（蓝色字体标出的代表专业词汇）

摘要：

As one of the most popular platforms for processing big data, Hadoop has low costs, convenience, and fast speed. However, it is also a significant target of data leakage attacks, as a growing number of businesses and individuals store and process their private data in it. How to investigate data leakage attacks in Hadoop is an important but long-neglected issue. This article first presents some possible data leakage attacks in Hadoop. Then an investigation framework is proposed and tested based on some simulated cases.

Hadoop（主语提前）作为处理大数据最流行的平台之一，具有低成本、方便快捷的特点（增词）。然而，随着越来越多的企业和个人在其中存储和处理他们的私人数据，它也是数据泄漏攻击的一个重要目标。如何研究Hadoop中的数据泄漏攻击是一个重要但长期被忽视的问题。本文首先介绍了Hadoop中一些可能的数据泄漏攻击。在此基础上，提出了一个调查框架，并通过模拟案例进行了验证（被动句式转换为主动句式|拆句）。

简介：

Hadoop is one of the most popular platforms for big data storage and analysis. It is widely used in many fields, such as manufacturing, healthcare, insurance, and retail because of its powerful processing capacity, huge storage capacity, scalability, and relatively low cost. Nowadays, a growing number of individuals and businesses store and process their private data in Hadoop, and this valuable data has become an important target of hackers.

Hadoop是最流行的大型数据存储和分析（修饰成分提前作定语）平台之一。它以其强大的处理能力、巨大的存储容量、可扩展性和相对低廉的成本（汉语中先因后果），在制造业、医疗保健、保险、零售等领域得到了广泛的应用。如今，越来越多的个人和企业在Hadoop中存储和处理他们的私有数据，这些有价值的数据已经成为黑客的重要攻击（增词）目标。

In order to prevent these types of attacks, investigating them and reconstructing the entire scenario is very important. Based on the forensic results [5] [6], the vulnerability of Hadoop can be found and the attackers can be accused. Although the Hadoop investigator often faces many challenges, there is still little research work in this area. Current challenges faced by the Hadoop investigator include, but are not limited to, how to locate the data leakage node among thousands of Hadoop nodes, how to obtain reliable evidence in a complex and rapidly changing Hadoop environment, and how to investigate attacks based on Hadoop audit logs, which usually contain a large quantity of redundant and multi-user data.

为了防止这些类型的攻击，对它们进行调查并重建整个场景是非常重要的。根据电子取证结果[5][6]，可以发现Hadoop的漏洞，并对攻击者进行指控（被动句式转换为主动句式）。尽管Hadoop研究者经常面临许多挑战，但在这方面的研究工作仍然很少。Hadoop调查员目前面临的挑战（被动句式转换为主动句式|主语提前）包括但不限于如何在数千个Hadoop节点中定位数据泄漏节点，如何在复杂且快速变化的Hadoop环境中获取可靠证据，以及如何基于Hadoop审核日志调查攻击，它通常包含大量的冗余和多用户数据。

Considering these challenges, in this article, we first present some possible data leakage attacks in Hadoop, then analyze the difficulties of investigation. After that, an investigating framework is proposed and tested based on simulated cases. The framework is composed of the data collector and data analyzer. The data collector collects Hadoop logs, Fsimage files, our own monitor logs, and other information from each node actively or on demand, and transmits them to the data analyzer. Then, the data analyzer analyzes the data with automatic methods to find the stolen data, find the attacker who stole this data, and reconstruct the crime scenario.

考虑到这些挑战，本文首先介绍了Hadoop中可能存在的数据泄漏攻击，然后分析了研究的难点。在此基础上，提出了一个基于模拟案例的调查框架并进行了测试（被动句式转换为主动句式）。该框架由数据采集器和数据分析器组成。数据采集器从每个节点主动或按需收集（修饰成分提前）Hadoop日志、Fsimage文件、我们自己的监视器日志和其他信息，并将它们传输到数据分析器。然后，数据分析器使用自动方法对数据进行分析，以查找被盗数据，找到窃取此数据的攻击者，并重建犯罪场景。

相关工作：

Current Hadoop security research mainly includes trusted audit mechanisms, access control, data encryption, and so on. In [1], an architecture was proposed to fight against advanced persistent threats (APT) targeted against data stored in the HDFS (Hadoop distributed file system). This architecture is based on the trusted platform group (TPM) and trusted computing group (TCG). With the help of this architecture, all of the operations triggered by users can be audited. In this way, suspicious actions can be discovered and evidence can be retrieved for future investigation. However, it results in a serious impact on performance and a huge size of logs, which make it impractical.

