**Hadoop当前安全威胁和防范措施**

**1.摘要**

云服务在全球范围内被广泛用于存储和分析大数据。Hadoop作为处理大数据的最流行的平台之一，具有成本低，便捷，速度快的特点。 但是，随着越来越多的企业和个人在其中存储和处理其私人数据，这也是数据泄露攻击的重要目标。 如何调查Hadoop中的数据泄露攻击是一个重要但长期被忽视的问题。本文首先介绍Hadoop中的一些可能的数据泄露攻击。 并分析了一个新的基于EnCoRe系统的大数据安全系统，该系统使用粘滞策略，以及Verizon和Twilio现有的安全体系结构，提出了这些公司采取的预防措施，以尽量减少安全问题，然后根据一些模拟 案例提出调查框架并进行测试。

**关键词：云服务，大数据，Hadoop，安全**

**1.1介绍**

在2008年由计算社区联盟三位成员撰写的一篇论文中，大数据被公认为是过去十年中最大的计算创新，很快成为外行熟悉的术语，因为其潜力在医疗保健，公共部门管理，零售业，制造业和个人位置数据等等。 随着社交媒体的普及和开放式API的引入，对数千TB数据的存储和分析成为计算机科学领域更为突出和迫切的挑战。

2009年被认为是计算机科学领域里程碑式的一年。 在今年，Web 2.0达到了相当程度的受欢迎程度，并激励微软和谷歌等技术巨头提供基于浏览器的企业解决方案。尽管亚马逊实际上已经在2006年发布了被认为是第一个云计算服务的EC2 / S3 ，标志着云技术的起飞。

2011年12月，Hadoop（最初于2005年作为雅虎Nutch搜索引擎项目的支持发布版本）发布了Release 1.0。 这个版本是第一个具有安全支持的版本，这个版本看起来很明显，也是必不可少的。 Hadoop已经成为利用集群横向扩展环境存储，处理和分析大数据的最尖端技术之一。Hadoop被全球广泛使用，包括Adobe，Facebook和Spotify在内的各种大公司。 根据国际数据公司的统计，预计在2016年的软件市场将会扩大到为8.128亿美元。

Hadoop是大数据存储和分析最流行的平台之一。由于其强大的处理能力，巨大的存储容量，可扩展性和相对较低的成本，广泛应用于制造业，医疗保险，保险，零售等诸多领域。 如今，越来越多的个人和企业在Hadoop上存储和处理他们的私人数据，这些有价值的数据已经成为黑客的重要目标。

为了防止这些类型的攻击，调查它们并重建整个场景是非常重要的。 基于取证结果，可以发现Hadoop的脆弱性，攻击者可以被指控。虽然Hadoop的调查人员经常面临很多挑战，但在这方面的研究工作还很少。 目前Hadoop研究人员面临的挑战包括但不限于如何在数千个Hadoop节点中定位数据泄露节点，如何在复杂且快速变化的Hadoop环境中获取可靠的证据，以及如何调查基于Hadoop的攻击审计日志，通常包含大量的冗余和多用户数据。

本文讨论了当前面向大数据存储和分析的云服务（包括Hadoop）所面临的安全威胁，介绍Hadoop中可能发生的一些数据泄漏攻击，然后分析调查的难点。之后，提出了一个基于模拟案例的调查框架。 该框架由数据收集器和数据分析器组成。 数据采集器主动或根据需要从每个节点收集Hadoop日志，Fsimage文件，我们自己的监控日志和其他信息，并将其传输到数据分析器。 然后，数据分析器利用自动方法分析数据，找到被盗数据，找到窃取这些数据的攻击者，重构犯罪场景。以及针对这些威胁采取的预防措施。

**1.2相关工作**

目前的Hadoop安全研究主要包括可信审计机制，访问控制，数据加密等。 在[1]中，架构是建议针对存储在HDFS（Hadoop分布式文件系统）中的数据针对高级持续威胁（APT）。 该体系结构基于可信平台组（TPM）和可信计算组（TCG）。 在这个架构的帮助下，用户触发的所有操作都可以被审计。 通过这种方式，可以发现可疑的行为，并可以找回证据以供将来调查。但是，这会对性能和日志的大小造成严重的影响，使其变得不切实际。 在[2]中，采用了一种复杂的访问控制机制来保证Hadoop的安全。 它保护Hadoop中的数据免遭未经授权的访问，意外泄漏和丢失以及违反租户机密性。 “ACL访问控制”和“Kerberos”是Hadoop新版本所采用的两种。 虽然这可以使Hadoop更安全，但如果犯罪分子使用合法的用户帐户，则无法防止攻击，而且在操作系统层无法直接进行数据访问，因为这些机制仅适用于应用程序层。

