cloudera®

白皮书

# 保护您的企业版 Hadoop 生态系统

## Cloudera 提供对企业的全面安全性

cloudera®

# 目录

cloudera®


## 摘要

企业通过统一的系统正在转变数据管理的方式，该系统综合了无限规模的分布式存储和计算技术，适用于任何数据量和任何类型，并具备强大、灵活的分析功能，包括批处理、交互式 SQL 和全文检索。然而，为了充分发挥其潜力，这些企业数据中心需要进行认证、授权、审计和数据保护控制。

Cloudera 公司通过提供部署、管理和整合当今监管环境和合规性政策所要求的必要的数据安全控制，使企业数据中心实现信息驱动业务。Cloudera 公司的企业数据中心提供了跨四大安全性支柱的综合安全和治理功能，包括针对集中式、自动化认证管理的 Cloudera Manager；针对统一的基于角色授权的 Apache Sentry；针对综合审计、数据沿袭和元数据管理的 Cloudera Navigator；以及针对透明加密和企业级密钥管理的 Navigator Encrypt 与 Navigator Key Trustee。


## 引言

信息驱动型企业认识到下一代数据管理方式应该是一种统一的集中式系统，该系统能够存储任何形式或数量的所有数据，并为企业用户提供广泛多样的工具和应用程序来充分利用该数据。数据管理方式的这种演变就是企业数据中心的出现，而 Apache Hadoop 是其核心。然而，随着企业机构在这个系统中不断地存储更多的数据，企业所面临的一项新增需求是业务敏感性和法规与监管控制的合规性。该信息要求 Hadoop 提供强大的功能和措施来确保其安全性和合规性。

本文旨在详尽地探析和总结 Hadoop 中常见的安全方法、工具和实践，以及 Cloudera 公司是如何在提供、加强和管理一个企业数据中心所需要的数据安全性方面的得天独厚的优势。

## 安全的重要性

数据安全性是大多数企业中业务和 IT 的头等大事。大多数企业的经营环境都面临着法规遵从性挑战，要求对数据提供控制和保护。例如，零售商和银行服务企业必须遵循支付卡行业数据安全标准（PCI DSS）来保护他们的客户和客户交易、服务和帐户信息。医疗服务提供商和保险公司，以及参与研究的生物技术和制药公司都必须遵循美国健康保险携带和责任法案（HIPAA）中规定的针对患者信息的合规性标准。而且，美国联邦组织和机构必须遵循 2002 年联邦信息安全管理法案（FISMA）的各种信息安全需求。

企业机构也有自身的业务任务和目标，旨在建立内部信息安全标准来保护他们的数据。例如，企业可能需要做出巨大的努力，将保持严格的客户隐私作为一种差异化的市场优势和业务资产。其它企业可能将数据安全性视为保护自己经过艰苦努力赢得的知识产权或政治敏感信息的一堵墙。这些内部政策也可以看作是应对由于违规或事故暴露造成的负面的公众关系和负面新闻报道的一种保险机制。

## 成长的烦恼

从历史上看，早期的 Hadoop 开发人员对于数据安全性并未特别重视，因为在这段时间内 Hadoop 的使用仅限于小规模的内部受众，并且是专为那些未被视为极端敏感和纳入监管的工作载荷和数据集而设计的。由此产生的早期控制措施旨在解决用户错误 - 例如，防止意外删除 - 而不是防止数据被滥用。为了实现更强大的保护功能，Hadoop 依靠现有的数据管理和网络基础架构固有的安全功能。随着 Hadoop 的不断成熟，该平台内的各种方案（例如 HDFS、MapReduce、Oozie 和 Pig），可以用来解决他们自身的特殊安全需求。正如分布式、开源项目的通常情况一样，这些方案构建了旨在配置相同类型的安全控制措施的不同方法。

如今，数据管理环境和需求已大不相同。现在，Hadoop 所具备的灵活、可扩展和具有成本效益的存储功能为企业提供了获取更大范围的数据以便进行分析的机会以及在更长的时间内保留这些数据的能力。数据资产的集中管理使得 Hadoop 自然成为了那些有恶意企图的人员的目标，但是随着当今 Hadoop 中数据安全和安全管理的快速进步，企业拥有强大而全面的功能，