目前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，我们（增词）提出了一种体系结构（被动句式转换为主动句式）来对抗针对HDFS（Hadoop分布式文件系统）中存储的数据的高级持久性威胁（APT）。该体系结构基于可信平台组（TPM）和可信计算组（TCG）。在这种体系结构的帮助下，用户触发的所有操作（被动句式转换为主动句式）都可以被审计。这样，就可以发现可疑的行为，检索证据（被动句式转换为主动句式）以备将来调查之用。但是，它会对性能造成严重影响，并且会导致日志的巨大规模，这使得它不切实际。

In [2], a complicated access control mechanism was adopted to ensure the security of Hadoop. It protects data in Hadoop from unauthorized access, accidental leakage and loss, and breach of tenant confidentiality. “ACL Access Control” and “Kerberos” are two of them that have been adopted by the new versions of Hadoop. Although this can make Hadoop more secure, it is not able to prevent attacks if criminals use legal user accounts, and nothing can be done for direct data access in the operating system layer because these mechanisms are only for the application layer.

在文[2]中，为了保证Hadoop的安全性，采用了一种复杂的访问控制机制（被动句式转换为主动句式）。它保护Hadoop中的数据免受未经授权的访问、意外泄漏和丢失，以及违反租户机密性的操作（增词）。“ACL访问控制”和“Kerberos”是Hadoop新版本所采用的（修饰成分（定语从句）提前翻译）两种方法。尽管这可以使Hadoop更加安全，但（增词）如果罪犯使用合法的用户帐户，它就无法防止攻击，而且无法在操作系统层直接访问数据（被动句式转换为主动句式|否定转移），因为这些机制仅适用于应用程序层。

In [3], a secure Hadoop architecture was proposed that adds encryption and decryption functions in the HDFS. Through this method, data in the HDFS will not be readable even if it is stolen because the attacker does not have the secret key. Data is protected in this way. Although it is a fundamental solution for securing Hadoop, what cannot be ignored is the impact on performance and the fact that this approach does not guard against attackers who use a legal account.

在[3]中，提出了一种安全的Hadoop体系结构（被动句式转换为主动句式），在HDFS中增加了加密和解密功能。通过此方法，因为（汉语中先因后果）攻击者没有密钥，所以即使HDFS中的数据被窃取，也无法读取。以这种方式保护数据。尽管它是保护Hadoop安全的基本解决方案，但（增词）不可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。

Unlike [1], our work concentrates on investigating (including evidence collection and analysis) data leakage attacks in Hadoop. Although [2] and [3] present some good solutions to prevent data leakage, they are far from silver bullets. This type of attack still takes places frequently, so research on how to investigate them after they happen is necessary and important. We have not found any other work on such an issue.

与[1]不同，我们的工作集中在调查（包括证据收集和分析）Hadoop中的（修饰成分提前）数据泄漏攻击。尽管[2]和[3]提出了一些很好的防止数据泄漏的（修饰成分提前）解决方案，但它们远不是（表示否定意义的词）灵丹妙药。这种类型的攻击仍然频繁发生，因此研究如何在攻击发生后对其进行调查是必要的（删词）。在这个问题上，我们还没有找到其他方法。

Hadoop中的数据泄漏攻击：

The data leakage attacks in Hadoop mainly include but are not limited to the following categories.