在[3]中，提出了一种安全的Hadoop架构，在HDFS中增加了加密和解密功能。 通过这种方法，HDFS中的数据即使被窃取也是不可读的，因为攻击者没有密钥。 数据以这种方式得到保护。 尽管这是保护Hadoop的根本性解决方案，但不能忽视的是对性能的影响，以及这种方法无法防范使用合法帐户的攻击者。

与[1]不同，我们的工作集中在Hadoop中调查（包括证据收集和分析）数据泄露攻击。 虽然[2]和[3] 提出了一些很好的解决方案来防止数据泄露，这远不是银弹。 这种类型的攻击仍然经常发生，因此研究如何在事件发生后对其进行调查是必要和重要的。 我们在这个问题上还没有找到任何其他的工作。

1. **大数据平台与模型面临的安全问题**

**2.1 大数据**

随着社交媒体比以往任何时候都更受欢迎，以及众多其他行业的数据量不断增加，大数据分析对于大型企业和小型企业来说变得越来越重要。大数据的分析提供了有助于增长的非常有用的商业洞察力，纠正了非生产性实践，突出强项和弱点等优点。现在，BigData越来越多地使用云服务进行存储和分析，虽然大数据的收集和分析可以产生非常积极的业务影响，但是如果没有适当的安全措施，也可能会证明它是极度的破坏性的。

数据存储和分析的绝对数量对于加密来说是一项具有挑战性的任务。 加密和解密大量数据的成本很高，影响了查看和使用数据的整体速度。 但是，加密对于避免云框架中数据集的混合以及避免可能导致数据失窃或丢失的漏洞是必要的。 一旦数据被加密和存储，应该在不解密数据的情况下进行分析，以确保客户的安全和隐私.

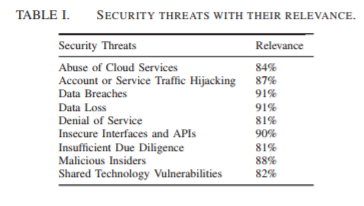
单个数据集中的不同类型的数据可能需要被访问和/或使用不同，并且处理对于各种数据类型的访问控制是困难的操作。 考虑到数据的快速性以及在节点之间没有固定的行进路径，跟踪访问和数据流可能是具有挑战性的，并且呈现出安全性问题。 但是，如果访问和数据流量没有得到适当的控制，数据可能会被盗用和丢失。

如果没有适当的数据流控制，云服务所必需的共享技术环境也存在风险。 如果一个数据集中有恶意或不恶意的恶意行为被允许移动到适当的范围之外，那么存在破坏其他客户数据集的风险，或者最糟糕的情况是会损坏云服务提供商的整个基础架构。 数据量大，速度快也可能导致备份和恢复功能对性能产生不利影响，但是如果要避免数据丢失，则不能将其排除在外。

**2.2云安全**

根据“福布斯”2013年发表的文章，当时美国有一半以上的企业正在使用云服务，而这一数字还在持续增长。随着这种技术的成熟，许多数据，安全和隐私问题，例如数据被窃取的可能性，都被识别出来了，必须加以解决。然而，最近的一份报告显示，用户对云服务提供商采取安全措施的整体信心仍然很低。 客户担心共享资源的基本云原则，缺乏对数据存储位置的控制，以及提供商对于采取何种安全措施缺乏开放性。

原则上，云服务依赖于服务提供商的技术环境。 因此，由服务提供者来解决与安全有关的许多问题和担忧。 表1中显示了由云安全联盟确定的“恶名昭彰的九种云计算威胁”，它们是云安全最紧迫的威胁及其在当今时代的相关性。