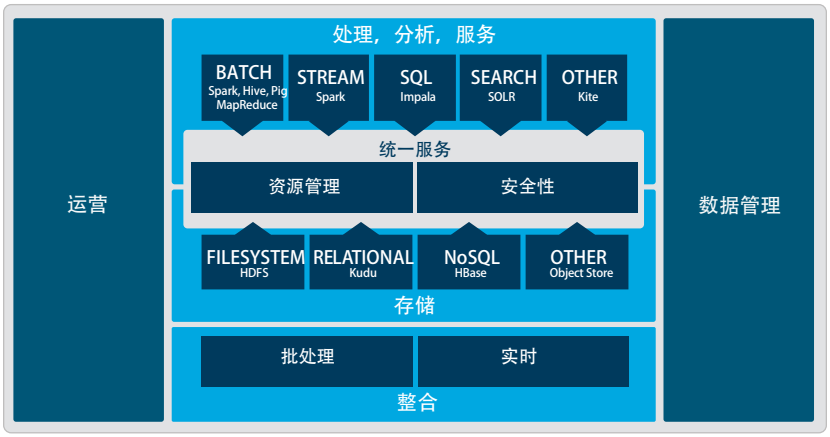
早期的 Hadoop 开发人员对于数据安全性并未特别重视，因为 Hadoop 最初的工作重心是工作载荷和数据集，而未考虑业务敏感性和监管合规性。安全控制解决了用户错误问题，例如，意外删除、非恶意使用。

以防止非法入侵和滥用。

现在，Hadoop 的成熟也意味着拥有很多流入和流出这些信息源的数据路径。企业在 Hadoop 中能够采用多种多样的方式来处理数据，包括交互式 SQL、全文搜索和先进的机器学习分析，确实向企业提供了前所未有的技术能力以及向他们多元化的业务受众和支持者提供了深刻的洞察力。企业安全标准要求对每条路径进行检查，这就对企业的 IT 团队造成了挑战，例如某一些路径和方式可能具有字段、表和视图的概念，而其它路径可能只能够处理与更精细的文件和目录一样的数据。

虽然所有这些功能可以将这些数据开放给更多的用户，提供各种框架来访问和分析数据，这些用户也可以拥有不同程度的必须强制执行的权限。此外，统一的数据存储意味着敏感的数据也会被存储，在必须进行适当的保护同时仍然确保企业能保持一定的灵活性，以便更充分地探索数据，并实现数据的真实价值。

Hadoop 中所固有的强大而深远的能力对于企业而言也是巨大的挑战。虽然 Hadoop 是大数据平台的核心，采用 Hadoop 作为企业数据中心的基础架构要求该平台的数据和流程必须满足企业的安全性需求。如果没有这样的保证，企业根本无法充分发挥企业数据中心的全部潜力，并利用可用于生产环境的企业数据中心来进行数据管理。



## 企业的安全性需求

然而企业怎样才能提供这些保证以确保企业级的数据安全性？非常重要的一点是，在评估 Hadoop 的安全性时，不仅仅是安装防火墙，需要基于安全性的四大支柱对平台进行评估：


由于 Hadoop 在数据、用户和访问框架方面庞大的规模特性，至关重要的一点是在各种系统和数据集之间集中和自动化管理安全控制措施，以便按照信息技术政策提供一致性的配置和应用程序，并方便和简化操作。该要求对于分布式系统而言尤为关键，因为其中部署和管理事件通常贯穿服务和底层基础架构的整个生命周期。

最后，企业需要将所有的系统和数据与现有的 IT 工具（例如，Kerberos）和目录（例如，微软的 Active Directory）或 LDAP，以及日志记录和治理基础架构（例如，SIEM 工具）集成在一起。如果无法充分利用和与现有的安全工具集成的能力，可能非常难以实施新的系统，重复地管理多个工具是非常耗费资源的。

## 安全功能和 Hadoop

在过去几年中 Hadoop 已经显著地成熟起来了，而这种成熟意味着 Hadoop 正在被用于生产以对任务关键型应用程序提供强大的支撑功能。这就是为什么 Cloudera 公司一直专注于增加必要的安全和治理功能，以便在生产环境中运行 Hadoop 和实现数据的真正价值。Cloudera 公司是 Hadoop 综合安全和综合治理领域的领导者。就安全性的四大支柱而言，Cloudera 公司提出的解决方案既保留了 Hadoop 的灵活性，同时为企业提供了所需的符合合规标准的安全性。

本文的剩余章节将重点介绍现在 Hadoop 是如何提供企业级的安全和治理功能的，以及企业 IT 团队是如何在面对频繁的法规遵循审计和治理要求环境中支持 Hadoop 的部署。要做到这一点，对我们有帮助的是更详细地探讨 Hadoop 中体现的企业安全的四个共同因素的技术基础、设计模式以及做法：边界、访问、可见性和数据保护。

