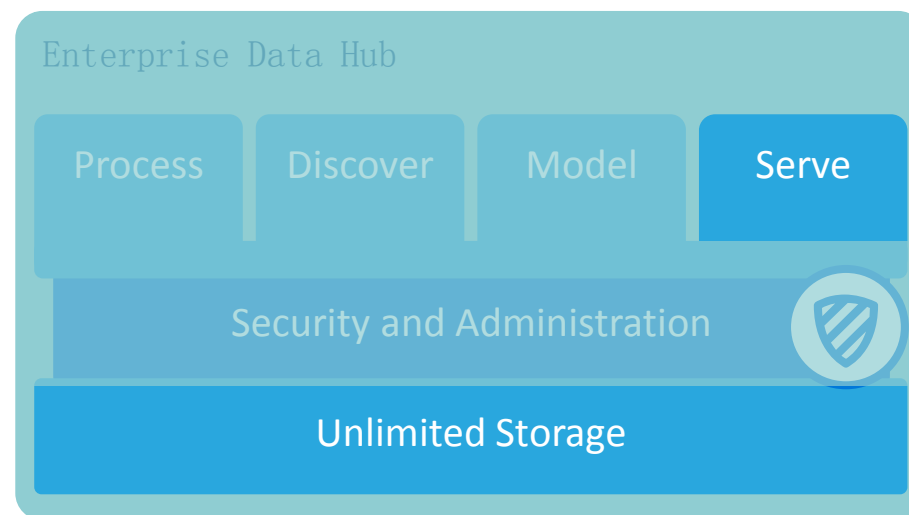
灵活性 多样化数据的统一存储

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

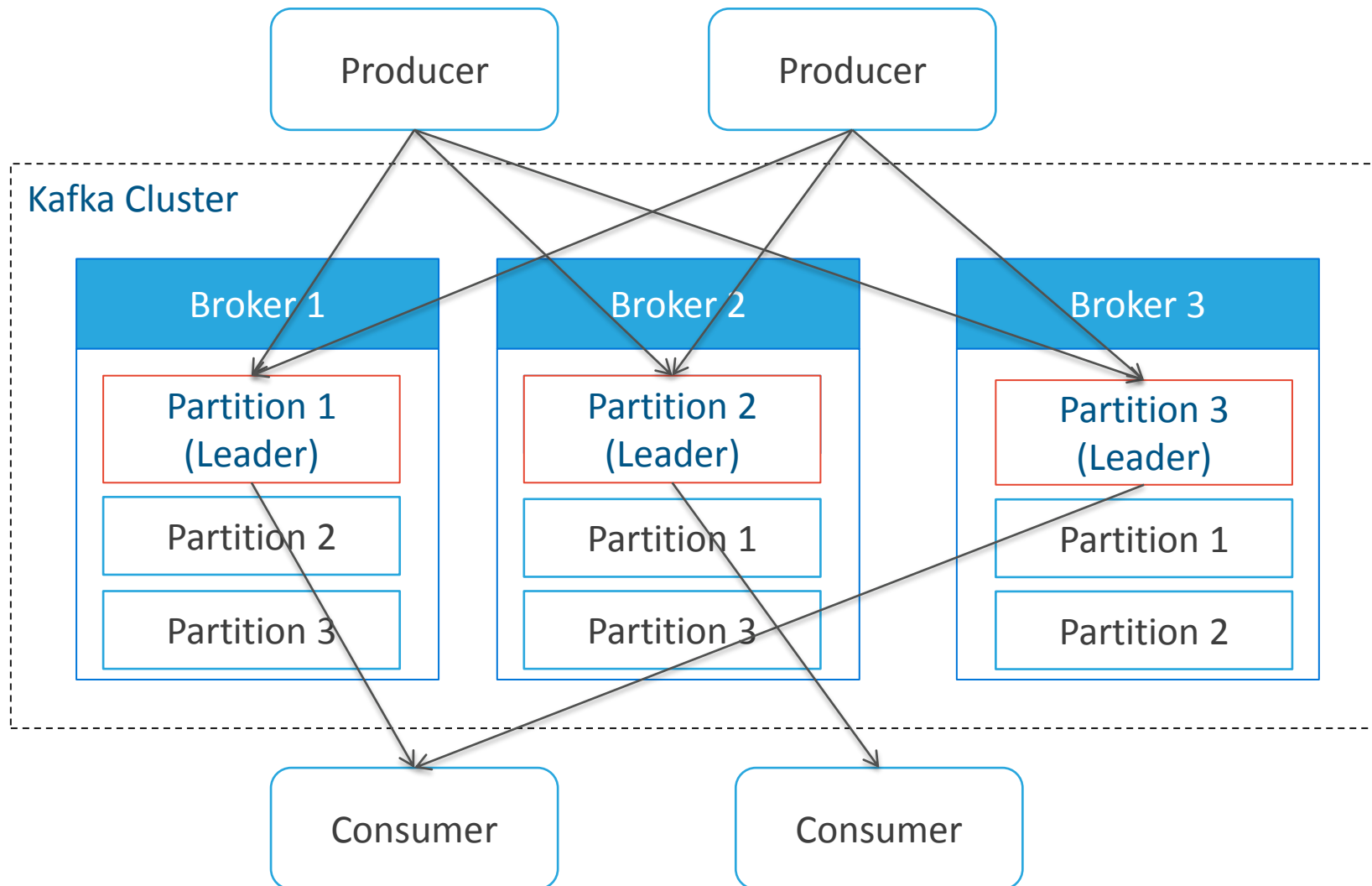
+ 在线数据服务

和HDFS紧密结合，适合高并发
随机读写



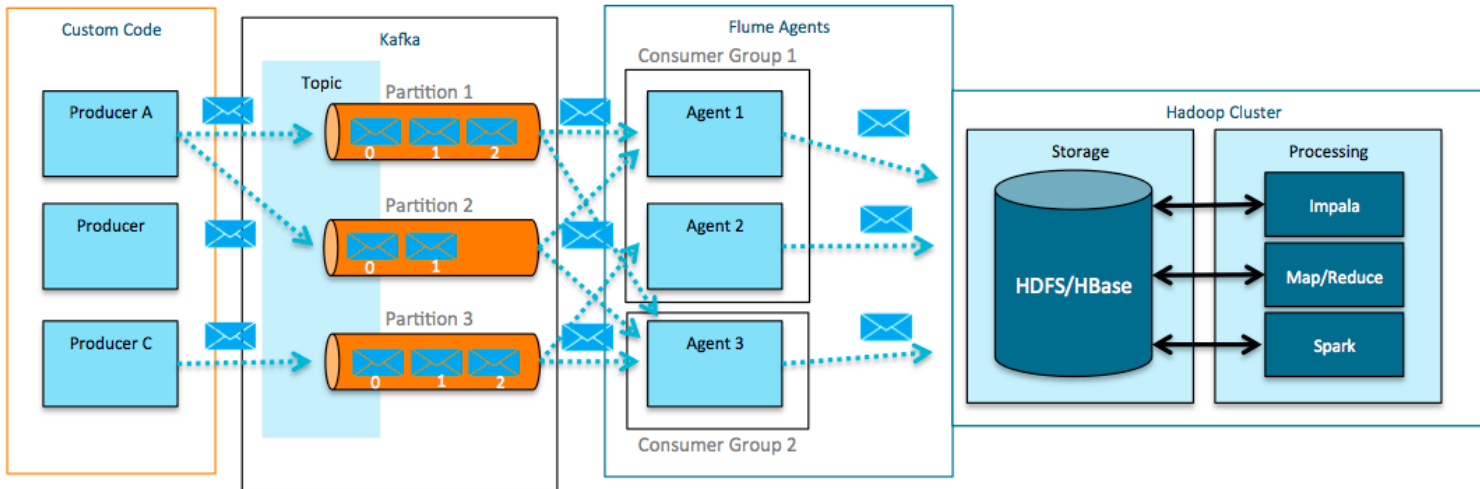
Apache Kafka

- 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



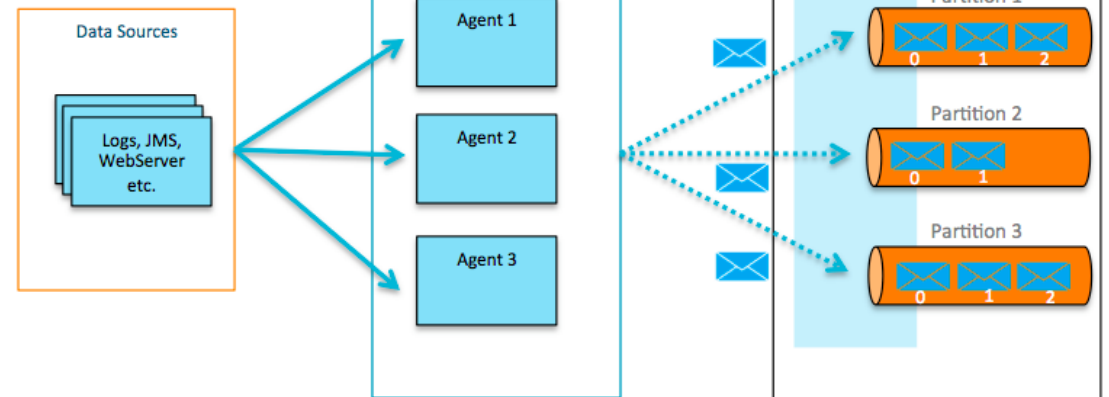
Kafka + Apache Flume

- 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



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

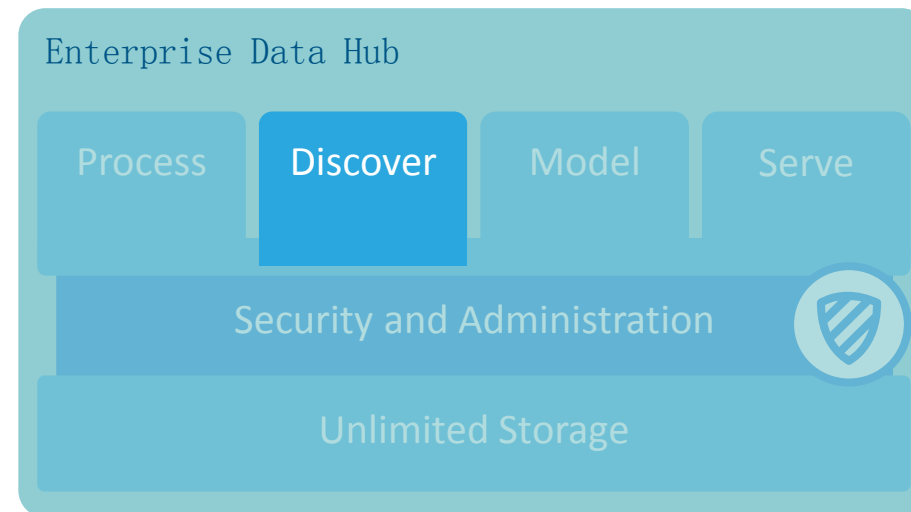


交互性

提供交互式的体验

原生

和Hadoop栈深度融合



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

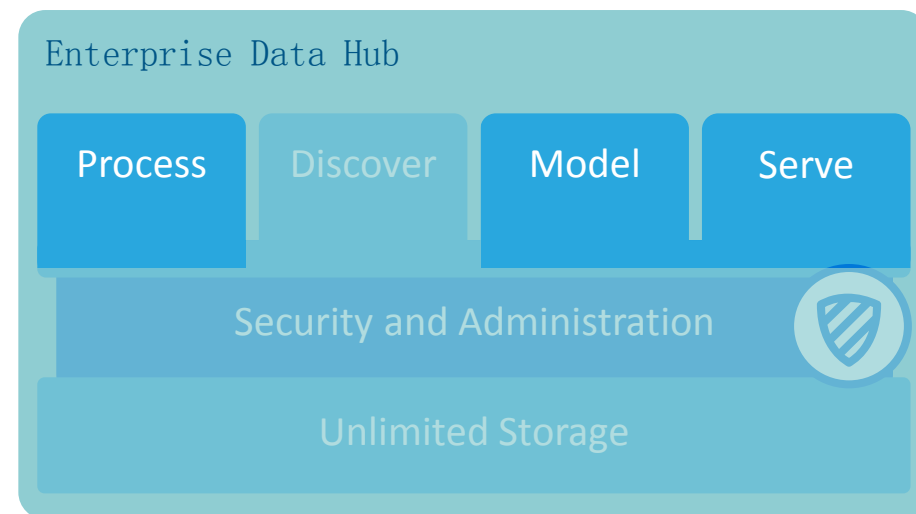
易用

好用且丰富的API

安全集成

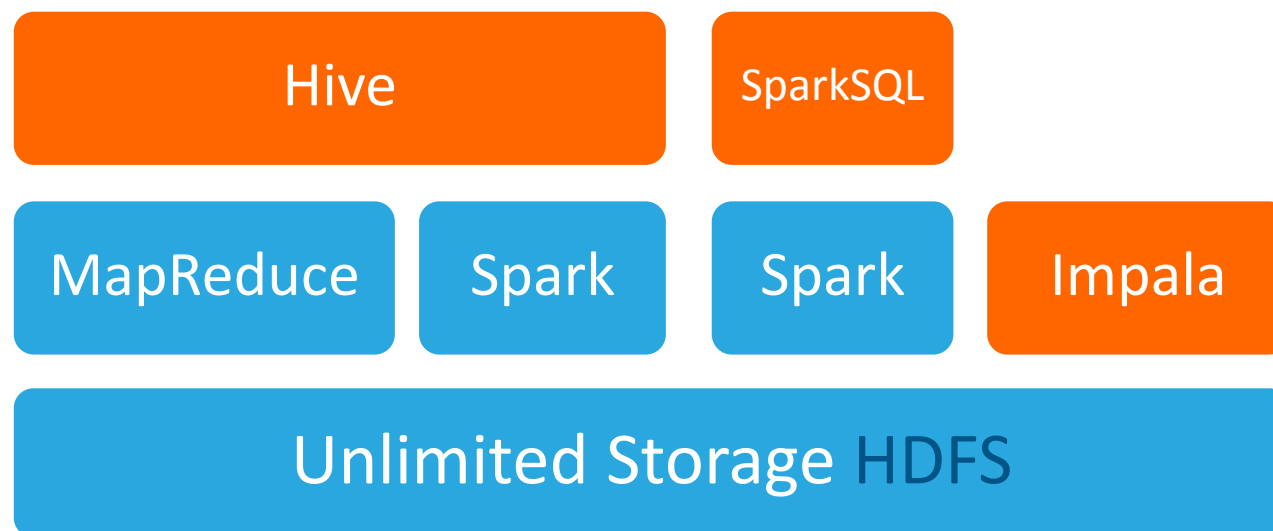
和数据平台的其他功能无缝集成

适合批处理、流计算以及迭代式计算



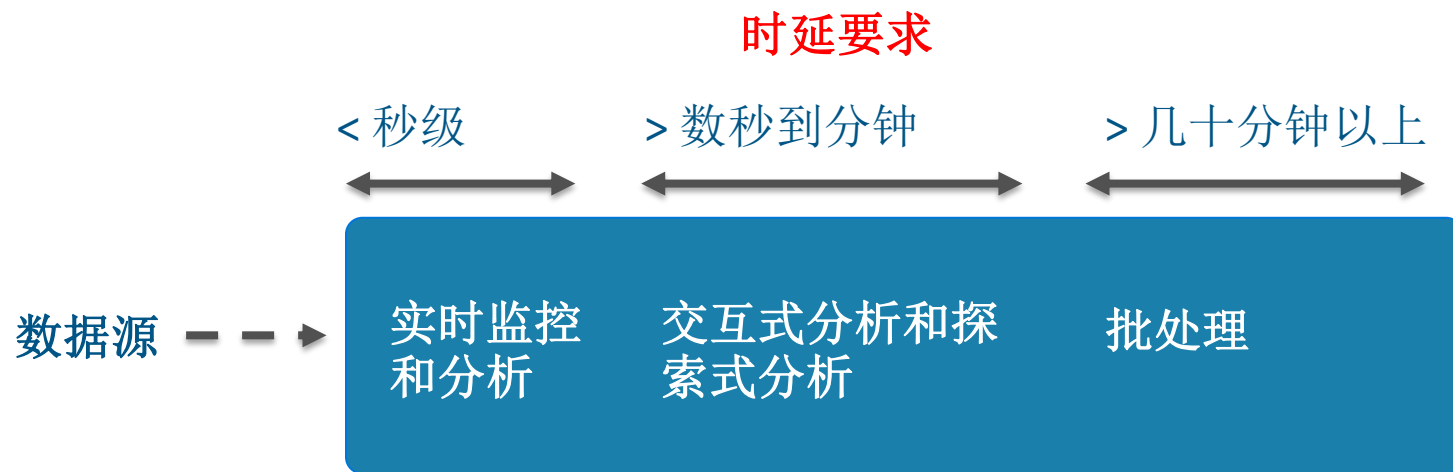
SQL的实现途径

- 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



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

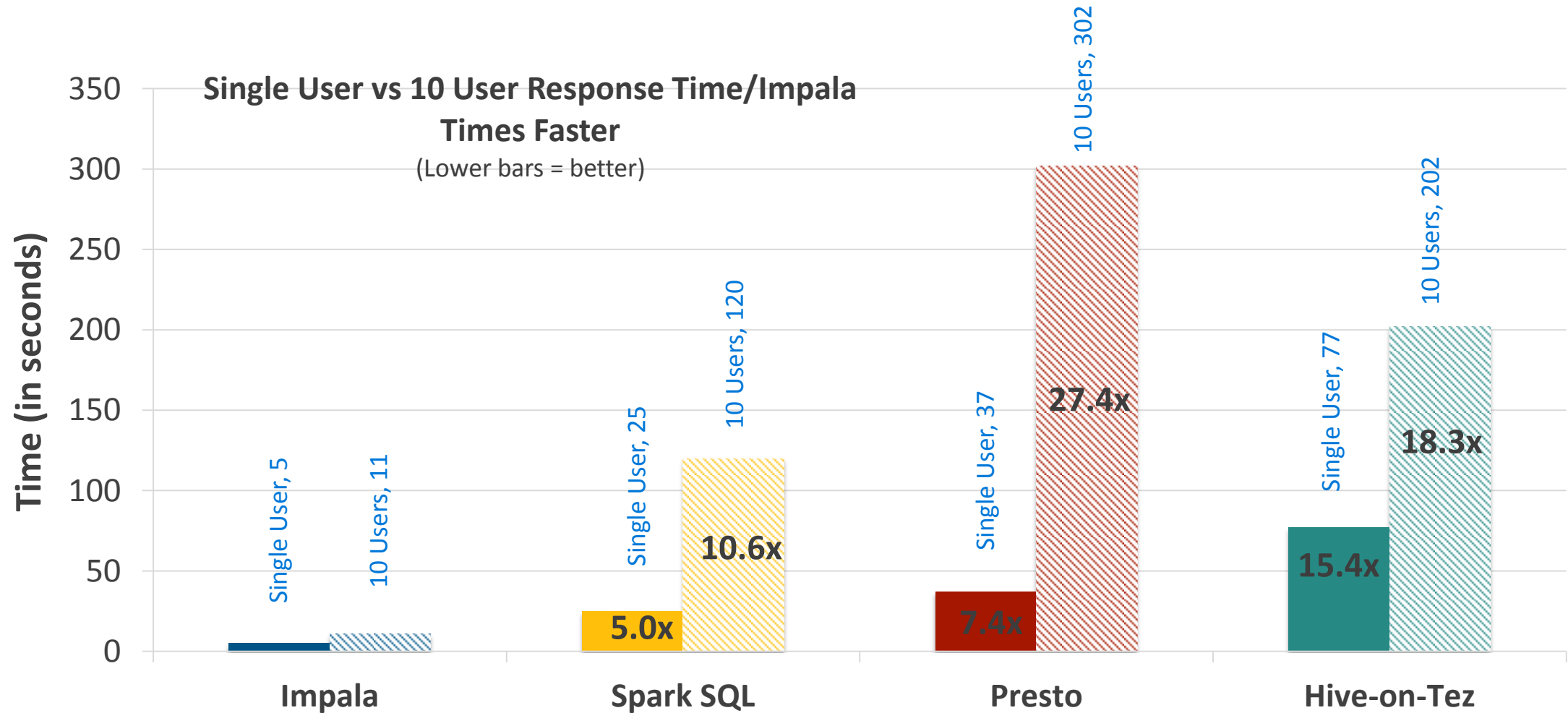


Batch
Processing



Spark developers

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 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