**2.3数据泄漏攻击HADOOP**

Hadoop中的数据泄漏攻击主要包括但不限于以下几类。

**应用层数据泄露。**

攻击者可以通过应用程序层漏洞或恶意软件获取私有数据。 例如，当前Hadoop审计机制中的一个漏洞是，它只记录操作类型，时间和内容，但没有关于谁做这个操作的信息。 假设在一家公 司 里 ， Alice ， Bob 和 Cindy 属 于 一 个 名 为Hadoop的组织，它负责管理他们的公司，NY的Hadoop系统。 有一天，Bob在HDFS中存储了一个名为star-project.txt的文件，并将该文件的权限设置为可读。 此后不久，Bob发现他的档案内容被他的对手所知，但内容没有被他泄露，那么谁是叛徒，Alice还是Cindy呢？ 我们无法从Hadoop日志中找到答案，因为他们没有记录是谁做的。 我们唯一的线索是Bob知道哪个文件被盗了。 另外，Hadoop审计日志可能已被篡改。

Hadoop框架的两个核心设计是HDFS（Hadoop Distributed File System）和MapReduce。HDFS是一个分布式文件系统，具有高容错性的特点，并且被设计用来部署在廉价的硬件上；而且它能够以高吞吐量来访问应用程序的数据，尤其适合那些有着超大数据集的应用程序。MapReduce是一个使用简易的软件框架，基于它编写出来的应用程序能够运行在由上千个商用机器组成的大型集群上，并以一种可靠容错的方式并行处理上T级别的数据集。

最近出现的针对MongoDB、ElasticSearch和Hadoop的勒索攻击模式都较为相似。在攻击过程中并没有涉及勒索软件和常规漏洞，而是利用相关产品不安全的配置，这为攻击者打开了方便之门。以MongoDB为例，这些受攻击的数据库没有采取任何身份验证，直接暴露在Internet公网上，一旦攻击者登录到这些开放的数据库就可以对其中的数据进行删除等恶意操作了；而针对ElasticSearch服务器的勒索攻击手段也是类似，ElasticSearch的TCP访问模式的默认端口为9300，HTTP访问模式的默认端口为9200，如果这些端口不做任何保护措施地暴露在公网上，那么对它的访问将没有任何身份认证，任何人在建立连接之后，都可以通过相关API对ElasticSearch服务器上的数据进行增删查改等任意操作。而黑客针对Hadoop的勒索攻击，也是利用了暴露在公网上的端口。Hadoop集群的使用者往往出于便利或者本身安全意识不强的缘故，会将Hadoop的部分端口，比如HDFS的Web端口50070直接在公网上开放。

**操作系统层数据泄露。**

这意味着攻击者是否有权限登录Hadoop节点的主机操作系统，他们可以绕过Hadoop的监视器，直接在 数 据 库 中 （ 不 仅 是 HDFS 文 件 块 ， 而 且 是MapReduce任务的临时结果） OS层。 例如，Bob在HDFS中存储了一个名为companyA.7z的文件，这个文件大于64Mb，所以它被分成了几个**块，每个**块可能被保存在不同的机器上。 Hadoop管理员和root用户都拥有对所有块的写入权限。 因此，如果攻击者想窃取文件，他们不必在HDFS中获得权限，只需要借助名称节点日志（或通过网络监视器等其他方式获取每个块的位置等等），如果他们拥有主机操作系统的root权限，则直接从物理机器上窃取数据块。 我们唯一的线索是哪个文件被盗。 Hadoop审计日志也可能被攻击者篡改。

为了调查上述情况，我们正面临着面临以下挑战。 首先，Hadoop集群通常包含数百或数千个节点。 当发生攻击时，调查所有节点几乎是不可能的，因为调查过程中需要的时间太长或证据已经改变。 大多数攻击只涉及少数节点，我们需要找到一种有效的方法来定位受到攻击的节点，并将资源集中在这些节点上。 其次，Hadoop环境复杂而且快速变化。 证据可能被新数据污染或被攻击者篡改。 那么，如何从这样的环境中获得可靠的证据呢？ 第三，目前的Hadoop审计日志在调查方面远远不够，因为他们缺乏关于谁是犯罪的信息。 此外，这些审计日志中包含大量的冗余和多用户数据，使得证据分析非常困难。