## 边界保护：验证

身份验证减轻了未经授权的服务使用的风险，正如在其它系统中的一样，Hadoop 中身份验证的目的就是为了证明某一个用户或某一项服务的身份确是其人或其事。

通常情况下，各企业将通过一个统一的分布式系统管理身份、配置文件和证书，例如轻型目录访问协议（LDAP）目录。LDAP 认证包括非常直接的用

户名 / 密码服务，由多种存储系统提供支撑，包括文件和数据库存储系统。

另一种常见的企业级身份认证系统是 Kerberos。Kerberos 提供了强大的安全优势，包括使攻击者无法使用的拦截认证报文的功能。Kerberos 几乎消除了早期版本的 Hadoop 中存在的身份假冒威胁，并且绝不会通过网络毫无保护地发送用户证书。

这两种身份认证方案都已经被广泛使用超过 20 年的时间，并且许多企业应用程序和体系框架都在其基础架构上利用这两种方案。例如，微软的 Active Directory 就是一种 LDAP 目录，同时其也向 Kerberos 提供了额外的身份认证安全性。事实上，几乎所有的 Hadoop 生态系统中的组件都倾向于使用 Kerberos 身份认证方案来管理和存储 LDAP 或 AD 中证书。

## 模式与实践

Hadoop 由于其自身的分布式特性，存在许多切入点、服务和操作流程，因此需要强大的身份认证功能，以便集群能够安全地运作。Hadoop 提供了实现这一目标的关键设施。

### 边界身份认证

第一种常见的方法关系到对 Hadoop 集群本身的安全访问：Hadoop 的外部或边界的访问。如上所述，在 Hadoop 内通常存在两种主要形式的外部认证，包括 AD/LDAP 和 Kerberos。

AD/LDAP 和 Kerberos 是企业内部的主要认证产品。通过 AD/LDAP，可以管理系统环境中的用户、组群和服务。用户在登录时提供用户名和密码认证，而组群用以控制用户可以访问哪些服务。例如：如果有一个组名称为 HR，并有一个薪酬管理系统，HR 组中的用户可以访问薪资信息，而非 HR 组中的用户则无法访问薪资信息。

最佳实践中建议客户端到 Hadoop（client-to-Hadoop）服务使用 Kerberos 进行身份认证，而且正如前面提到的，几乎所有的生态系统客户端现在都支持 Kerberos。当 Hadoop 的安全配置为使用 Kerberos 时，客户端应用程序通过 Kerberos RPC 对集群边缘进行身份认证，而 Web 控制台使用 HTTP SPNEGO Kerberos 进行身份认证。尽管客户端使用不同的方法来处理 Kerberos 票据，他们都充分利用了一个共同的核心 Kerberos 以及相关的用户和组的中央目录，从而保持整个集群及之外的所有服务的身份认证连续性。

通过利用现有的标准进行身份验证，Cloudera 公司不仅集成了 AD 和 Kerberos，而且通过 Cloudera Manager 可自动完成许多繁琐的设置过程。AD 中有一个内置的 Kerberos 服务器，Cloudera Manager 直接与其集成在一起 – 因此无需安装一个单独的 Kerberos 服务器。这样可使用户能够继续直接对 AD 进行身份认证，可以在 AD 中直接定义 Hadoop 服务作为整体服务管理层的一部分，并且可以通过 AD 组来控制用户对 Hadoop 服务的访问。

此外，Cloudera Manager 能够自动执行 Kerberizing 集群的流程。通过向导，已授权用户能够配置单独服务器的 Kerberos（使用推荐的默认设置），并触发自动工作流程以确保该单独服务器的集群的安全性。Cloudera Manager 也可用于调整用于集群的 Kerberos 服务器和主要 Kerberos 服务的配置，并监控他们的所有服务情况。最后，Cloudera Manager 可以管理和部署 Kerberos 客户端配置

并拥有 Hadoop SSL 相关配置。

为进一步支持 IT 团队实现访问集成，Cloudera 公司更新了 Hue 和 Cloudera Manager，通过使用安全断言标记语言（SAML）进行身份认证以提供额外的单点登录（SSO）选项。

### 集群和 RPC 认证

几乎所有的内部 Hadoop 服务都使用 Kerberos RPC 进行彼此之间的相互身份认证。这些内部的检查机制可以防止通过身份假冒的方式将恶意服务引入到集群活动中，并随后捕获潜在的关键数据作为分布式作业或服务的成员。