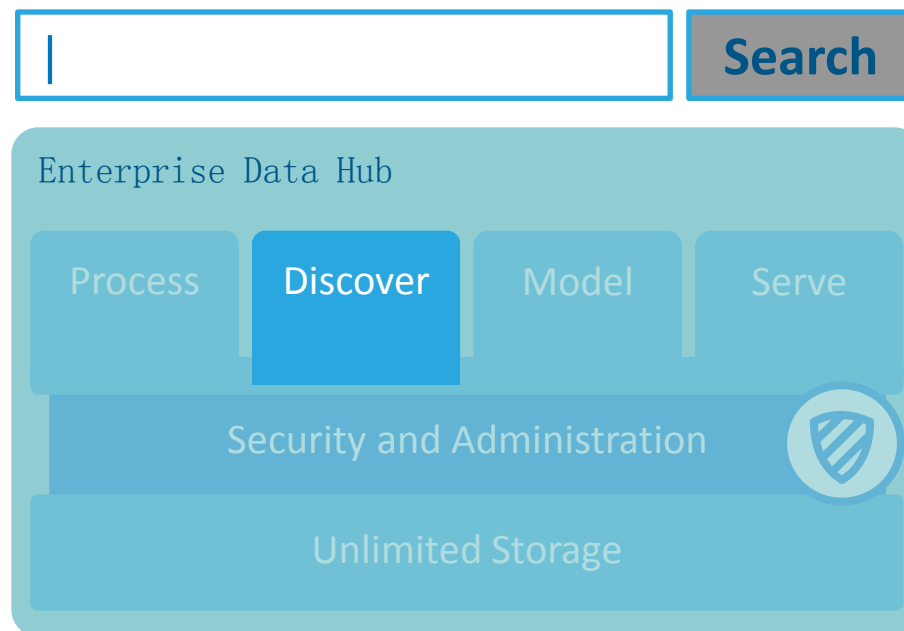
灵活性

实现了多种索引的构建方式

安全和集成

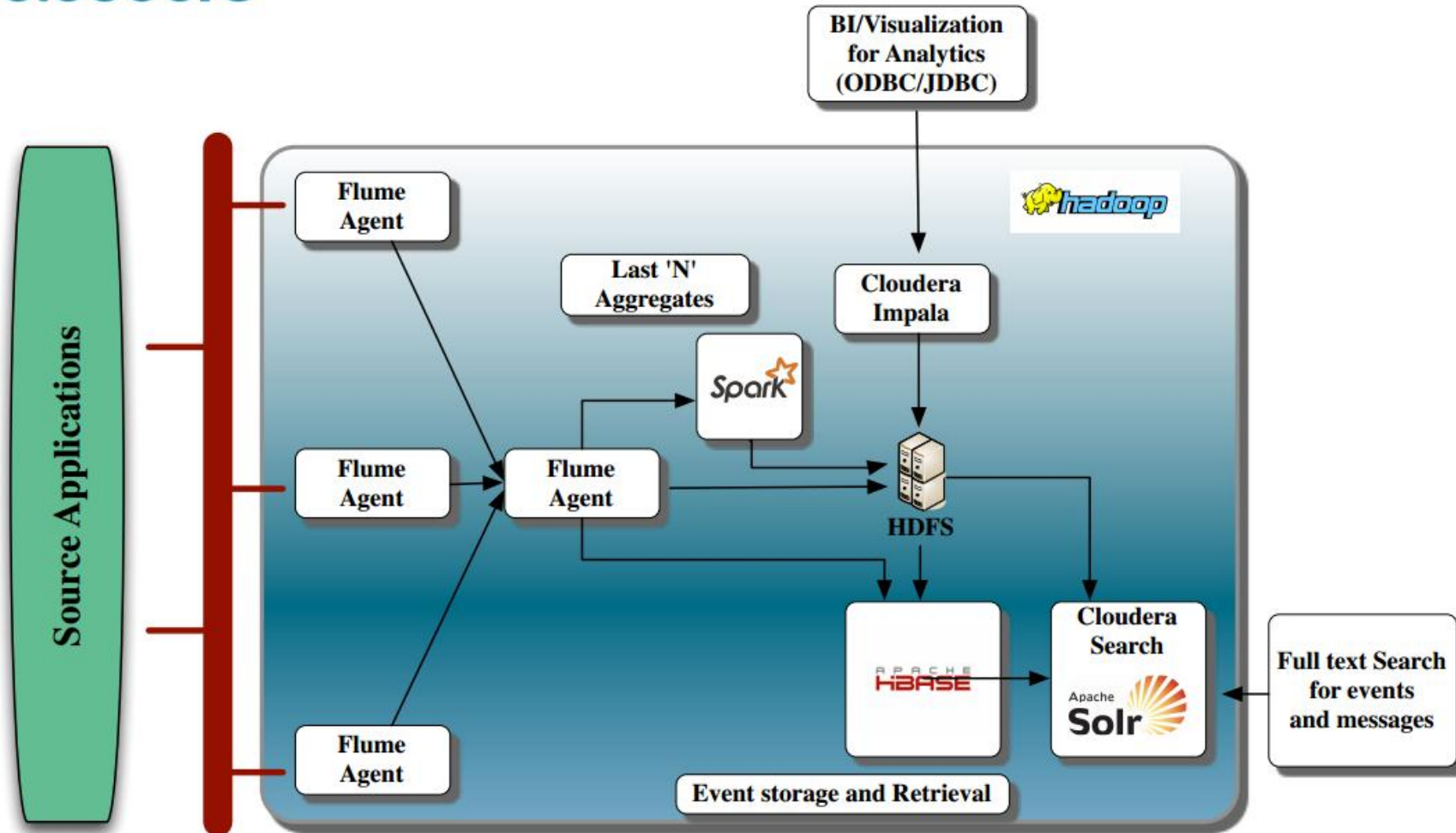
和Cloudera企业级功能的紧密集成

所有人都知道怎么搜索



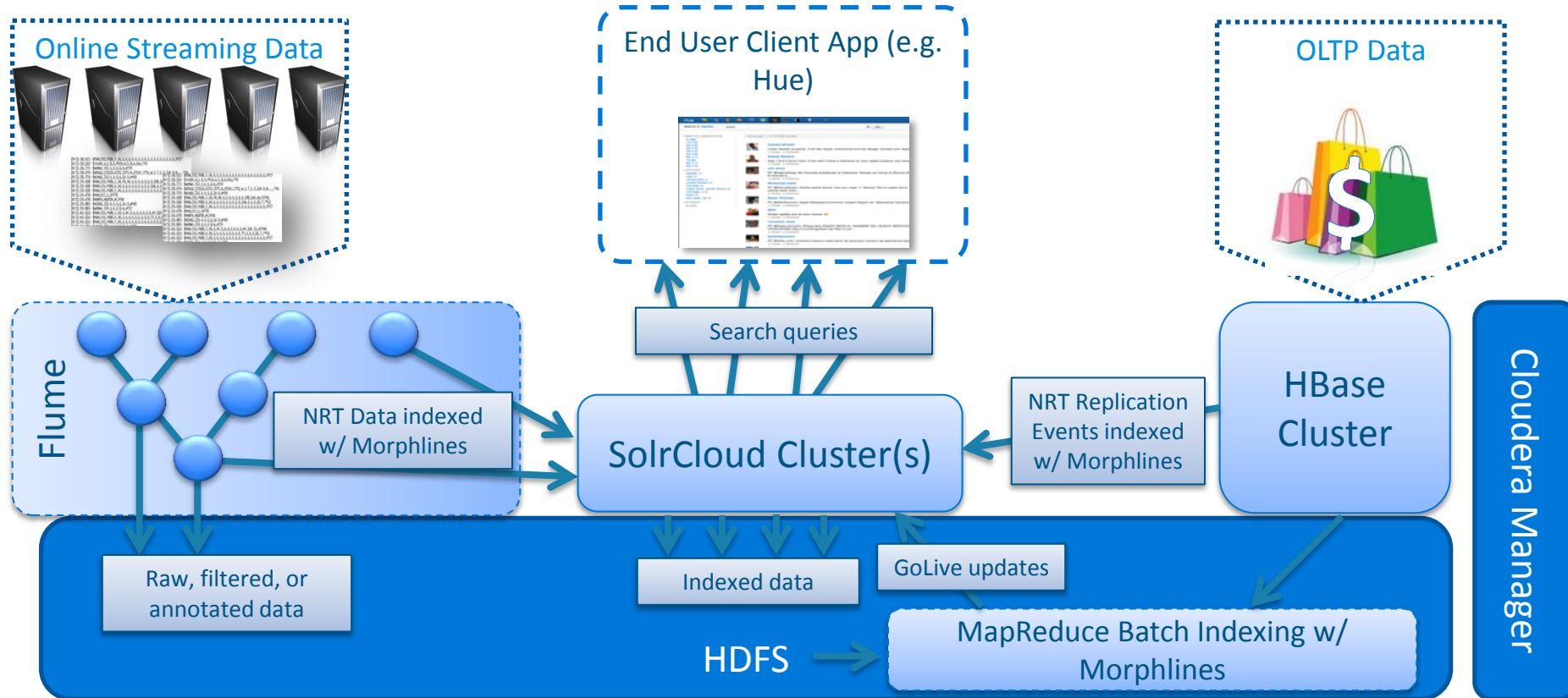
实时数据处理

cloudera®

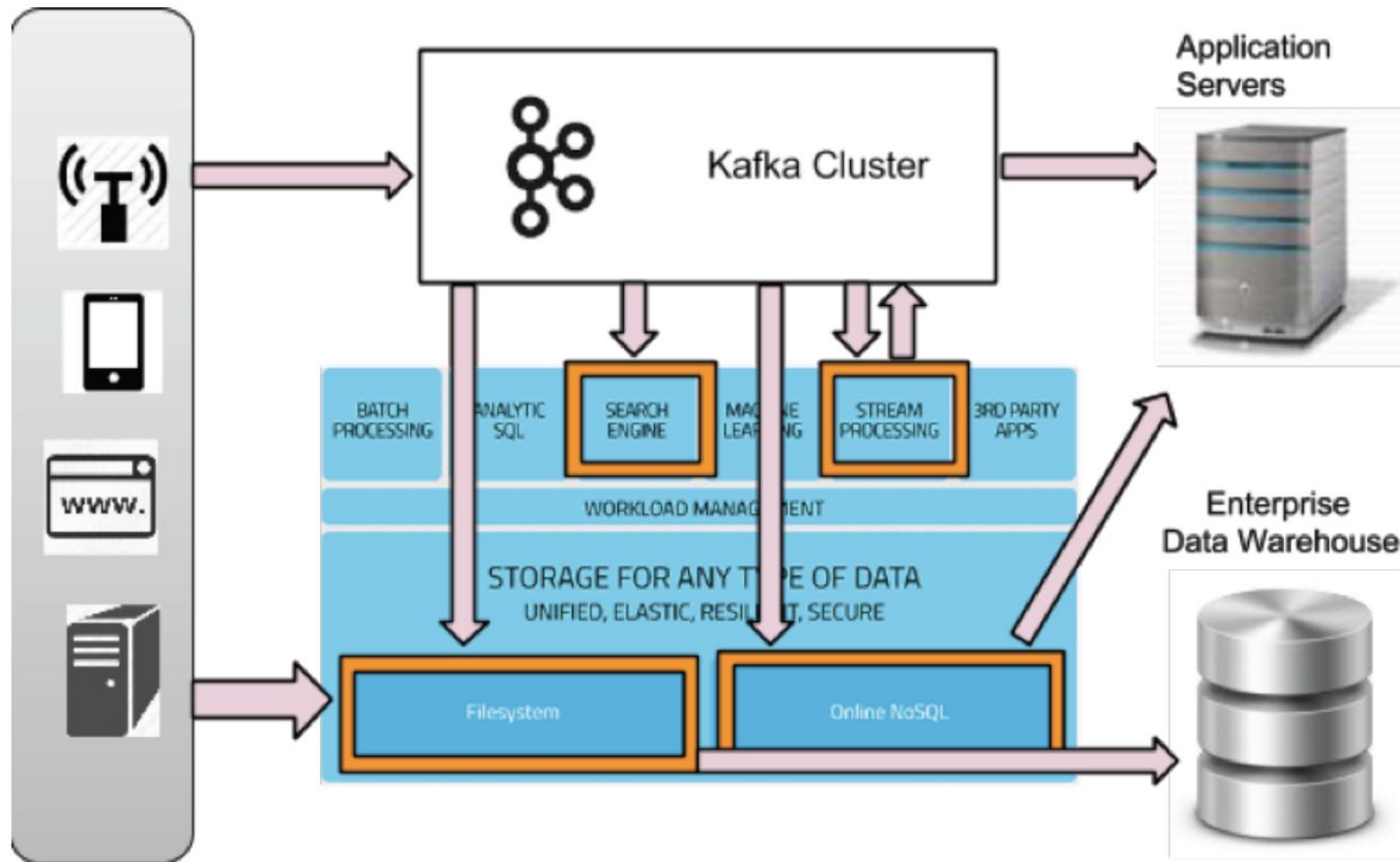


cloudera

实时搜索



企业消息总线

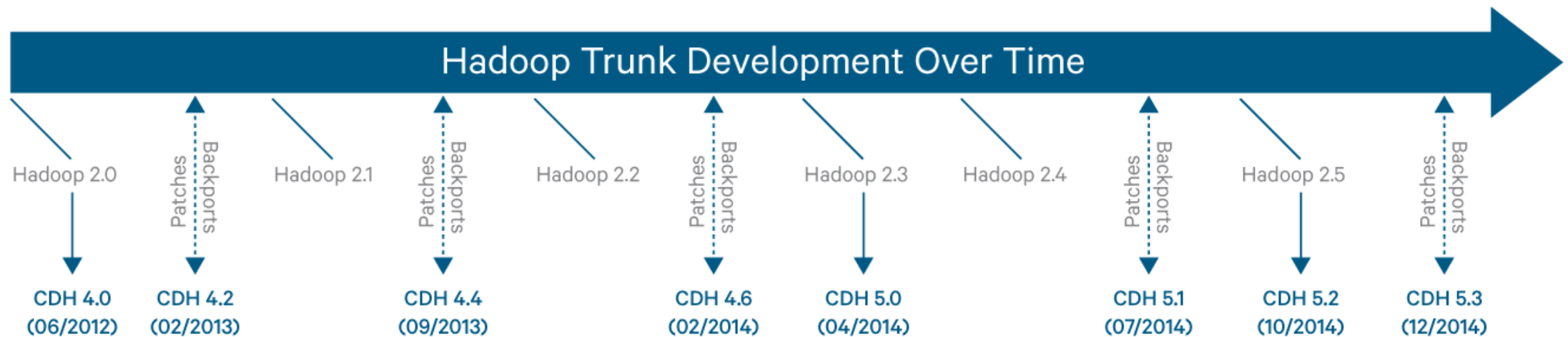


Hue

- 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