1. **防范措施**

**3.1 大数据**

本文提出一个用于保护大数据应用程序的粘性策略框架的体系结构。

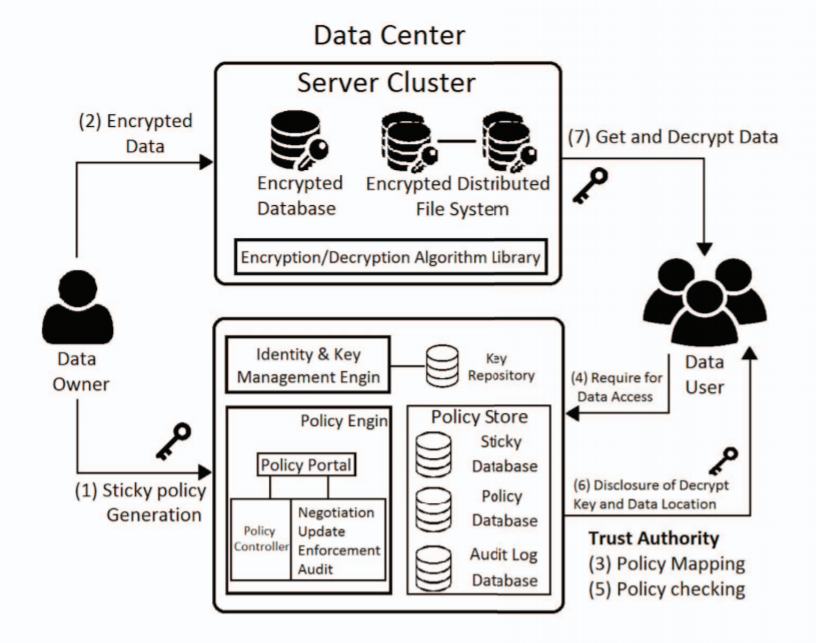
1. **粘滞政策**

Li李等人根据数据隐私管理项目EnCoRe ，提出了一个独特的架构框架来帮助保护大数据应用。EnCoRe执行粘性政策，对数据应用限制和条件，以确定使用限额和义务。这样可以控制项目内多个边界对机密数据的访问和披露。 EnCoRe用公钥加密作为管理粘滞策略的核心机制。

Li和他的合作者提出的大数据框架使用了松散耦合（loose-couple）绑定，一种将粘性策略和数据碎片存储在不同地方的方法。 这增加了基础设施的合规性，并使系统更难以违反。 该体系结构有两个子域，即可信权限域和数据中心域，目的是保持粘性策略和数据片段的合法性。 可信权限域拥有身份和密钥管理引擎和策略引擎，而加密数据位于数据中心域内。

关于用户的信息，包括他们的认证和授权信息，以及给定用户对每个数据片段持有的权限被存储在身份和密钥管理引擎中。 策略引擎作为整个域的核心。 这个引擎保持对正在访问的数据的控制，并跟踪各方在访问数据方面的权限。 最重要的是，存储的数据是加密的，只有在策略被批准，接受和满意之后才能被访问。

策略引擎有几个子组件：策略门户，策略控制器，策略协商组件，策略更新组件，实施组件和策略存储。 策略门户作为引擎的入口，接收数据访问请求，并将最终响应发回给用户。 策略控制器决定拒绝请求或将其转发给响应组件。 诚如其名称所述，政策协商组件协商政策，将政策与有权执行的用户数据库相匹配。 可以通过策略更新组件更新安全策略。 执行组件检查数据用户是否满足了所需的粘滞策略。 最后，策略存储在数据和访问它所需的策略之间保持一个映射关系，并维护所有激活策略和访问数据的审计日志。 为了实现其功能，策略库支持三个子数据库，策略数据库，粘性数据库和审计数据库。简单地说，策略数据库存储策略规则，粘性数据库存储粘滞策略和它们“卡住”的数据之间的映射，审计数据库记录和跟踪数据访问。 该框架提供了一个理论上彻底的设计，以在收集，分析，存储和共享大数据的领域内产生更安全的交易。 这个框架对于那些与存储医疗记录的医院或执法组织分析绝密调查报告等机密数据的用户尤其相关。