### 分布式用户帐户

很多的 Hadoop 项目都可以在 HDFS 内的数据集上进行操作，而此项活动要求在集群内所有的处理和存储节点（NameNode、DataNode、JobTracker/Application Master 和 TaskTracker/NodeManager）上存在一个给定的用户帐户，以便验证访问权限并提供资源分割。主机操作系统提供了用于用户帐户传播的机制 - 例如，在 Linux 主机上的可插拔式认证管理器（Pluggable Authentication Manager，PAM）或系统安全服务守护进程（System Security Services Daemon，SSSD） - 允许可以通过 LDAP 或 AD 集中管理所有的用户帐户，这样有利于开展所需要的进程隔离操作。（请参阅“授权：进程隔离”章节）

## 访问控制：授权

授权关注的是谁或什么对于给定的资源或服务拥有访问或控制权限。Hadoop 整合了之前五花八门的多种分离 IT 系统的各种功能，并在其中添加了各种开源和商业化软件工具。这些不同的数据访问方法在多粒度层次对数据进行处理。Hadoop 管理员面临的挑战是能提供一组可在所有这些基础架构之间进行操作的授权策略。

例如，Hadoop 和传统数据管理环境中共同的数据流要求在每个处理阶段采用不同的授权控制。首先，商用 ETL 工具可以收集源自各种来源和各种格式的数据，并将这些数据提取到一个公共存储库中。在这个阶段，ETL 管理员具有完全访问权限，可以查看输入数据的所有元素，以验证其格式和配置面向数据的句法分析。然后，将处理后的数据通过与公共存储库之间的 SQL 接口公布给商业智能（BI）工具。商业智能（BI）工具的最终用户将根据他们所分配的角色获得不同的访问数据的权限级别。最后，某一些用户可以通过一种商用的基于 GUI 的工具（负责运行 MapReduce 作业）来查询此相同的数据。

Apache Sentry 可以允许创建在所有这些访问方法中强制执行的策略。

## 模式与实践

Hadoop 中针对数据访问的授权通常包含三种形式：针对文件和目录的 POSIX 式授权，针对服务和资源管理的访问控制列表（ACL），以及针对某些服务的且具备对数据先进访问控制的基于角色的访问控制（RBAC）。针对 Hadoop 中不同组件之间的授权都可以越来越多地通过一个单一的工具即 Apache Sentry 实现。例如，数据现在可以在 Hive 和其它作业（例如 MapReduce、Apache Spark 和 Apache Pig）之间进行共享，并且可以通过 HDFS 直接访问此等数据，同时只需要设置一组针对该数据的权限。

cloudera®

## 进程隔离

作为一种共享服务环境，Hadoop 依赖于集群内计算进程彼此之间的隔离，从而可以严格保证每一项进程和每一个用户针对一组给定的资源都具有明确的授权。这一点在 MapReduce 内尤为重要，由于一项给定作业的任务（Tasks）可以在主机服务器上运行 Unix 进程（即 MR 流），独立的 Java VMs，甚至是任意代码。

在与 MapReduce 类似的体系框架内，任务代码是通过使用作业（Job）所有者的 UID 中主机服务器上执行的，因此在 Hadoop 集群的操作系统级别上提供了可靠的进程隔离和资源分割。

## Hadoop 中基于角色的访问控制（RBAC）

由于存在众多的访问路径和用户，访问策略的集中管理才非常关键，因此用户可以访问做他们的作业所需要的数据。Cloudera 5.x 提供了 Apache Sentry（孵化中）中统一的授权。Sentry 是一个开源代码的项目，通过细粒度的基于角色的访问控制（RBAC）对 Impala、Apache Hive、Apache Spark、MapReduce、Apache Pig 和 Search 提供了统一的授权。Apache Sentry 已经迅速成为 Hadoop 中开放性的授权标准，对于维持持续性的工程质量，可移植性的多厂商支持，以及第三方兼容性整合做出了广泛的贡献，所有这些已经被业界广泛采用，即使是在最为规范的垂直行业中。