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
 -

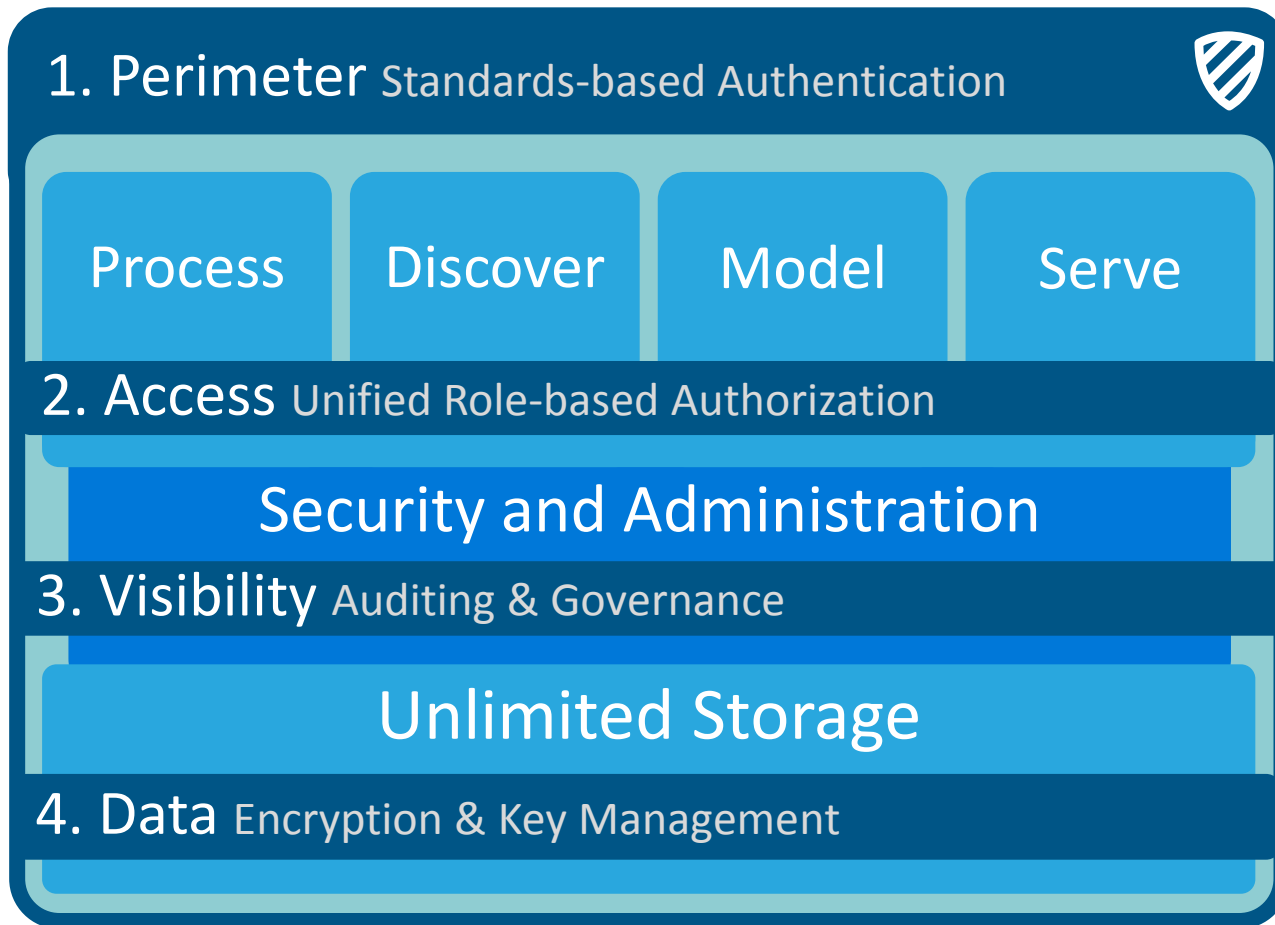
Security Infrastructure in Cloudera Enterprise – Apache Sentry & Cloudera Navigator

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

全面的安全管控

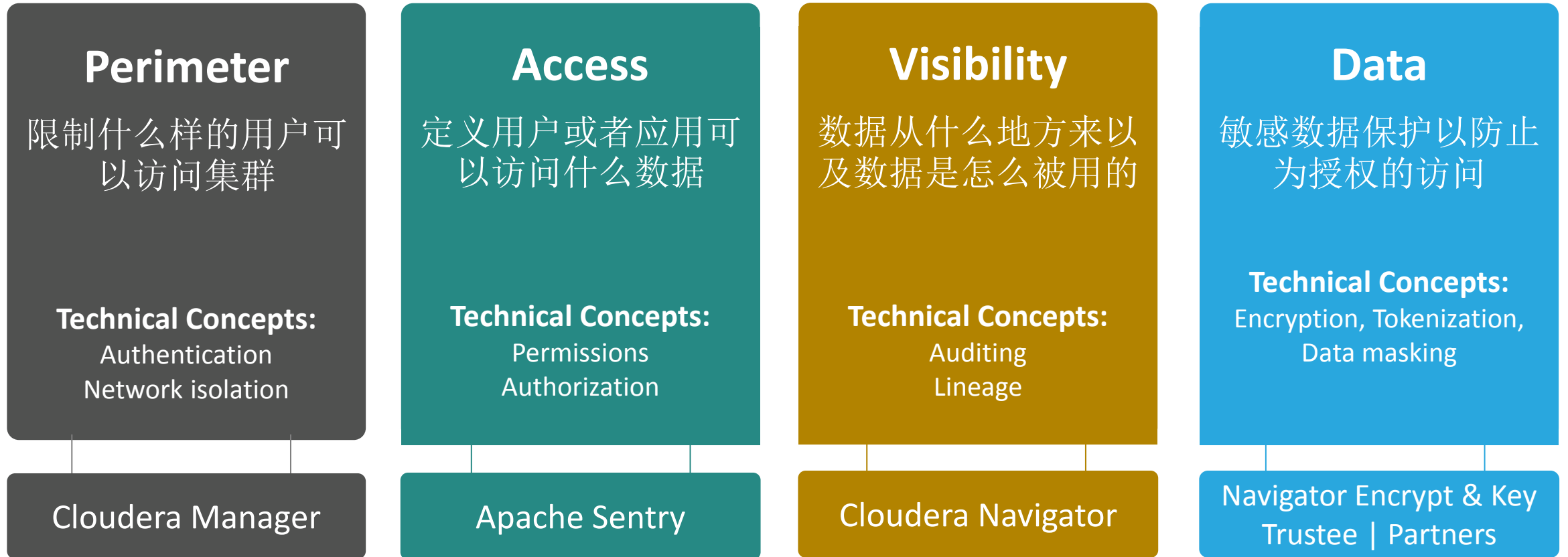
Apache Sentry, HDFS Encryption, Cloudera Navigator, Key Trustee