**3.2云服务**

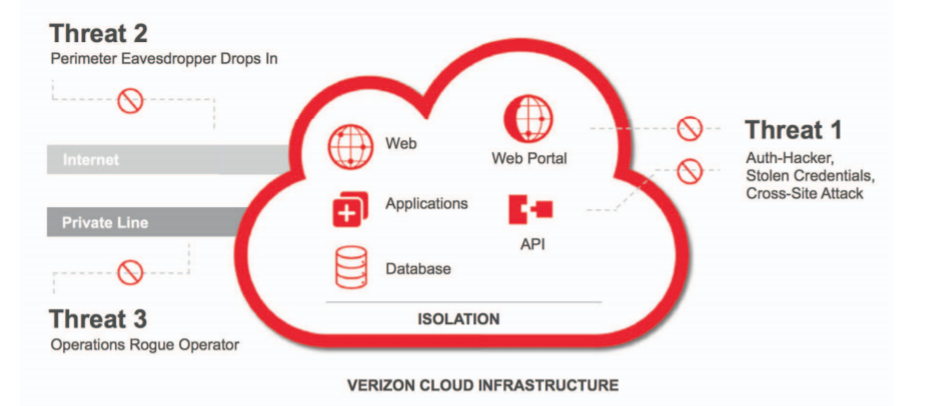
针对云服务，研究了究Verizon的基础设施，如图1所示.Verizon使用分层安全模型来保护其云，这是一种向客户提供的服务，用于存储在设备中共享的个人数据和信息。 这四个层次是基本安全，逻辑安全，增值安全以及治理，风险和合规。

**A.基本安全**

这一层关注于独立于技术本身的物理或外部安全。Verizon已经努力确保他们的数据中心拥有最高水平的物理安全，利用90天视频支持的全天候视频监控，生物识别阅读器，不间断的现场保安服务以及入境运输安全等措施。只接受预先通知的软件包。 此外，访问控制通过“需要知道”和“逐个事件”协议来维护。 这些措施有助于通过硬件的物理攻击来降低物理数据被盗和数据丢失的风险。员工定期接受有关最新安全程序的培训。 所有员工也必须通过背景调查，以降低恶意的内部攻击风险。

1. **逻辑安全**

逻辑安全级别确保网络，资源和数据的机密性，完整性和可用性。 这个级别进一步分为计算，网络，存储和管理子层，共同协作保护基础设施。 在计算子层中，实现了密码策略，操作系统安全，管理员和用户认证等措施。网络子层负责分布式拒绝服务（DDoS）检测和缓解，固件级别的“MAC-in-MAC”封装以及集成的防火墙功能。 存储子层支持使用对**称**AES-256密码加密数据，并通过SSL和高级消毒技术进一步增强安全性。 这一层还允许用户通过Microsoft SQL和Oracle添加数据库加密。 最后，管理子层处理身份和访问控制。 Verizon希望在其下一个版本中支持安全声明标记语言（SAML）2.0框架。



**图1.Verizon Cloud Infrastructure**

1. **增值安全**

除了基本和逻辑安全性外，该层还封装了Verizon支持的一些附加安全功能。 这些功能包括具有高度可定制结构的防火墙和VPN功能，包括大数据应用程序和软件开发功能的预配置安全解决方案，以及能够检测安全漏洞并识别缓解选项的智能管理系统。 增值安全层还为专用IP网络提供了一个选项，该专用IP网络支持端到端环境，以实现与用户工作负载的安全连接。

1. **治理，风险和合规**

这一层确保了前三个层次中的所有安全措施都符合完善的和知名的标准。 使用敏捷开发技术完成更新，增强功能和错误修复。 敏捷开发的本质是强有力的控制，确保快速的创新。 更新和调试过程中还使用高端更改管理来支持验证，回滚过程和先决条件。