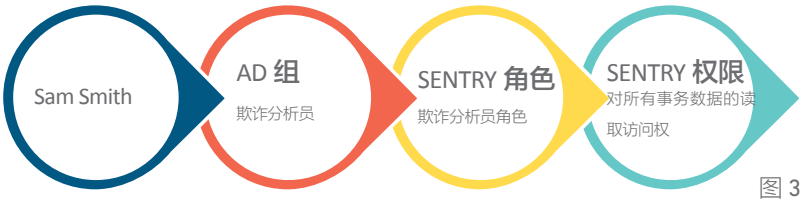



图 3

图 3 显示了 Sentry 授权是如何工作的流程概述。其中，设定了一个 Sentry 角色（欺诈分析员角色），并且该角色具有以下权限 - 对于本例，具备对所有事务数据的读取访问权，然而该权限既可以是针对整个数据库的访问权限，也可以是针对一组特定的列和一组特定的行的访问权限。与上述边界安全的定义相同，AD 中还存在一个称之为欺诈分析员的组，与 Sentry 欺诈分析员角色相关联。Sam Smith 是此 AD 组的一名成员，继承了该 Sentry 角色和相应的权限。

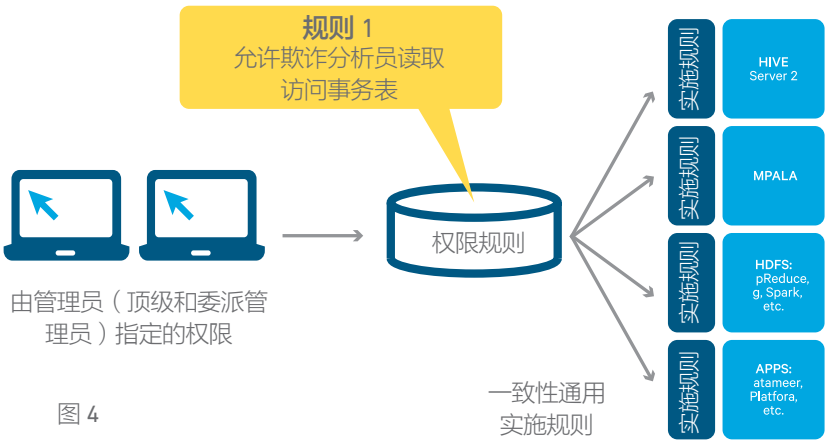



图 4

Sentry 可以被配置为使用 AD 来确定用户的组分配，因此在 AD 中组分配的任何变化都将被 Sentry 自动采集到，继而更新 Sentry 角色分配。这样就消除了两者之间的手动镜像，允许企业机构继续利用现有的 AD 工具和技能组。

为了管理 Sentry 策略，Cloudera 公司提供了一种 UI，可以在 HUE 中完成所有功能。用户可以进入并选择表（TABLE）或数据库（DATABASE），定义角色和权限，或者创建组关联 - 而所有这一切都可以在一个统一的可视化界面内完成。

图 5

### HDFS 权限

Hadoop 生态系统中的许多服务，从与 CLI shell 类似的客户端应用程序到使用 Hadoop 的 API 编写的工具，都可以直接访问存储在 HDFS 中的数据。针对目录、文件和 HDFS ACL，HDFS 采用 POSIX 式的权限；每一个目录和文件都可以被分配成多个用户和组。每一个分配都有一组基本可用的权限；文件权限只包含基本的读、写、执行，而目录具有一项额外的权限来确定对子目录的访问权。当文件级粒度足够好时，以及当无需使用 SQL 组件（例如，Hive 和 Impala）来共享数据时，HDFS 权限 /ACL 就已经足够了。

Sentry 与 HDFS 集成在一起，使客户能够方便地在 Hive、Impala 以及所有与 HDFS（MapReduce、Spark、Pig、Sqoop、等等）进行互动的 Hadoop 组件之间共享数据，同时确保只需要设置一次用户的访问权限，并且确保一致地执行这些权限。

例如，具备 Hive 表 SELECT 访问权限的用户将自动具有该表的所有基本文件的读取访问权。

### 访问控制列表

除了每一项服务和 HDFS 中的数据之外，Hadoop 还保持服务本身的通用访问控制权。服务访问控制列表（ACL）的控制范围从 NameNode 访问到 client-to-DataNode（客户端到 DataNode）的通信。如果是 MapReduce 和 YARN 的情况，用户和小标识符构成了确定作业提交或修改权限的基础。

Apache HBase 也采用访问控制列表（ACL）来进行数据级的授权。HBase 访问控制列表（ACL）通过列、列族、列族限定符以及单元格级元素来对各种操作（包括读、写、创建、管理）进行授权。HBase 访问控制列表（ACL）可

授予和撤销给用户和组。

## 提升可见性：Hadoop 的数据治理

对于集群中的数据，非常关键的一点是了解数据来自哪里以及该数据是如何被使用的。这是在审计中面临的经典问题。审计的目的是捕获系统内所有活动的完整的和不可改变的记录。审计中企业内三大主要活动方面发挥着非常关键的作用。