- 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

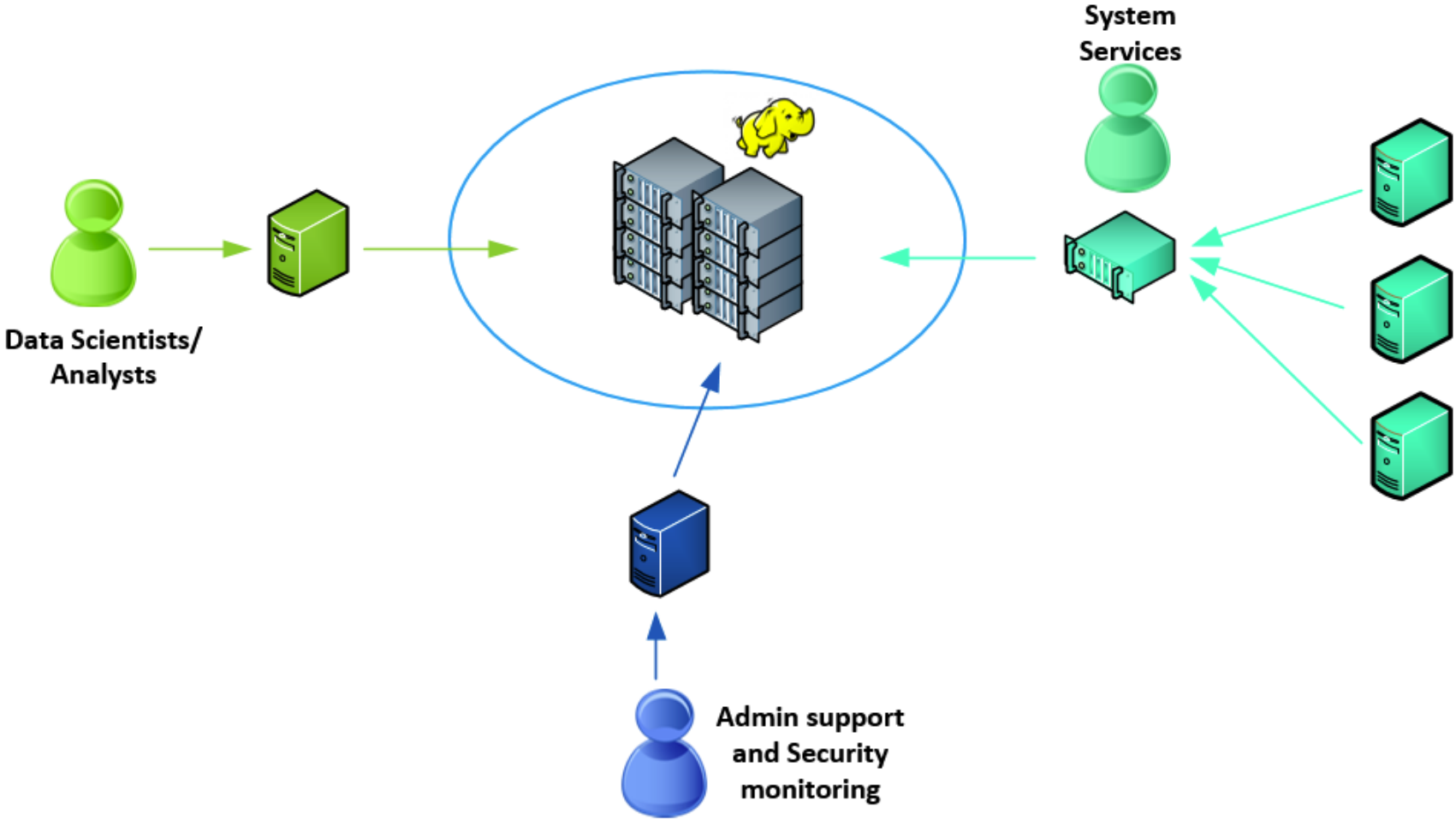
认证, 授权, 审计, 以及行业监管规范



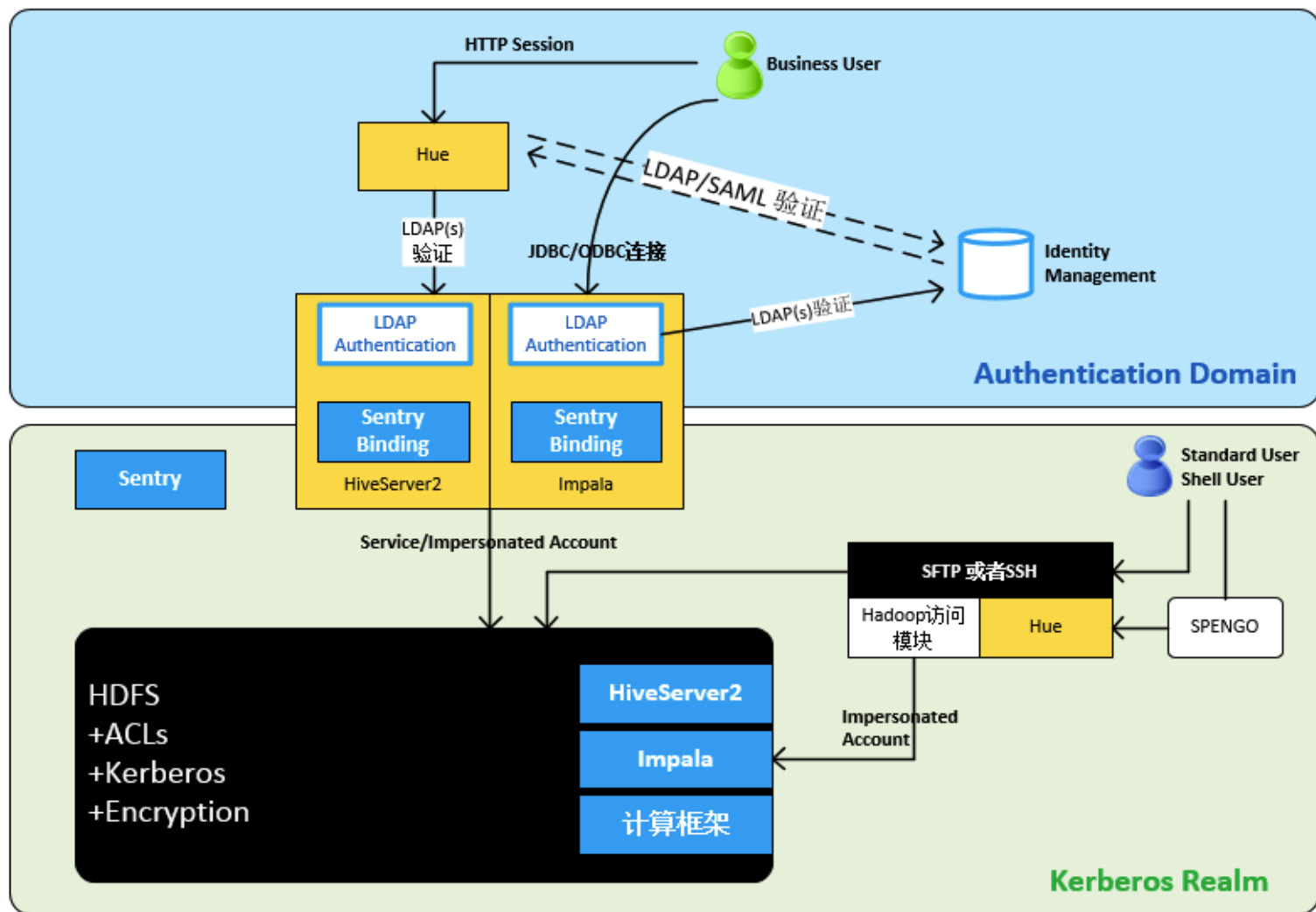
多层次的安全

- 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

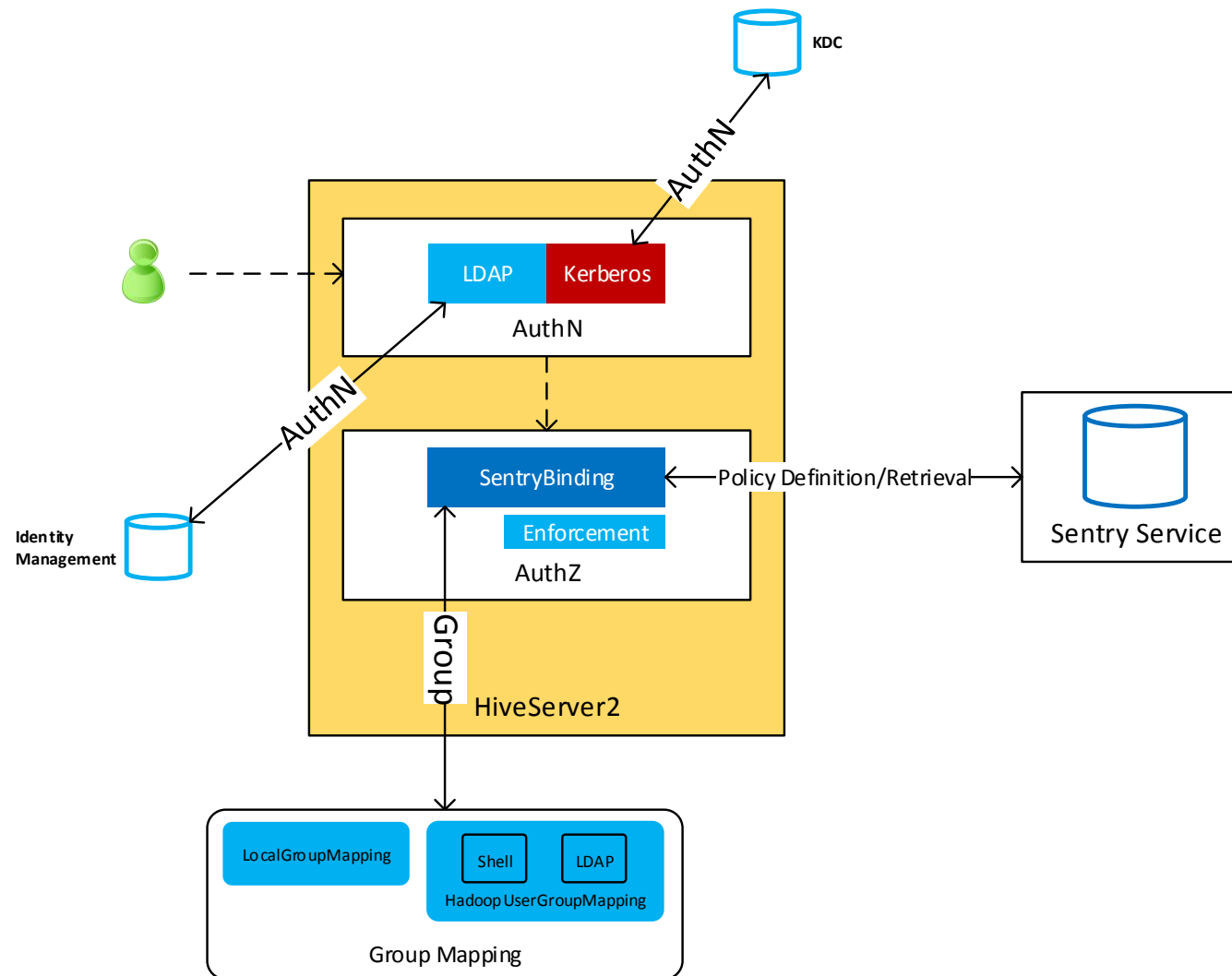
使用者视图



认证和授权



认证和授权

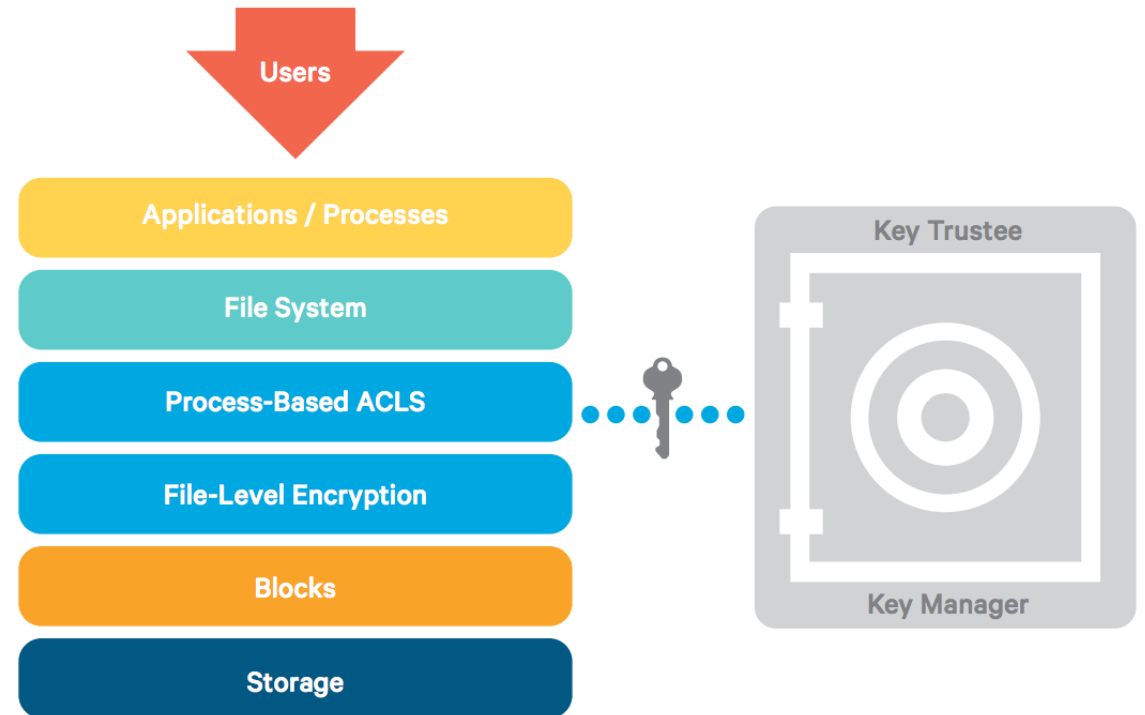


数据保护

- 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
 - 集中化的密钥管理
 - 灵活的部署方式
 - on-premise或者SaaS