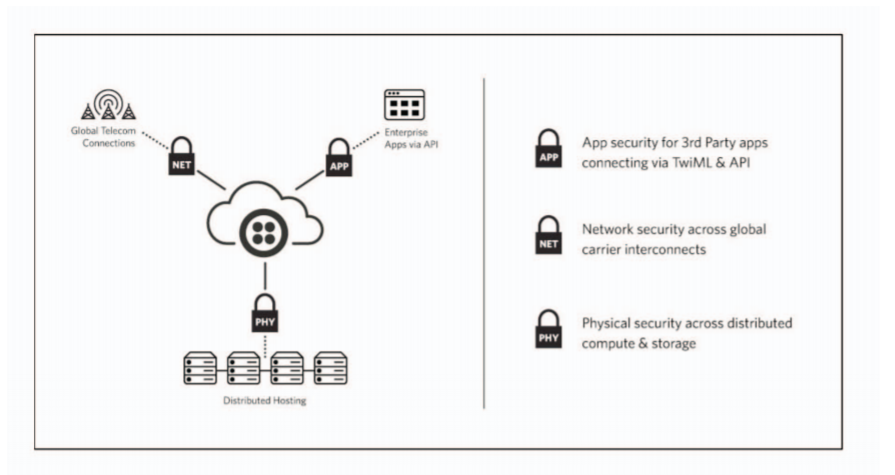
**3.3HADOOP**

Hadoop的开源大数据分析框架自推出以来已经越来越受欢迎。 随着Hadoop普及率的提高，Hadoop的安全性不断受到严格的审查，特别是考虑到并发处理架构。 针对身份验证漏洞，Hadoop通过Kerberos引入了中央身验证。 在2013年，英特尔推出了“犀牛计划”，通过提供对各种加密和认证方法的支持，加强了Hadoop框架的安全性。 但是，基于数据湖传输的数据量和关键信息的大量增加，仍然需要加强Hadoop框架的安全措施，以防止网

络犯罪。 Hadoop框架支持传统分布式系统以及作为云环境一部分的系统。 本文重点介绍通过云连接的系统的安全漏洞和预防措施。 Hadoop的体系结构允许将文件分解成块，分布在各个节点上同时处理。 但是，在云环境中，定位保存给定数据块的特定节点是非常困难的。 该系统的一个主要安全威胁是关键数据的保护，这需要采取额外的安全措施。 正如第二节所提到的，单一数据集中的安全级别的多样性使安全程序变得复杂。 Hadoop块系统只能解决这个问题。 此外，节点可以包含来自多个文件块并具有管理权限。 “块窃取”和“块注入”，在没有适当的许可的情况下进行数据修改或修改是另外两个主要的安全问题。

本节研究Twilio的安全体系结构，这是一家云计算通信公司，它以使用Amazon S3服务可靠地实现Hadoop框架而闻名。 Twilio使用多租户通信平台来确保资源权限的隔离，如图3所示。该策略进一步扩展以识别关键数据并执行必要的安全检查。 为了了避免“块窃取”，Twilio使用基于作业角色的访问控制，并结合Amazon S3的存储桶策略和访问控制列表。 Twilio遵守安全港标准有助于第三方供应商的高度信任，其中许多供应商被要求具有单独的标准安全合规性。 进一步的攻击，如拒绝服务和分布式拒绝服务，通过维护冗余的DNS服务器和使用iptables阻止流氓IP，以及通过支持AWS专有的DDoS缓解技术，最大限度地减少。 数据的物理盗

窃仍然是一种可能性，但是可以避免使用类似于Verizon在保护其Cloud以及多个不同地理区域中的多个冗余区域时使用的基本安全性措施。 整体安全性通过第三方公司的渗透测试每六个月评估一次。



**一个hadoop取证框架**

实际上，它也可以扩展到调查Hadoop中的其他类型的攻击。 这个框架由许多数据收集器和一个数据分析器组成。 数据收集器位于Hadoop节点（主要是数据节点）中的主机操作系统的内核中。 它是根据需求收集政策设计的。 首先，它主动监视每个节点上对重要数据的访问，并将这些行为日志传送给数据分析器。 然后收集所需的证据（如磁盘图像 -es和网络流量日志）根据分析结果从特定的节点。数据分析器位于特定的取证服务器中。 利用自动算法对数据进行分析，找到被盗数据，找到偷窃这些数据的攻击者，重构整个场景。 它还发送命令给特定的收藏家以获得更多的证据。

上述设计有ser advantages的优势。 通过数据采集器和数据分析器的配合，实时监控和自动分析算法可以立即检测到数据泄漏攻击，并根据采集到的行为日志中的信息快速定位发生的节点。 需求集合策略可以降低成本，因为大规模的证据（如磁盘映像）只能根据分析结果在多个节点上收集。 活动的内核级监视器确保在污染之前收集证据，并且收集器本身对于正常的Hadoop用户和攻击者是透明的。 而且，所有的证据都存储在一个独立的，受到仔细保护的法医服务器中。 这使得证据更可靠。