首先，审计是一个系统的安全机制的一部分，并且可以解释在发生违规或其它恶意企图的情形时，将发生什么，何时，何时或者何事。例如，如果一个流氓管理员删除了一个用户的数据集，审计提供了该动作的细节信息，并且可从备份中检索到正确的数据。

第二大活动是合规性，审计旨在满足敏感性数据或个人身份信息（PII）相关的法规的核心要求，例如健康保险流通与责任法案（Health Insurance Portability and Accountability Act，HIPAA）或支付卡行业（Payment Card Industry，PCI）数据安全标准。审计提供了构建何人、如何、何时以及多久生成、查看和操作数据的踪迹的必要的切入点。

最后，审计提供了数据取证所需的历史数据和背景情况。审计信息将有助于我们了解各种不同的群体如何各种使用不同的数据集，并可以有助于建立针对这些数据集的访问模式。这种检查（例如趋势分析）在范围上比合规性审计更为广泛，可以帮助内容和系统所有者实现他们的数据优化和提升投资回报率。

审计面临的风险是可靠、及时和防篡改性地捕获所有的活动，其中包括行政活动。直到最近，本地 Hadoop 生态系统也主要依靠使用日志文件。日志文件对于企业中大多数审计使用案例来说是不可接受的，因为实时监控是不可能实现的，并且日志结构是不可靠的 - 在系统崩溃之前或日志提交写入期间可能会导致日志的完整性遭受破坏和数据发生丢失。其它替代方案已经加入了特定于应用程序的数据库和其它单一用途的数据存储设施，但是这些方法无法捕捉整个集群的所有活动。

Cloudera Navigator 是针对 Apache Hadoop 的唯一的本地端到端的治理解决方案。Cloudera Navigator 设定了管理员、分析员和数据科学家等角色，并且对于 Hadoop 中存储的海量的不同数据提供了统一的可视性。

## 模式与实践

Cloudera Navigator 是 Cloudera 公司平台的核心组成部分，对于满足合规性要求是至关重要的。其主要功能包括：

综合审计

Cloudera Navigator 将 HDFS、Impala、Hive、HBase 和 Sentry 的完整的审计历史保留在同一个地方。Cloudera Navigator 跟踪记录了每一次访问尝试，一直延伸到用户 ID、IP 地址和全查询文本。这提供了谁正在访问什么数据的可视性，提供了查看某一时间点的权限、它们如何改变（利用 Sentry）以及审查和验证 HDFS 权限的能力。同时，Cloudera Navigator 还提供了即装即用的集成功能，具备业界一流的企业元数据、数据沿袭和 SIEM 应用程序。

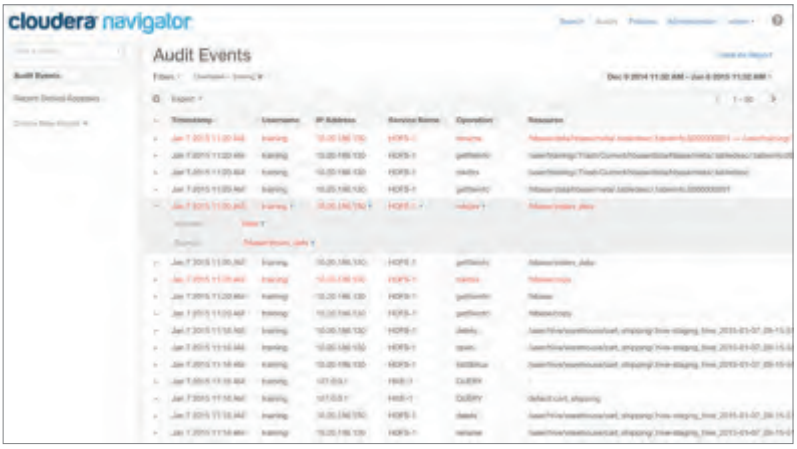



图 6

## 元数据管理

对于元数据和数据发现而言，Cloudera Navigator 具备完整的元数据存储功能。首先，Cloudera Navigator 将 Hadoop 内部所有数据的技术性元数据整合成一个单一的、可搜索性接口，并且可以根据进入集群的外部数据源对数据进行自动标记。例如，如果存在一个外部 ETL 过程，数据在其进入 Hadoop 时就可以被自动标记。其次，Cloudera Navigator 支持基于用户的标记功能，以增加具备自定义业务环境、标签和键 / 值对的文件、表格和单独列。综上，这样可以很容易地发现、分类和定位数据，不仅支持数据治理与合规性标准，而且还方便 Hadoop 内的用户发现。

