B12 Computer—Leakage

张昊宇 2019.11.12

An Investigation Framework for Hadoop Forensics

Hadoop取证调查框架

Considering the above challenges, an investigation framework aimed at data leakage attacks in Hadoop was proposed. In fact, it can also be extended to investigate other kinds of attacks in Hadoop. This framework is composed of （固定结构，被动语态）many data collectors and one data analyzer. The data collector is located in the kernel of the host operating system in Hadoop nodes (mainly data nodes). It is designed based on a demand collection policy. First, it actively monitors the accesses to important data on each node and transmits these behavior logs to the data analyzer. Then it collects the required evidence (such as disk images and network traffic logs) from specific nodes according to the analysis results. The data analyzer is located in a specific forensic server. It analyzes the data with automatic algorithms to find the stolen data, find the attackers who stole this data, and reconstruct the entire scenario. It also sends commands to specific collectors to obtain more evidence.

考虑到上述挑战，我们（补充主语）提出了针对Hadoop中数据泄漏攻击的研究框架（被动转主动）。实际上，它也可以用作（词义扩展）研究Hadoop中的其他类型的攻击。该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。它是根据需求收集策略设计的。数据收集器的工作流程如下：（添加语句使中文语意连贯）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输到数据分析器。然后根据分析结果从特定节点收集所需的证据（先因后果）（例如磁盘映像和网络流量日志）。数据分析器位于特定的取证服务器中。它使用自动算法对数据进行分析，以找到被盗的数据和（合译）偷窃该数据的攻击者，并重建整个场景。它还将命令发送给特定的收集器以获得更多证据。

There are serval advantages in the above design. Through the cooperation between the data collectors and the data analyzer, the data leakage attacks can be detected immediately by the live monitor and the automatic analysis algorithm, and the node where it happened can be located quickly according to the information in the collected behavior logs. The demand collection policy can reduce cost because the large sized evidence (e.g., disk images) is only collected on several nodes based on analysis results. The live kernel-level monitors ensure the evidence is collected before contamination, and the collector itself is transparent to normal Hadoop users and attackers. Moreover, all evidence is stored in an independent and carefully protected forensic server. This makes the evidence more reliable.

上述框架（异译便于理解）有不少优势。通过数据采集器和数据分析器的协作，可以通过实时监控器和自动分析算法立即检测到数据泄漏攻击，并根据收集到的行为日志中的信息快速定位发生泄漏的节点（被动转主动）。仅根据分析结果就可以在在多个节点上收集大型证据（例如磁盘映像）可以为需求收集策略降低成本（先因后果）。实时的内核级监控器确保在复写（专业术语）之前收集证据，并且收集器本身对一般的Hadoop用户和攻击者透明。此外，所有证据都存储在独立且经过精心保护的电子取证（专业术语）服务器中。这些设计（增词）使证据更加可靠。

Data Collector

数据收集器

The architecture of the data collector is shown in Fig. 1. The data it collects includes Hadoop logs, the Fsimage file, our own monitor (i.e., HProgger) logs and images or logs of files, processes, networks, and system. In the live monitoring stage, data collectors on each node obtain Hadoop logs, HProgger logs, and the fsimage first. Then, the log filter changes them to a unified format. Finally, before transmitting them to the server, the log encryptor encrypts them to keep their credibility. In the demand collecting stage, the data collector and even some other forensic tools will collect file logs, network logs, process logs, and system images, and then transmit them to the server for analyzing. This data can also be encrypted before transmission. The two-stage design gives our framework the ability to collect data according to what the investigation needs, with reduced consumption and improved efficiency.

图1展示了数据收集器的体系结构（被动转主动）。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，日志过滤器将它们编译（专业术语）为统一格式。接着日志加密器会对它们进行加密以保持其可信度，最后再将它们传输到服务器中（语序调整）。在需求收集阶段，数据收集器甚至其他一些电子取证工具将收集文件日志，网络日志，过程日志和系统映像，然后将它们传输到服务器进行分析。也可以在传输之前对该数据进行加密（被动转主动）。这种两阶段设计使我们的框架能够根据调查需要收集数据，从而（递进）减少了消耗并提高了效率。

Hadoop logs include name node logs, job tracker logs, data node logs, and task tracker logs. Our framework will mainly focus on name node logs and data node logs because data leakage attacks are mainly recorded in them.

The Fsimage is the image file of the Hadoop system. Unlike general system image files, the fsimage only records the data directory and data distribution. In simple words, the fsimage records the meta information of files in the file system and the distribution of blocks.