1. **数据收集器**

数据收集器的体系结构如图3所示。它收集的数据包括Hadoop日志，Fsimage文件，我们自己的监视器（即HProgger）日志以及文件，进程，网络和系统的映像或日志。 在实时监控阶段，每个节点上的数据收集器首先获取Hadoop日志，HProgger日志和fsimage。 然后，日志过滤器将其更改为统一的格式。 最后，在将它们传送到服务器之前，日志加密器将它们加密以保持其可信性。 在需求采集阶段，数据采集器甚至一些其他的取证工具将采集文件日志，网络日志，进程日志和系统映像，然后传送给服务器进行分析。这些数据也可以在传输之前加密。 两阶段设计使我们的框架能够根据调查需要收集数据，同时降低消耗和提高效率。

Hadoop日志包括名称节点日志，作业跟踪器日志，数据节点日志和任务跟踪器日志。 我们的框架将主要集中在名称节点日志和数据节点日志上，因为数据泄露攻击主要记录在其中。Fsimage是Hadoop系统的映像文件。 与一般系统映像文件不同，fsimage只记录数据目录和数据分布。 简而言之，fsimage记录了文件系统中文件的元信息和块的分布情况。HProgger是我们自己的监视工具，它是基于Progger 实现的。 Progger是一个在Linux内核中运行的开源数据监视工具。 它修改系统调用表中某些系统调用的地址，允许它记录这些调用的调用。 Progger能够帮助我们获得更可靠**的证据，因为**Hadoop中的数据泄漏攻击将不可避免地调用某些系统调用来访问某些文件和目录。尽管Progger在审计方面很强大，但却导致了严重的业绩下滑，并拖**延**了每一项操作。 此外，Progger生成的日志数量也是不可接受的，也就是说，它只需几分钟即可生成数百兆字节的日志。为了使其适用于Hadoop取证，必须提高性能，并尽可能减少日志的数量。 因此，我们设计HProgger（即Hadoop-Progger）。

它在几个方面得到了改进：

1.只有三个系统调用（即OPEN，COPY，

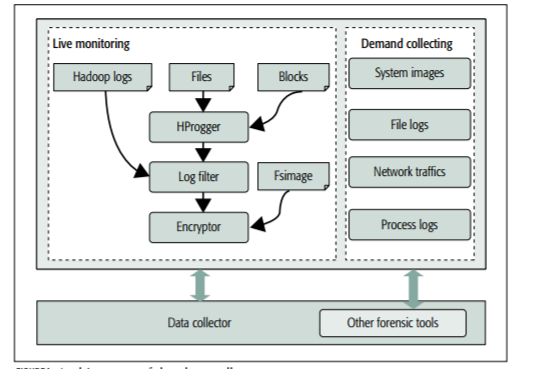
MOVE）将被录制。

2. 日志格式被最小化; 只剩下操作类型，时间，

操作员等必要的信息。

3. 而不是监视所有的目录，只保留那些与

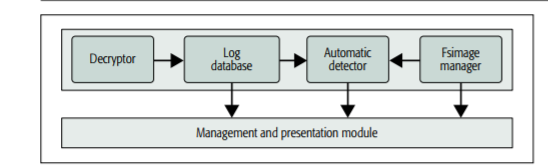
Hadoop系统相关的目录，比如Hadoop存储HDFS文件的目录。 因此，与Progger相比，HProgger更轻，更适用于Hadoop取证。



**图3..数据收集器的体系结构**

1. **数据分析器**

数据分析器由一个解密器，一个日志数据库，一个fsimage管理器，一个自动检测器以及管理和表示模块组成（如图2所示）。 数据到达分析器时，首先由解密器进行处理。 然后将Hadoop日志和HProgger日志存储在日志数据库中，并将fsimage文件存储在fsimage管理器中。 用户可以 在 管 理 和 演 示 模 块 中 查 询 和 管 理 日 志 和fsimage。 此数据在演示之前格式化。 自动检测器自动分析日志，查找被盗数据，窃取数据的攻击者以及攻击过程。 结果也将显示在管理和演示模块中。



**图4.数据分析器的体系结构**

为了尽快发现数据泄露攻击，设计了一些自动检测算法。 对于OS层数据泄露攻击，检测算法分为四个维度：异常目录，异常用户，可疑块比例和异常操作。 该算法检测文件数据集中的异常目录和异常用户，判断是否存在可疑块，并计算出可疑块的比例。 同时，该算法通过监测块数据中的运行趋势来检测异常值，并计算可疑块的比例。 如果这四个维度中的任何一个发出警告，则可能发生攻击，调查人员可以根据警告信息继续调查。