Hadoop中的数据泄漏攻击主要包括但不限于以下几类。

Application layer data leakage. This means attackers can obtain private data by application-layer vulnerability or malware. For example, a vulnerability in the current Hadoop audit mechanism is that it only records the operation type, time, and content, but no information on who did this operation. Suppose in a company, Alice, Bob and Cindy belong to a group named Hadoop, which is responsible for managing their company’s Hadoop system. One day, Bob stored a file named star-project.txt in the HDFS, and set the permission of this file to group readable. Soon after that, Bob found the content of his file was known by his rival, yet the content was not leaked by him, so who is the traitor, Alice or Cindy? We cannot find the answer just from Hadoop logs because they do not record who did this. The only clue we have is that Bob knows which file was stolen. In addition, the Hadoop audit logs might have been tampered

应用层数据泄漏。这意味着攻击者可以通过应用层漏洞或恶意软件（修饰成分提前）获取私有数据。例如，当前Hadoop审计机制中的（修饰成分提前）一个漏洞是，它只记录操作类型、时间和内容，而不记录谁执行了此操作（否定转移）。假设在一家公司中，Alice、Bob和Cindy属于一个名为Hadoop的组，该组负责管理他们公司的Hadoop系统（拆句）。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组可读。不久之后，鲍勃发现他的档案内容被他的对手知道了，但他并没有泄露任何内容（被动句式转换为主动句式），那么谁是叛徒，爱丽丝还是辛迪？我们无法仅仅从Hadoop日志中找到答案，因为它们没有记录是谁做的。我们唯一的线索是鲍勃知道哪个文件被偷了。此外，Hadoop审计日志可能已被篡改。

Operating system layer data leakage. This means if attackers have the permission to log-in to the host operating system of a Hadoop node, they can bypass the monitor of Hadoop and steal data (not only the HDFS file block but also the temporary result of the MapReduce task) directly in the OS layer. For example, Bob stored a file named companyA.7z in the HDFS, and this file is larger than 64 Mb, so it was divided into several blocks, each block possibly being saved in different machines. Both the Hadoop administrator and the root user have write permission to all the blocks. Hence, if the attacker wants to steal the file, they do not have to get permission in the HDFS, they just need to get the locations of every block with the help of the name node logs (or through other ways such as the network monitor and so on), then steal the block directly from the physical machine if they have root permission of the host OS. The only clue we have is which file was stolen. The Hadoop audit logs might also have been tampered with by the attackers.

操作系统层数据泄漏。这意味着，如果攻击者有权限登录到Hadoop节点的主机操作系统，他们可以绕过Hadoop的监控，直接在OS层窃取数据（不仅是HDFS文件块，也是MapReduce任务的临时结果）。例如，Bob在HDFS中（修饰成分提前）存储了一个名为companyA.7z的文件，该文件的大小超过64mb，因此它被分成几个块，每个块可能保存在不同的机器中。Hadoop管理员和根用户都拥有对所有块的写权限。因此，如果攻击者想要窃取文件，他们不需要在HDFS中获得权限，（删词）只需要借助名称节点日志（或通过网络监视器等其他方式）（修饰成分提前）获取每个块的位置，然后如果拥有主机操作系统的根权限，他们（改变主语位置）能直接从物理机上窃取块。我们唯一的线索是哪个文件被偷了。Hadoop审计日志也可能被攻击者篡改。

To investigate the cases above, we are faced with the following challenges. First, Hadoop clusters often contain hundreds or thousands of nodes. When attacks happen, it is almost impossible to（固定句式，it is无需翻译） investigate all the nodes because either it requires too much time or the evidence has been changed during the investigations. Most attacks involve only a few nodes, and we need to find an efficient way to locate the nodes being attacked and concentrate our resources on them. Second, the Hadoop environment is complex and fast changing. The evidence may be tainted by new data or tampered with by attackers. Hence, how can we obtain reliable evidence from such an environment? Third, current Hadoop audit logs are far from sufficient in investigations as they lack the information on who did the crimes. Furthermore, these audit logs contain a large quantity of redundant and multi-user data, making evidence analysis extremely difficulty.