Cloudera Navigator 还包括元数据策略管理功能，其可以根据到达或预定的间隔触发特定数据集的操作（例如元数据的自动分类）。这样用户能够轻松地设置、监控和执行数据管理策略，同时，也可以与常见的第三方工具集成在一起。

## 数据沿袭

Cloudera Navigator 提供了上行和下行数据沿袭的自动采集和方便的可视化，以验证其可靠性。对于每一个数据源来说，Cloudera Navigator 显示了该数据源内部的列级信息，其确切的上行数据来源是什么，产生该数据源所执行的变换，以及该数据对下游工件的冲击。


图 7

## 数据保护

数据保护的目的是确保只有被授权的用户才可以查看、使用和创建数据集。这些安全控制规定不仅增加了终端用户对于潜在威胁的另一层保护，而且也阻止了网络上或数据中心中管理员和恶意行为人的攻击。这意味着实现这一目标包含两大要素：一方面，保护保存在磁盘或其它存储介质中的数据，通常被称为“静态数据”；另一方面，保护从一个进程或系统转移到另一个进程或系统的数据，通常被称为“传输中数据”。

除了在本文中讨论的其它安全控制规定之外，多个通用的合规性法规都要求提供数据保护。Cloudera 公司通过 TLS 和其它机制提供了传输中数据的安全性保护 - 通过 Cloudera Manager 进行集中部署。通过 HDFS Encryption、Navigator Encrypt 和 Navigator Key Trustee 的综合组合，实现了透明的静态数据保护。

## 模式与实践

Hadoop 作为一个企业机构内多种数据的中央数据中心，自然具有不同的数据保护需求，这都取决于受众和流程，同时使用一个给定的数据集。只有 Cloudera 公司方可提供企业级的加密和密钥管理功能。

CDH 提供了透明的加密功能，可以确保所有的敏感性数据在被存储在磁盘之前都进行加密。此功能配合上 Navigator Key Trustee 的企业级的加密密钥管理功能，提供了必要的数据保护，以满足大多数企业的合规性要求。

HDFS Encryption 与 Navigator Encrypt（Cloudera Enterprise 可支持该功能）一起提供了 Hadoop 中对于数据和元数据的透明加密。这些解决方案能够自动地对数据进行加密，而集群将继续照常运行，并且对于性能的影响非常小。其具有高度可扩展性，能够支持针对所有数据节点的并行加密操作 - 随着集群的发展，加密措施也随之增强。此外，Cloudera 的透明加密特性针对英特尔（Intel）芯片组进行了优化，已实现高性能优势。英特尔（Intel）芯片组包括 AES-NI 协处理器，提供了特殊的功能，使得加密工作载荷的运行速度极快。并非所有的 AES-NI 都是相同的，而 Cloudera 能够充分利用最新的英特尔（Intel）技术进步实现更快的性能优势。

此外，HDFS Encryption 和 Navigator Encrypt 具备职责分离的特性，防止甚至是 IT 管理员和根用户访问他们无权查看的数据。

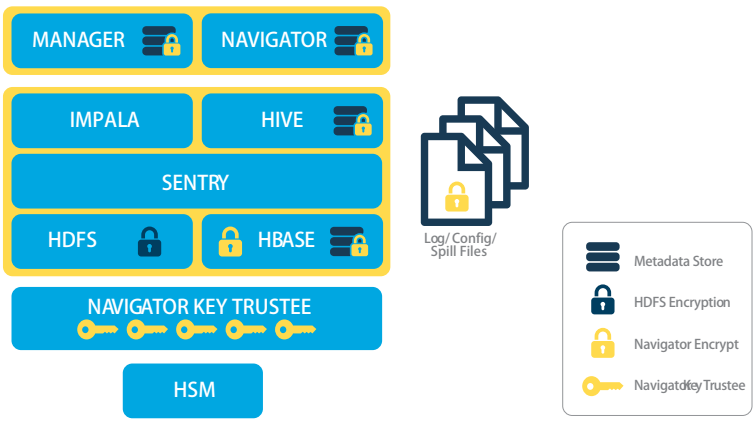



图 8

在某些情况下，如果一家企业机构的安全性需求超出了合规性规定，他们可能会选择实施除了加密之外的控制措施，例如静态数据屏蔽和标记化。Cloudera 公司的合作伙伴提供了上述两个选项。