**异常目录（AD）：**

正常Hadoop系统文件操作将涉及固定目录，因为这些目录是由Hadoop系统配置文件配置的。 因此，在收集的日志中，如果找到超出此范围的任何目录，就可能发生攻击。 通过分析包含这些异常目录的记录，可以发现可疑块。异常用户（AU）：通过研究HProgger日志，可以得出结论，除了将文件从HDFS复制到本地机器外，所有的Hadoop系统文件操作都只涉及一个操作员，即Hadoop超级用户。 因此，如果在HProgger日志中找到其他用户而是Hadoop超级用户，则可能发生攻击。 通过分析包含这些异常目录的记录，可以发现可疑块。另一方面，如果攻击者想直接从操作系统窃取数据，则必须具有root权限或Hadoop超级用户权限。 但是，由于不止一个用户可以切换到根目录，那么如何才能找到以root权限执行操作的真实用户呢？ 幸运的是，只要HProgger登录到系统，HProgger就可以跟踪所有用户的数据，因此可以分析HProg-ger日志以找到真正的用户。

**异常操作（AO）：**

每个HadoopHDFS中的系统操作由读取HDFS文件和将文件写入HDFS组成。 通过研究各种Hadoop系统操作产生的HProgger日志，可以得出结论：每个块只写入一次，在其生命周期中至少读取两次（不考虑该块的其他备份）。 但是，如果攻击者想要从物理机器上窃取一个数据块，他们需要将其复制（移动）到另一个目录，有时可能会重命名该数据块，因此如果发现这些操作中的任何一个，就可能发生攻击。 而且攻击场景可以根据这些操作进行重构。 攻击者和丢失的文件也可能基于它们被识别。

**块比例（BP）：**

让SusBlockNum（文件 -ID ） 表 示 文 件 数 据 集 中 可 疑 块 的 数 量 ，AllBlockNumber（FileID）表示文件数据集中块的总数。 我们将块比例（BP）定义为通过AllBlockNum-ber （ FileID ）划分 SusBlockNum（FileID）的结果。 在正常情况下，BP的值应该是0，因为不应该发现可疑的块。 BP越大，发生攻击的概率就越大，所以如果BP大于0，算法会给出警告。

以上描述了OS层数据泄露攻击的自动检测算法。 根据HProgger日志，检测应用程序层攻击要容易得多。 例如，对于使用合法用户帐户的窃取行为，由于缺少必要的信息，我们无法找到基 于 Hadoop 审计日志的小偷。 但是基于HProgger，盗贼可以按照以下步骤进行识别：

步骤1：根据。找到HDFS块文件名并从Hadoop日志或fsimage文件中获取它们的块ID。

第2步：筛选出包含的HProgger日志copyToLocal”或其他必要的键操作类型。

第3步：过滤掉包含的Hadoop日志在步骤1中找到的块ID，然后根据操作时间将这些日志映射到步骤2中获取的HProgger日志。 如果找到这样的映射记录，可以在相关的HProgger日志中找到攻击者名称，这些映射记录可以相互支持。

1. **总结**

大数据收集和分析以及Hadoop等云计算服务正在日益广泛和全球范围内得到利用。 即使在过去的几年里，数据量的大幅度增长也意味着越来越多的技术正在被开发来收集，分析和存储这些数据，每个数据都有自己的新漏洞和安全威胁。 在今天的新闻中，从零售商店到政府数据库的大量实体数据泄露事实证明，在开发这些大数据系统时，安全性不可能是事后考虑的事情。 今天，在确定现有系统中的安全问题方面还有不断的工作，并制定了减少或消除这些安全问题的措施。 随着新系统的开发在最初的设计和实施中把安全作为重中之重是至关重要的。虽然安全是优先的，但是新的安全漏洞肯定会被识别出来。此外，完全制止网络犯罪的统一制度尚未设计，尽管像粘性政策架构这样的理论体系的实施可能会有很大的改进。大数据安全领域必将在技术世界中继续发挥重要作用，因为云服务越来越受到人们的欢迎，从为个人数据备份提供移动客户到大型企业大规模存储和分析大数据

。