为调查上述案件，我们面临以下挑战。首先，Hadoop集群通常包含数百或数千个节点。当攻击发生时，几乎不可能对所有节点进行调查，因为这需要太多的时间，或者在调查过程中（修饰成分提前）证据会发生变化。大多数攻击只涉及几个节点，我们需要找到一种有效的方法来定位被攻击的节点，并将我们的资源集中在这些节点上。其次，Hadoop环境复杂且变化迅速。证据可能被新数据污染或被攻击者篡改。因此，我们如何从这样的环境中获得可靠的证据？第三，目前的Hadoop审计日志在调查中远远不够，因为它们缺乏犯罪者的信息。此外，这些审计日志包含大量的冗余和多用户数据，使得证据分析极为困难。

总结：

①专业词汇较多

例子：

Application layer data leakage应用层数据泄漏

trusted audit mechanisms 可信审计机制

forensic results 电子取证结果

②英语中的被动句式在汉语中多转换为主动句式

例子：

A complicated access control mechanism was adopted to ensure the security of Hadoop.

为了保证Hadoop的安全性，采用了一种复杂的访问控制机制。

An investigating framework is proposed and tested based on simulated cases.

提出了一个基于模拟案例的调查框架并进行了测试。

In [1], an architecture was proposed to fight against advanced persistent threats (APT) targeted against data stored in the HDFS (Hadoop distributed file system).

在[1]中，我们提出了一种结构来对抗针对HDFS（Hadoop分布式文件系统）中存储的数据的高级持久性威胁（APT）。

In this way, suspicious actions can be discovered and evidence can be retrieved for future investigation.

这样，就可以发现可疑的行为，并检索证据以备将来调查之用

③英语中常常果前因后，但在汉语中常常因前果后

例子：

It is widely used in many fields, such as manufacturing, healthcare, insurance, and retail because of its powerful processing capacity, huge storage capacity, scalability, and relatively low cost.

它以其强大的处理能力、巨大的存储容量、可扩展性和相对低廉的成本（汉语中先因后果），在制造业、医疗保健、保险、零售等领域得到了广泛的应用。

Data in the HDFS will not be readable even if it is stolen because the attacker does not have the secret key.

因为攻击者没有密钥，所以即使HDFS中的数据被窃取，也无法读取。

④修饰成分多数情况下需要翻译在被修饰词之前，英语中的修饰成分多用在被修饰词之后，而汉语中的修饰成分多用在被修饰词之前

例子：

This is a vulnerability in the current Hadoop audit mechanism.

这是当前Hadoop审计机制中的一个漏洞。

The evidence has been changed during the investigations.

在调查过程中证据会发生变化。

⑤否定转移，英语中的否定多用在主语或宾语中，汉语中的否定多用在谓语中

例子：

No information on who did this operation.

不记录执行了此操作的人。

Nothing can be done for direct data access in the operating system layer.

无法在操作系统层直接访问数据。

⑥需要改变主语位置，有时需要将主语提前至句子的开头，否则会“头重脚轻”，有时也需要将主语提后；

例子：

As one of the most popular platforms for processing big data, Hadoop has low costs, convenience, and fast speed.

Hadoop作为处理大数据最流行的平台之一，具有低成本、方便快捷的特点。

Hence, if the attacker wants to steal the file, they do not have to get permission in the HDFS, they just need to get the locations of every block with the help of the name node logs (or through other ways such as the network monitor and so on), then steal the block directly from the physical machine if they have root permission of the host OS.

因此，如果攻击者想要窃取文件，他们不需要在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，然后如果拥有主机操作系统的根权限，他们能直接从物理机上窃取块

⑦并列句或主从句可以拆句翻译

例子：

Suppose in a company, Alice, Bob and Cindy belong to a group named Hadoop, which is responsible for managing their company’s Hadoop system.

假设在一家公司中，Alice、Bob和Cindy属于一个名为Hadoop的组，该组负责管理他们公司的Hadoop系统。

Then an investigation framework is proposed and tested based on some simulated cases.

在此基础上，提出了一个调查框架，并通过模拟案例进行了验证。

⑧科技论文中的用词非常精炼，大部分情况下无需删词，需要增词的情况可能出现得更多

例子：

This valuable data has become an important target of hackers.

这些有价值的数据已经成为黑客的重要攻击目标。

⑨科技论文中需要翻译的复杂句式不多，要力求逻辑的简洁明了