下表对这些不同的方法进行了比较，展示了透明加密是如何用作数据合规性的基准保护的，并且针对特殊的使用案例可能添加了其它选项。

| | 数据类型 | 用途 | 实施 |
|---|---|---|---|
| 标记化 | • 信用卡号<br>• SSN | • PCI-DSS范围缩减 | • 截取应用程序中的数据流<br>• 修改代码来调用标记化API |
| 静态数据屏蔽 | • 信用卡号<br>• SSN<br>• 姓名<br>• 地址 | • 生成测试数据<br>• 准备发送给第三方的数据集 | • 识别所有被屏蔽的字段的位置<br>• 处于透明位置的数据，直到屏蔽批处理作业运行 |
| 透明加密 | 数据和元数据：<br>• 所有的数据类型<br>• 所有的数据格式<br>（结构化数据、非结构化数据、已发现数据、未被发现数据） | • 确保在磁盘上未曾存在敏感性数据<br>• 满足合规性要求 | • 无需进行整合<br>• 通过Cloudera Manager启用透明加密 |

### 密钥管理

加密操作需要提供加密密钥。密钥的生成、存储和访问都需要进行妥善管理，以避免出现安全漏洞或数据丢失的情况。例如，密钥万万不可与数据存储在一起，这样即使万一密钥被盗窃之后被加密的数据仍然没用。Cloudera Navigator Key Trustee 针对安全加密密钥和其它的 Hadoop 安全性工件提供了企业级的密钥管理功能。

Navigator Key Trustee 是一个基于软件的密钥管理系统，提供了针对 Navigator Encrypt 的密钥管理功能。为了让 Hadoop 满足合规性法规要求，需要将加密密钥存储单独的专用机器上，保持在集群之外并且与数据隔离 - 这就是 Navigator Key Trustee 的用途。同时，Navigator Key Trustee 还集成了一流的硬件安全模块（HSM），因此 Navigator Key Trustee 可以使用 HSMs 作为系统环境中的所有密码技术的信任根。这也意味着围绕 HSM 构建的所有密匙管理策略将继续在 HSM 内执行。

虽然 Navigator Key Trustee 的主要使用案例是存储加密密钥，但是其本质上是一个集中式虚拟保险箱，可以在 Hadoop 架构中存储与工件相关的任何敏感性安全信息。

## 遵循合规性

对于许多行业而言，存储敏感性数据还意味着需要遵循各种法规要求，例如用于存储信用卡数据的 PCI 和用于存储患者医疗记录的 HIPAA。这些法规规定不仅意味着系统在存储这些数据时必须符合规范的要求，并且与其集成在

一起的任何系统都必须遵循相同的要求。

Cloudera 是第一个也是唯一一个满足 PCI 合规性要求的 Hadoop 平台。在充分利用安全性四大支柱的基础之上，Cloudera 平台有充分的能力通过最常见的合规性法规的技术审查，并且在后续过程中其职员中将有专家为用户提供帮助。

## 总结

企业级数据中心作为下一代数据管理设施，需要具备认证、授权、治理和数据保护控制功能特性，从而建立起一个存储和操作企业内所有数据的地方，从批处理到交互式 SQL 直至高级分析。Hadoop 就是其基础，但是 Hadoop 的核心要素仅仅只是一套完整的企业解决方案的一部分。随着更多数量的数据和更多种类的数据迁移到企业数据中心中，则更有可能的是这些数据需要遵循合规性和安全性要求。因此，Hadoop 中全面而综合的安全性就变成了建立新的数据管理中心的基石。

Cloudera 的企业数据中心具有内置的综合、透明和遵循合规性的安全性解决方案。在安全性的四大支柱中，Cloudera 提供了一整套的安全和治理功能，这在 Hadoop 环境和生态系统中无法比拟的。

## 关于 Cloudera

Cloudera 公司提供了现代化的数据管理和分析平台。全球领先的企业机构都信赖 Cloudera 公司可以帮助他们通过利用 Cloudera 企业版产品，以解决他们所面临的最具挑战性的业务问题。Cloudera 企业版是 Cloudera 公司开发的基于 Apache Hadoop 的最快速、最便捷、最安全的数据平台。我们的客户可以有效地捕捉、存储、处理、分析大量的数据，使其能够利用先进的分析技术以迅速、灵活、更低成本的方式推动业务决策。为了确保客户取得成功，我们将为客户提供全面的支持、培训和专业服务。

欲了解更多信息，请登录访问：cloudera.com。

cloudera.com
**电话：**+86-21-62369001
**地址：上海长宁区延安西路 2299 号世贸商城 26 楼 2612 室**