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

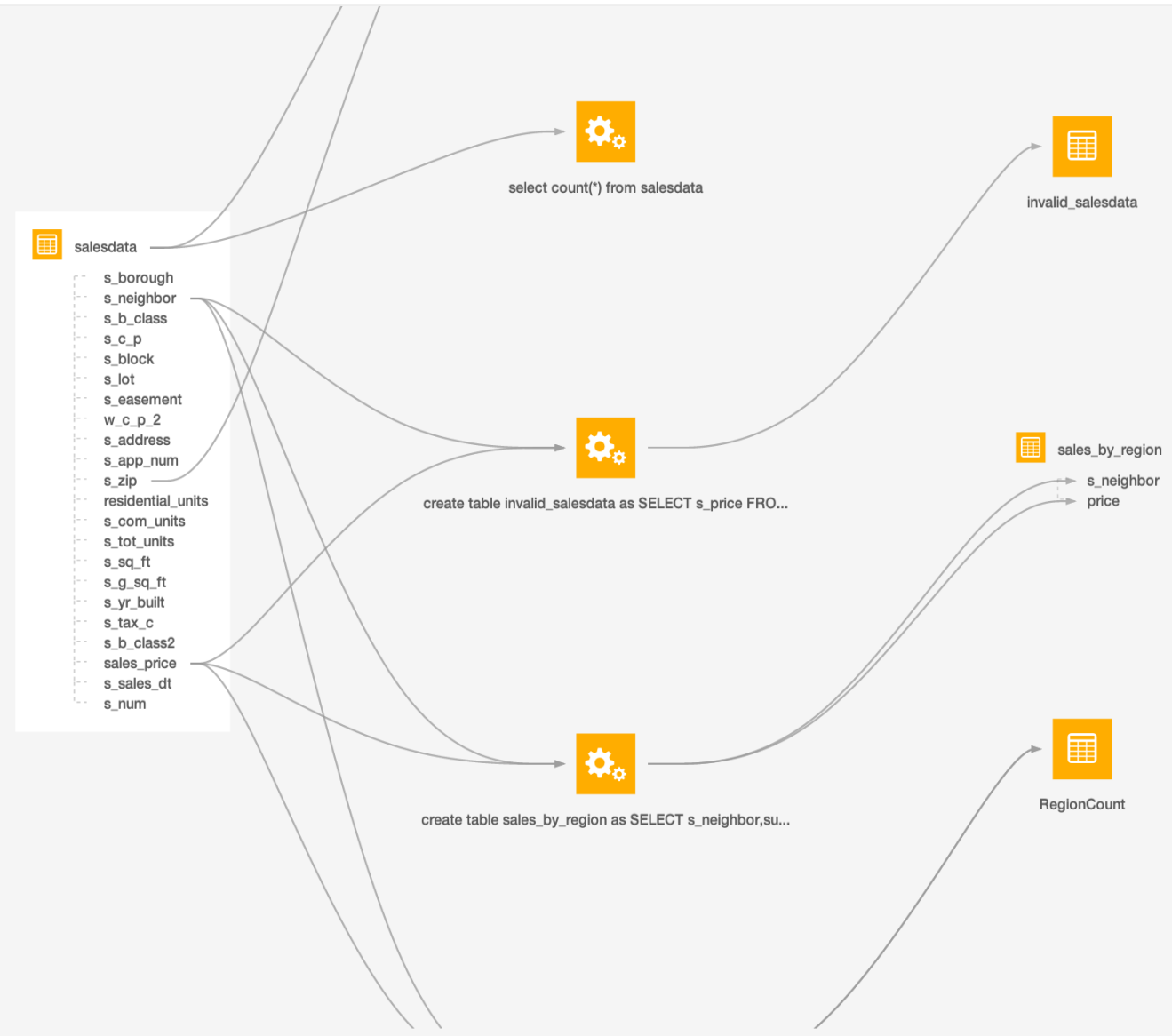
origin: mls

month: 03-2014

state: NY

retain-until: 5-15-2018

Lineage Schema

 Instances

sales_by_region

tags:

source type: HIVE

type: TABLE

owner: training

created: Apr 3 2014 2:55 AM

source: HIVE-1

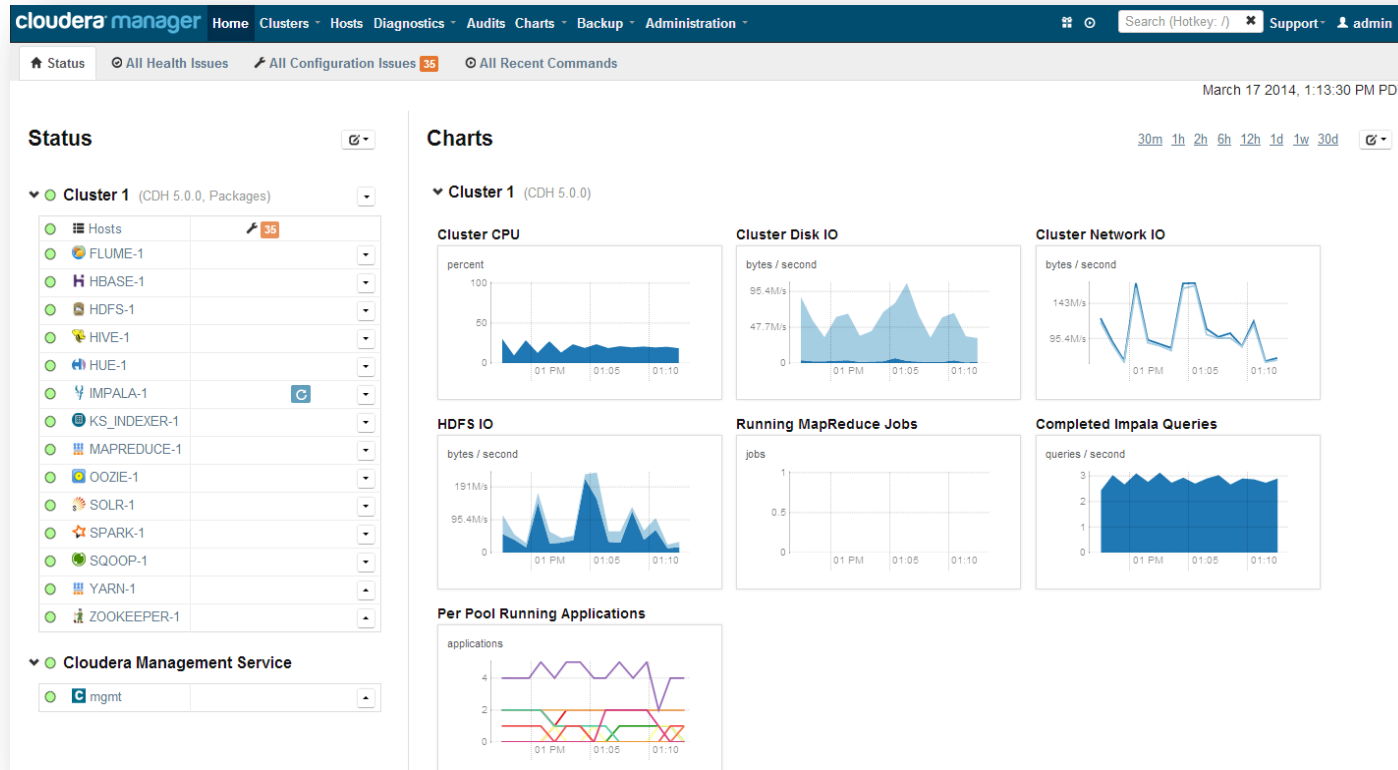
sensitivity: low (except for Brian)

expires: 2014

Most powerful Hadoop platform Management – Cloudera Manager

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
-

滚动重启和升级

The screenshot displays the Cloudera Manager interface for two clusters. Cluster 3 is the primary focus, showing several CDH components. The top row includes three CDH components (5.3.3-1.cdh5.3.3.p0.5, 5.3.2-1.cdh5.3.2.p0.10, 5.3.0-1.cdh5.3.0.p0.30) and one active CDH component (5.2.4-1.cdh5.2.4.p0.3). The bottom row includes CDH (5.2.1-1.cdh5.2.1.p0.12), KAFKA (0.8.2.0-1.kafka1.2.0.p0.2), and CDAP (2.7.1-1). Cluster 2 is partially visible at the bottom, showing four CDH components with various states (active, downloaded).

The screenshot shows the 'Rolling Restart' dialog box in Cloudera Manager. The 'Rolling Restart' menu item is highlighted in the left-hand navigation pane. The dialog box contains the following settings:

- Restart Roles with Stale Configs Only
- Restart Roles with Old CDH Versions Only
- Restart Role Types**
 - SecondaryNameNode
 - NameNode
 - DataNode
- Number of Slave Roles to Restart Together**: 1
- Stop Rolling Restart when this number of Slave Batches Fail (For Advanced Users Only)**: 0

Buttons: Confirm, Cancel

集群灾备

复制 计划 源

所有复制

+ 创建

搜索

hdfs3 (cluster3)



目标: hdfs2 (Cluster 2)
路径: /user/root/simpleData2 > /user/root
消息: 已复制 0 个文件, 5 个未更改。
上次运行: 2015年4月14日晚上6点30
上一次成功: 2015年4月14日晚上6点30
下一次运行: 2015年4月14日晚上6点35

操作

- 编辑配置
- 试运行
- 立即运行
- 禁用
- 删除



消息: HDFS replication succeeded.
开始时间: 2015年4月14日晚上6点30
结束时间: 2015年4月14日晚上6点30
命令: [详细信息](#)
可复制文件: 5 ([详细信息](#))
未复制的文件数: 5 ([详细信息](#))
可复制字节: 309.9 MiB
已删除文件数: 0 已复制文件数: 0
已复制字节: 0 B 已跳过文件数: 5
失败的文件数: 0
MapReduce 作业: [详细信息](#)

运行复制

没有运行的复制。

即将进行的复制



源: hdfs3 (cluster3)
目标: hdfs2 (Cluster 2)
路径: /user/root/simpleData2 > /user/root
开始时间: 2015年4月14日晚上6点35

hdfs3 (cluster3)



目标: hdfs2 (Cluster 2)
路径: /user/root/simpleData2 > /user/root
消息: 已复制 0 个文件, 5 个未更改。
上次运行: 2015年4月14日晚上7点05
上一次成功: 2015年4月14日晚上7点05
下一次运行: 2015年4月14日晚上7点10

操作

- 2015年4月14日晚上6点30 成功 可复制文件: 5 (309.9 MiB)
已复制文件数: 0 (0 B) 失败的文件数: 0 已删除文件数: 0 已跳过文件数: 5
- 2015年4月14日晚上6点25 成功 可复制文件: 5 (309.9 MiB)
已复制文件数: 0 (0 B) 失败的文件数: 0 已删除文件数: 0 已跳过文件数: 5
- 2015年4月14日晚上6点20 失败
- 2015年4月14日晚上6点15 失败
- 2015年4月14日晚上6点10 失败
- 2015年4月14日晚上6点05 失败
- 2015年4月14日晚上6点04 失败

配置历史

cloudera manager 主页 群集 主机 诊断 审核 图表 备份 管理

Cluster 1 30分钟 在 2015年4月1日, 上午10点28 CST 之前

HDFS 状态 实例 配置 命令 审核 文件浏览器 图表

配置和角色组历史记录

消息	日期	用户使用
当前版本		
Configured as part of enabling Kerberos. These changes should be reverted only if you want to disable Kerberos. (详细信息...)	2015-3-30 22:38:53 CST	
过去的版本: 全部显示 在选定的时间范围内显示		
已更新服务和角色类型配置。(详细信息...)	2015-3-28 10:03:43 CST	admin
已更新服务和角色类型配置。(详细信息...)	2015-3-28 9:59:34 CST	admin
已更新服务和角色类型配置。(详细信息...)	2015-3-28 9:58:42 CST	admin
已更新服务和角色类型配置。(详细信息...)	2015-3-28 9:42:13 CST	admin
Express wizard autoconfigured (详细信息...)	2015-3-27 12:12:29 CST	admin

版本回滚

版本详细信息

消息: Configured as part of enabling Kerberos. These changes should be reverted only if you want to disable Kerberos.

日期: 2015-3-30 22:38:53 CST

用户:

创建的组: 无

删除的组: 无

配置值	组成员资格	
属性	值	说明
DataNode Group 1 设置		
DataNode HTTP Web UI 端口	1006 50075 默认值	DataNode HTTP Web UI 的端口。结合 DataNode 的主机名称建立其 HTTP 地址。
DataNode 收发器端口	1004 50010 默认值	DataNode 的 Xceiver 协议的端口。结合 DataNode 的主机名称建立其地址。
DataNode 数据目录权限	700 755	DataNode 存储其块的本地文件系统中的目录权限。权限必须为八进制。755 和 700 是常用值。
DataNode Group 3 设置		
DataNode 收发器端口	1004 50010 默认值	DataNode 的 Xceiver 协议的端口。结合 DataNode 的主机名称建立其地址。

还原配置更改

关闭

智能的配置警告

cloudera manager 主页 群集 主机 诊断 审核 图表 备份 管理

主页 状态 所有运行状况问题 所有配置问题 ✖ 5 所有最新命令 🕒 1

所有配置问题

其他

- **C** Cloudera Management Service: [Service Monitor 的 Java 堆栈大小 \(字节\)](#)
推荐的堆大小为 256.0 MiB 字节, 大于配置 27.0 MiB。
- **C** Cloudera Management Service: [Service Monitor 的最大非 Java 内存](#)
建议的非 Java 内存大小为 1.5 GiB, 大于配置 98.0 MiB。
- **C** Cloudera Management Service: [Host Monitor 的 Java 堆栈大小 \(字节\)](#)
推荐的堆大小为 256.0 MiB 字节, 大于配置 27.0 MiB。
- **C** Cloudera Management Service: [Host Monitor 的最大非 Java 内存](#)
建议的非 Java 内存大小为 1.5 GiB, 大于配置 98.0 MiB。

Cluster 1

- **H** HDFS: [Namenode 的 Java 堆栈大小 \(字节\)](#)
Java Heap Size of Namenode in Bytes is recommended to be at least 1GB for every million HDFS blocks. Suggested minimum value: 1073741824

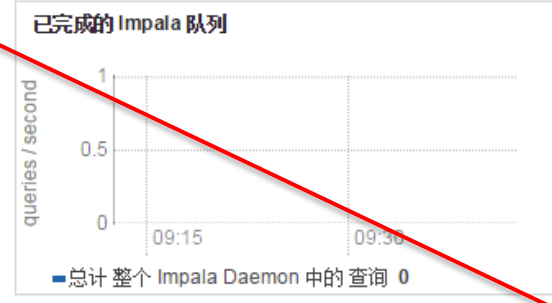
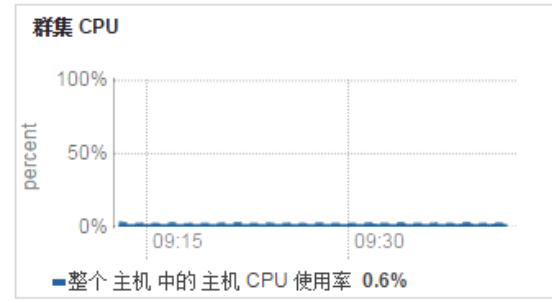
智能决策

主页 状态 所有运行状况问题 所有配置问题 ✖ 6 所有最新命令

Cluster 1 (CDH 5.3.2, Parcel)

主机		
HBase		C
HDFS	✖ 1	C
Hive		C
Hue	✖ 1	C
Impala		C
Key-Value Stor...		C
Oozie		C
Solr		C
Spark		C
Sqoop 2		C
YARN (MR2 Inc...		C
ZooKeeper		C

配置过期需要重启



客户端配置过期

Cluster 1 过期配置

查看更改后，调用 重启群集 向导将更改传播到所有角色，重新部署客户端配置并重启群集。

审核更改

Filter by 所有文件 Hue 所有角色 ✖ 删除筛选器

```
文件: hbase-conf/hdfs-site.xml
... .. @@ -27,9 +27,9 @@
27 27 <value>3</value>
28 28 </property>
29 29 <property>
30 30 <name>dfs.blocksize</name>
31 - <value>134217728</value>
31 + <value>67108864</value>
32 32 </property>
33 33 <property>
34 34 <name>dfs.client.use.datanode.hostname</name>
35 35 <value>>false</value>

文件: hive-conf/hdfs-site.xml
... .. @@ -27,9 +27,9 @@
27 27 <value>3</value>
28 28 </property>
29 29 <property>
30 30 <name>dfs.blocksize</name>
31 - <value>134217728</value>
31 + <value>67108864</value>
32 32 </property>
33 33 <property>
34 34 <name>dfs.client.use.datanode.hostname</name>
35 35 <value>>false</value>

文件: hue.ini
... .. @@ -65,10 +65,8 @@
65 65 [sqoop]
66 66 server_url=http://sb-node3:12000/sqoop
67 67 [search]
68 68 solr_url=http://sb-node3:8983/solr
69 69 [-hbase]
70 70 -hbase_clusters=(HBase[sb-node4:9090])
71 69 [proxy]
72 70 whitelist=(localhost|127.0.0.1):(50030|50070|50060|50075)
73 71 [shell]
```

hue: 配置问题

Hue: 在 HBase Thrift Server 属性中选择服务器以使用 Hue HBase Browser 应用程序。

全局时间线控制方便诊断

The screenshot displays the Cloudera Manager interface for Cluster 1. At the top, a navigation bar includes 'cloudera manager', '主页', '集群', '主机', '诊断', '审核', '图表', '备份', and '管理'. Below this, a timeline shows a period from 01:35 to 02:55, with a red box highlighting the interval from 01:55 to 02:55. A red arrow points from this timeline to a red-bordered box containing a grid of performance charts.

HDFS 汇总

配置的容量 2.6 太字节/25.5 太字节

快速链接 [复制, 报告, NameNode Web UI \(活动\)](#)

事件搜索 [警报](#), [严重](#), [全部](#)

状态摘要

- SecondaryNameNode ● 1 运行状况良好
- NameNode ● 1 运行状况良好 (活动)
- DataNode ● 4 运行状况良好
- Balancer ? 未知

运行状况测试 全部展开

▶ ● 7 良好。

运行状况历史记录

时间	事件	操作
▶ ● 3月27日下午2点09	HDFS Canary 良好	显示
▶ ● 3月27日 2:08:01 下午	1 变成不良 5 变成良好	显示
▶ ● 3月27日 2:07:51 下午	1 变成存在隐患 4 变成未知 1 变成良好	显示
▶ ○ 3月27日下午1点57	7 变成已禁用	显示

图表

30分钟 1小时 2小时 6小时 12小时 1天 7d 30d

HDFS 容量

bytes

18.2T

02 PM 02:15 02:30 02:45

使用 840G 使用... 2.1T 总 25.5T

各 DataNode 中的总读取的字节

bytes / second

58.6K/s

39.1K/s

19.5K/s

02 PM 02:15 02:30 02:45

各 DataNode 中的总... 1b/s

各 DataNode 中的总写入的字节

bytes / second

381M/s

191M/s

02 PM 02:15 02:30 02:45

各 DataNode 中的... 445M/s

各 DataNode 中的总读取块

blocks / second

0.4

0.2

02 PM 02:15 02:30 02:45

各 DataNode 中... 0.02

各 DataNode 中的总已写入块

blocks / second

4

2

02 PM 02:15 02:30 02:45

各 DataNode 中的总... 4.1

各 DataNode 中的总收发器

transceivers

80

60

40

02 PM 02:15 02:30 02:45

各 DataNode 中的... 86

整个 DataNode 中的收发器

transceivers

25

20

10

02 PM 02:15 02:30 02:45

整个 DataNode ... 21.5

整个 DataNode 中的数据包确认往返的平均时间

nanos

15ms

10ms

5ms

02 PM 02:15 02:30 02:45

整个 DataNode 中的数据包确认往... 12ms

整个 DataNode 中的发送数据包传输的平均时间

nanos

200µs

100µs

02 PM 02:15 02:30 02:45

整个 DataNode 中的发送数据包... 15.05µs

整个 DataNode 中的发送网络阻止数据包的平...

整个 DataNode 中的磁盘刷新

2K

整个 DataNode 中的平均磁盘刷新时间

300µs

极方便的全局时间线控制

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 用于 Cluster 1

KDC 信息

指定有关 KDC 的信息。Cloudera Manager 使用下面的属性生成在群集中运行的 CDH 守护程序的主体。

KDC 类型

- MIT KDC
 Active Directory

用于在 CDH 群集中进行身份验证的 KDC 类型。

KDC Server 主机

kdc

sb-node1.example.com



KDC Server 所在的主机。

Kerberos 安全领域

default_realm

EXAMPLE.COM



Kerberos 安全所用领域。注：更改该设置将清除 Cloudera Manager 中现有的所有凭据和 Keytab。

Kerberos 加密类型

rc4-hmac



KDC 支持的加密类型。备注：如果想使用 AES 加密，确保按[此处](#)的说明部署 JCE 无限强度政策文件。

Kerberos Principal 最大可更新生命期

7

天



Cloudera Manager 生成的 Kerberos 主体最大可更新生存期。只有在使用 MIT KDC 时才使用该属性。如果 KDC 应该提供最大可更新生存期，则将该属性设为零。备注：建议不要采用票证不可更新的主体，因为会妨碍 Hadoop 服务运行。

启用Kerberos

KRB5 配置

指定生成群集的 `krb5.conf` 所需的属性。可以使用安全属性指定高级 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`。

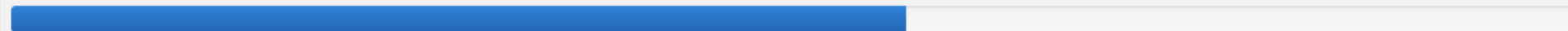
启用Kerberos

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

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?

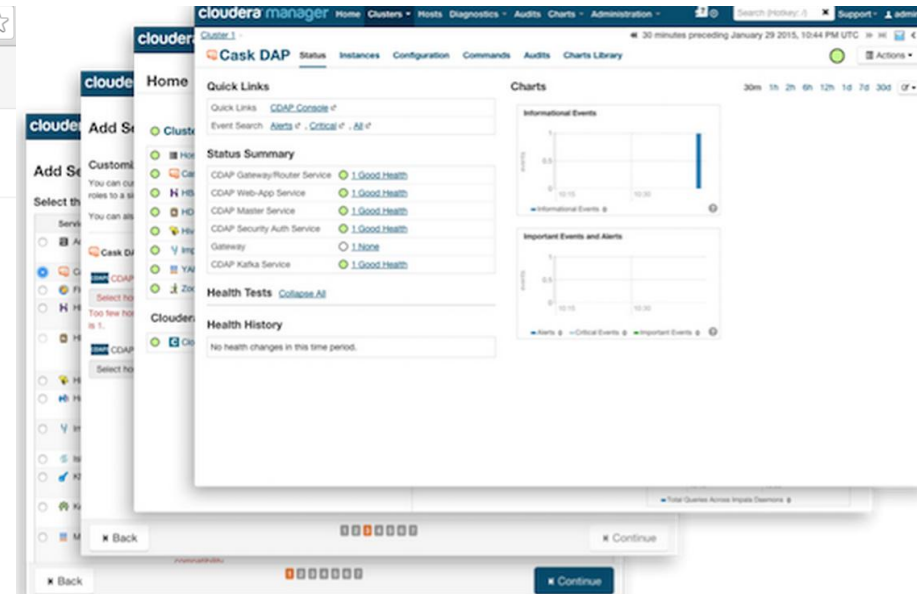
At the highest level, Cloudera Manager concerns itself with managing the lifecycle of various services running on a cluster. This can be broken down into two primary areas:

- 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
- Service Descriptor Language



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 Manager REST API. The API uses JSON Object Notation (JSON).

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

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 RESTful system, expand the URL path to include the user identifier. '/users/bar' identifies user 'bar'.



Cloudera Manager API

[View On GitHub](#)

- GETTING STARTED
 - Introduction
 - Quick Start
 - Cloudera Manager Concepts
 - Full API Docs

API access to Cloudera Manager

Cloudera Manager's REST API lets you work with existing tools, and programmatically manage your Hadoop clusters. The API is available in both Cloudera Express and Cloudera Enterprise, and comes with open-source client libraries.

[Download Cloudera Manager](#)

Inc. [US] https://github.com/cloudera/cm_api

This repository Search Explore Gist Blog Help

cloudera / cm_api Watch 48

Cloudera Manager API Client

268 commits 14 branches 3 releases 12 contributors

branch: master cm_api / +

[api] OPSAPS-23971 Add new CDH Upgrade arguments

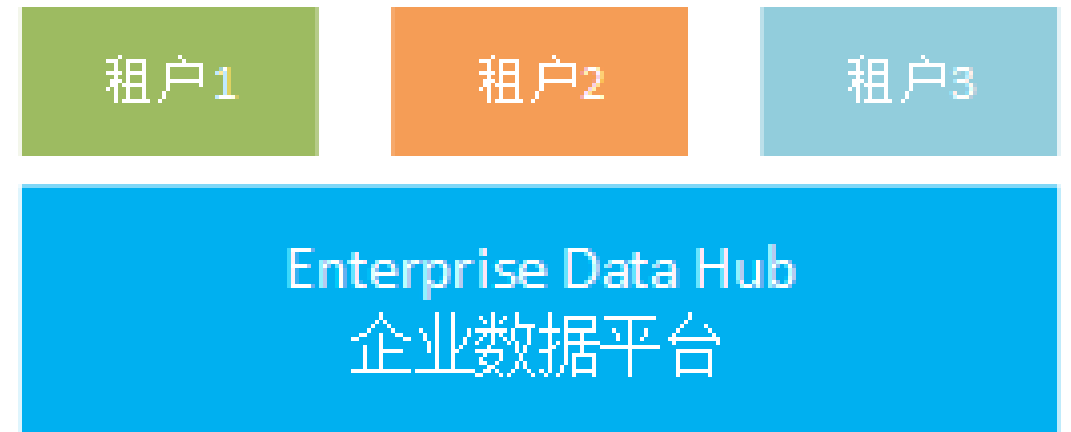
Darren Lo authored on 25 Nov 2014 latest commit dbad44c15b

phillip committed 10 days ago

- java [java] Updated with v8 java source 5 months ago
- nagios Fix Markdown formatting in nagios/README.md 3 years ago
- python [api] OPSAPS-23971 Add new CDH Upgrade arguments 10 days ago
- .gitignore typos and updated gitignore for PyDev 11 months ago
- LICENSE.txt Add ASL2 LICENSE.txt 3 years ago
- README.md Fix links to Java, Python, and the license. a year ago

多租户管理

- 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



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

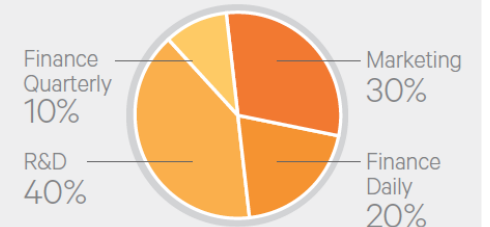
- 配额管理

- 磁盘空间配额

- 文件、目录数量配额，以优化文件系统元数据

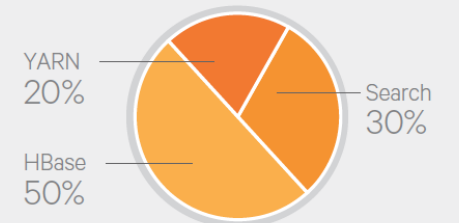
Dynamic Partitioning

- Shared concept of “consumers” between services
- Execution across services by cumulative usage per consumer



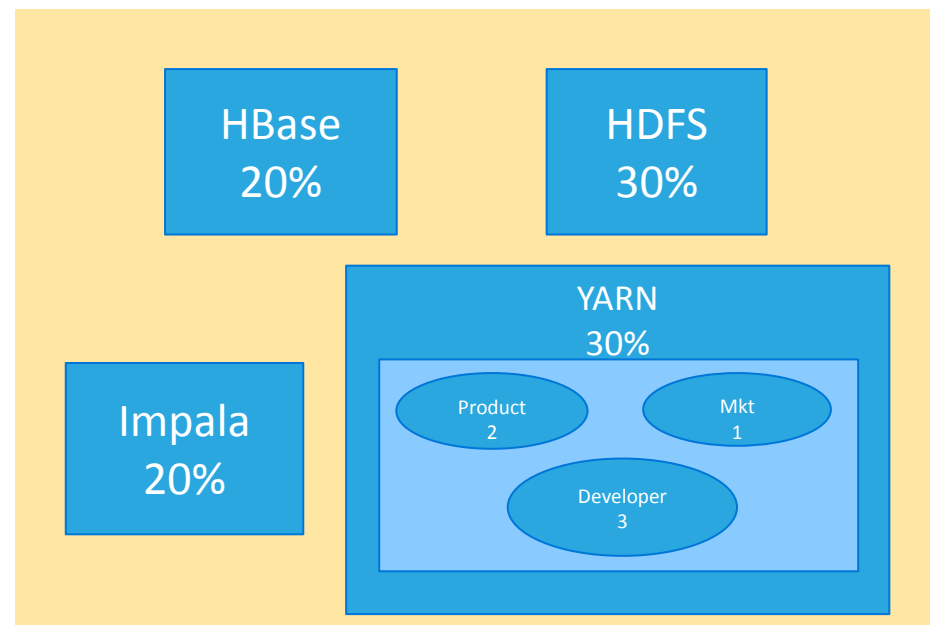
Static Partitioning

- Focused on separate and distinct services within a cluster
- Execution within a service per designated resource allocations



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



静态资源管理配置

cloudera manager 主页 群集 主机 诊断 审核 图表 备份 管理

Cluster 1 »

静态服务池 状态 配置

第 2 步, 共 4 步: 直接更改设置

高级

服务	分配 %
HBase	20 %
HDFS	20 %
Impala	30 %
YARN (MR2 Included)	30 %
总计	100 %

详细信息

Impala 用于资源管理的 YARN 服务

YARN (MR2 Included) 将 CGroups 用于资源管理 始终使用 Linux Container Executor

Cgroup CPU 共享

	主机数	值	小计	%
HDFS: DataNode	1 主机	400	400	100.0%
总计			400	100.0%

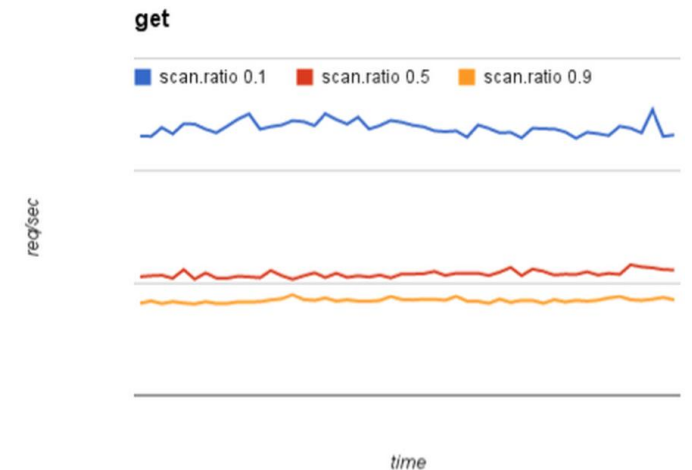
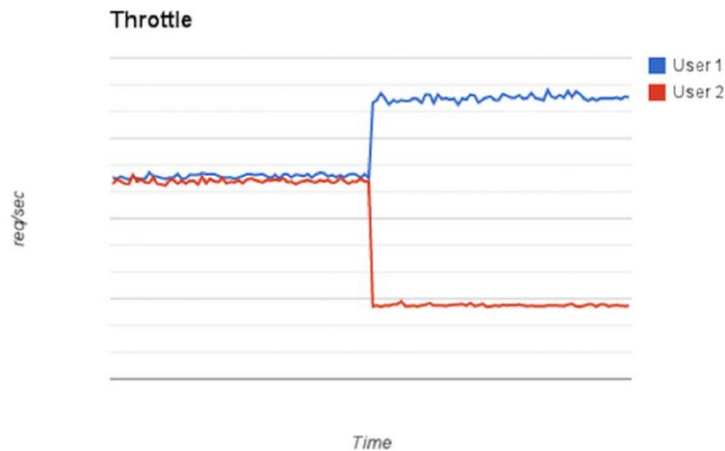
Cgroup CPU 共享

	主机数	值	小计	%
YARN (MR2 Included): NodeManager	1 主机	600	600	100.0%

返回 继续

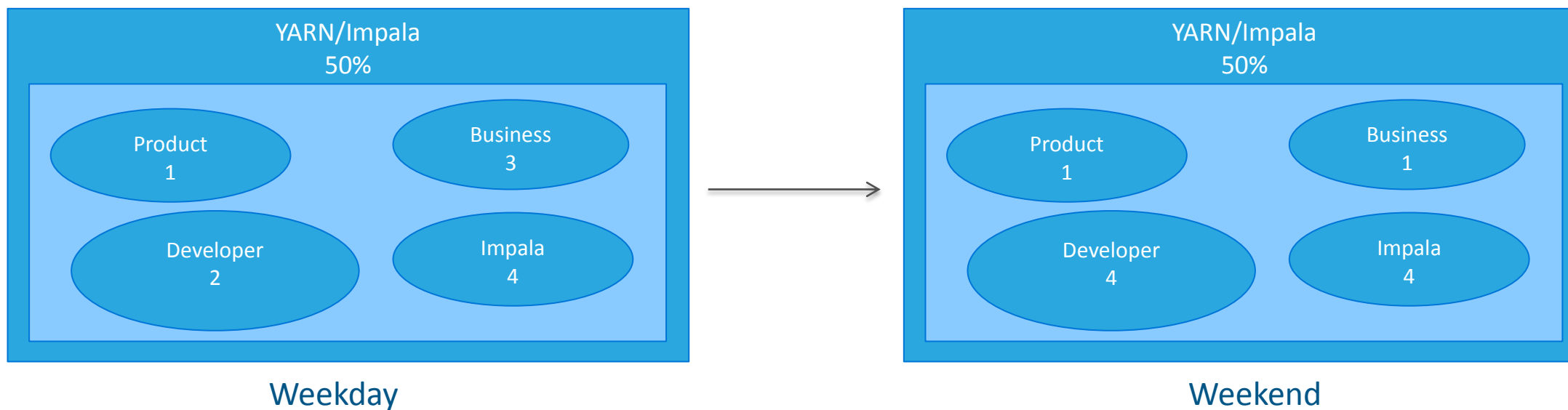
HBase内部的资源管理

- 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



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



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

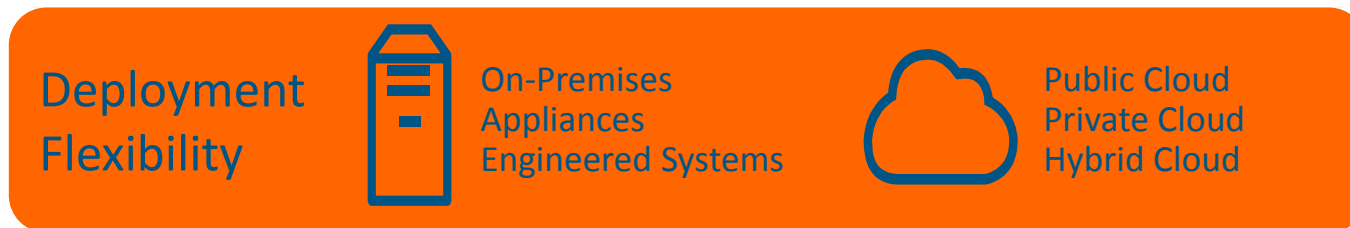
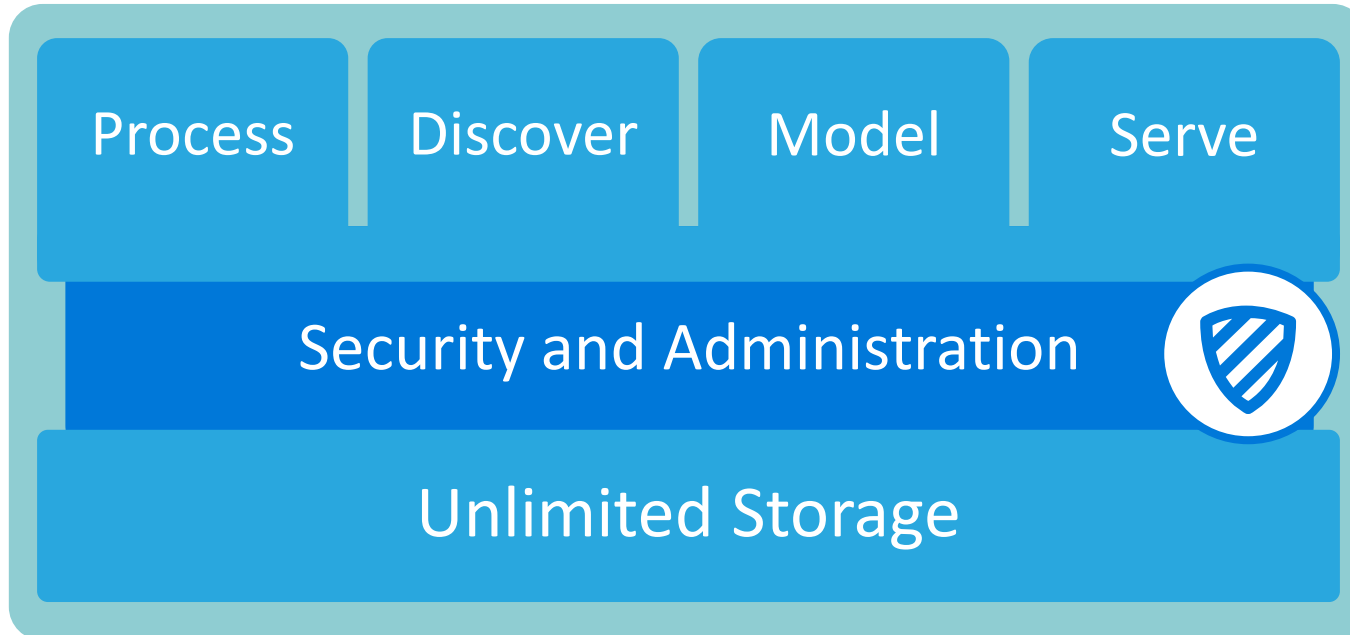
Reports

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 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 Director

**Portability: Multiple
Deployment Options**

Private Cloud

Physical

vmware



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

Ensure Customer Success – Industry-Leading Support

Cloudera技术支持

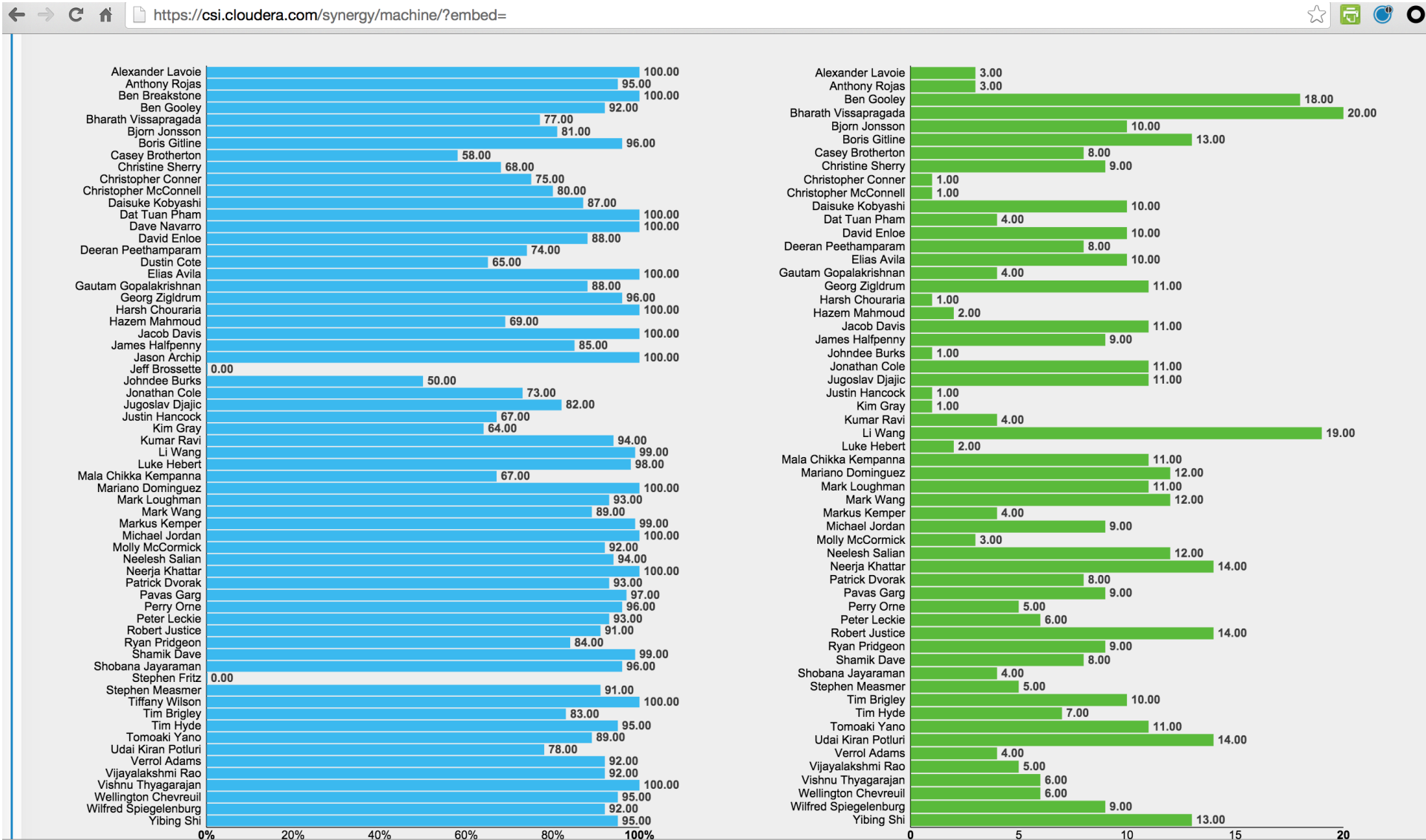
- 专业服务
- 近百人的专业技术支持团队
- 丰富的知识库
- 基于大数据技术的预测支持及主动支持
- 严格的问题修复流程

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

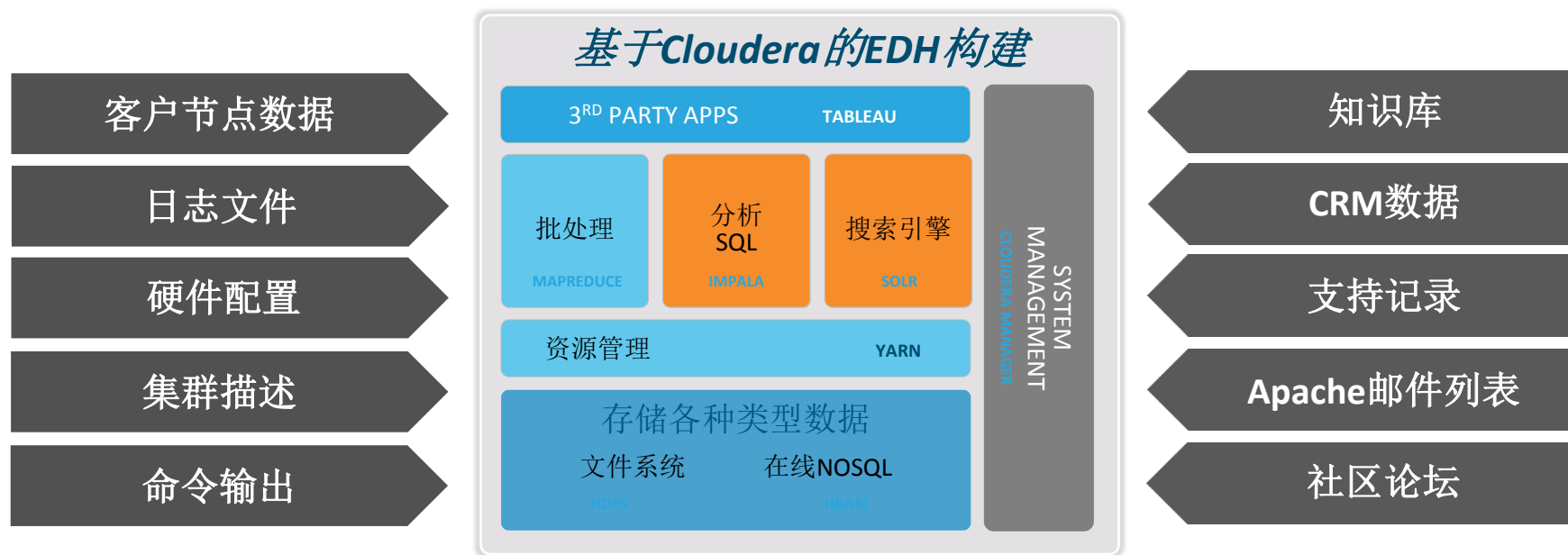


Cloudera 客户支持中心 (CSI)

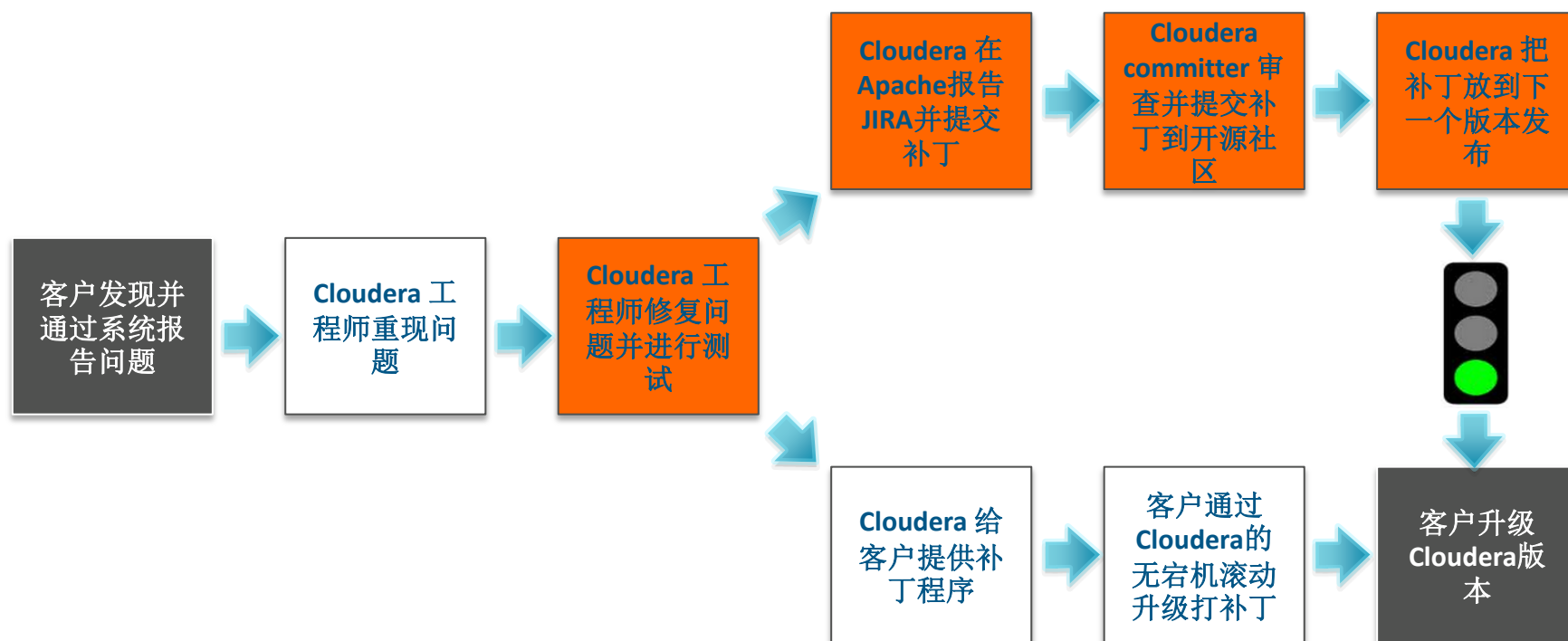


预测、主动技术支持

- 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



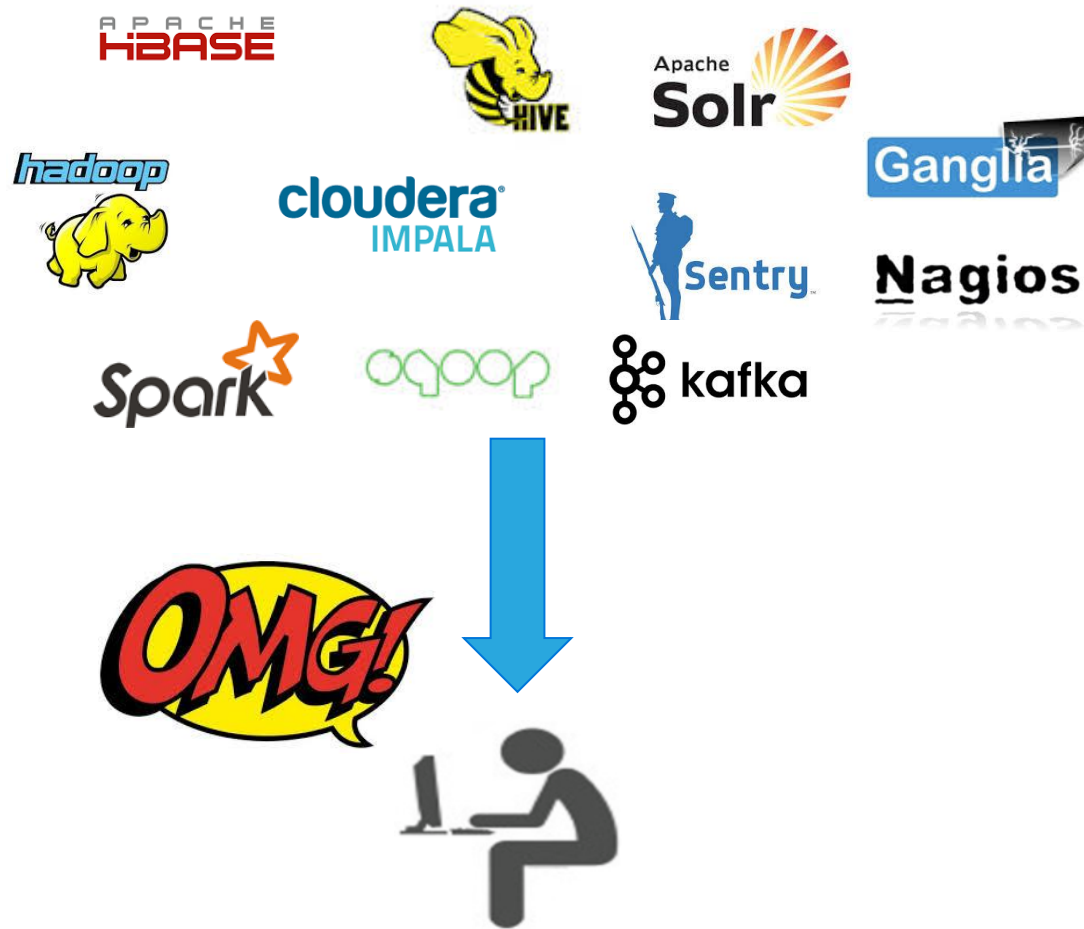
严格的问题修复过程



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

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	<ol style="list-style-type: none"> 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	<ol style="list-style-type: none"> 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	<ol style="list-style-type: none"> 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	<ol style="list-style-type: none"> 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

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

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）

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

Thank you

@Contact Info

Backup

灵活的版本选择 (1)

	CLUDERA EXPRESS	CLUDERA 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		✓	✓	✓

灵活版本选择 (2)

	CLUSTERA EXPRESS	CLUSTERA 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		✓	✓	✓

灵活版本选择 (3)

	CLUDERA EXPRESS	CLUDERA 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	✓	✓	✓	✓

灵活版本选择 (4)

	CLUDERA EXPRESS	CLUDERA 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

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**.

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

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

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

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
Thank you