Hadoop日志包括名称节点日志，作业跟踪程序日志，数据节点日志和任务跟踪程序日志。我们的框架将主要关注名称节点日志和数据节点日志，因为数据泄漏攻击主要发生（异译）在其中。

Fsimage是Hadoop系统的映像文件。与常规系统映像文件不同，fsimage仅记录数据目录和数据分布（专业术语）。简单来说，fsimage记录的是文件系统中文件的元信息和文件块的分布。

HProgger is our own monitor tool which is implemented based on Progger [4]. Progger is an open source data monitor tool running in the kernel of Linux. It modifies the addresses of certain system calls in the system call table, allowing it to record the invocations of these calls. Progger is able to help us acquire more reliable evidence because data leakage attacks in Hadoop will inevitably invoke some system call to access certain files and directories. Although Progger is powerful in auditing, it causes a serious downgrade in performance and delays every operation. Additionally, the amount of logs that Progger generates is unacceptable, i.e. it generates hundreds of megabytes of logs in just a few minutes. To make it applicable in Hadoop forensics, performance must be improved and the number of logs must be reduced as much as possible. Thus, we design HProgger (i.e. Hadoop-Progger). It isimproved in several aspects:

HProgger是我们自己的基于Progger [4]实现的监视工具（状语从句合译）。Progger是在Linux内核中运行的开源数据监视工具（语序调整）。 它修改了系统调用表中某些系统调用的地址，从而（递进）使其可以记录这些调用的调用。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特定文件和目录。所以Progger能够帮助我们获得更加可靠的证据。（先因后果）尽管Progger的审核功能强大，但它会严重降低性能并延迟每次操作。另外，Progger它在短短几分钟内生成了数百兆的日志，这种巨大的日志数数据不可接受的（语序调整）。为了使其适用于Hadoop电子取证，必须提高性能并且必须尽可能减少日志数量（被动转主动）。因此，我们设计了HProgger（即Hadoop-Progger）。相较于Progger（增句），在几个方面进行了改进：

1. Only three system calls (i.e. OPEN, COPY, MOVE) will be recorded.

2. The log format is minimized; only necessary information such as operation type, time, operator, and so on remains.

3. Instead of monitoring all the directories, only those associated with the Hadoop system, such as the directory where Hadoop stores HDFS files, are retained. Hence, compared to Progger, HProgger is lighter and more applicable for Hadoop forensics.

1.仅记录三个系统操作（即OPEN，COPY，MOVE）（被动转主动）。  
2.将日志格式最小化；仅保留必要的信息，例如操作类型，时间，操作员等（被动转主动）。  
3.保留所有与Hadoop系统关联的目录（例如Hadoop存储HDFS文件的目录），而不是监视所有目录。因此，与Progger相比，HProgger更智能（一词多义），更适用于Hadoop电子取证。

Data Analyzer

数据分析仪

The data analyzer is composed of a decryptor, a log database, a fsimage manager, an automatic detector, and the management and presentation module (as shown in Fig.2). When data arrives at the analyzer, it is first processed by the decryptor. Then Hadoop logs and HProgger logs are stored in the log database and the fsimage files are stored in the fsimage manager. Users can query and manage the logs and the fsimage in the management and presentation module. This data is formatted before presentation. The automatic detector analyzes the logs automatically to find the data that was stolen, the attacker who stole the data, and the process of the attacks. The results will also be presented in the management and presentation module.

数据分析器由解密器，日志数据库，fsimage管理器，自动检测器以及管理和显示模块组成（如图2所示）。数据到达分析器后，首先由解密器处理。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。用户可以在管理和显示模块中查询和管理日志和fsimage。此数据在演示前已经被格式化。自动检测器会自动分析日志，以查找被盗的数据，偷走数据的攻击者以及攻击过程（合译）。管理和显示模块中也将显示结果（被动转主动）。

In order to detect data leakage attacks as soon as possible, some automatic detection algorithms are designed. For an OS-layer data leakage attack, the detection algorithm works in four dimensions: abnormal directory, abnormal user, suspicious block proportion, and abnormal operation. The algorithm detects abnormal directories and abnormal users in the file datasets to find out if suspicious blocks exist, and it calculates the suspicious block proportion. In the meantime, the algorithm monitors the trend of operations in the block datasets to detect abnormal values, and also calculates the suspicious block proportion. If any of these four dimensions gives a warning, then attacks may have happened and the investigators can continue their investigation according to the warning information.

为了尽快检测数据泄漏攻击，设计了一些自动检测算法（被动转主动）。对于操作系统层（专业术语）的数据泄露攻击，检测算法的工作范围包括四个方面：目录异常，用户异常，可疑块比例和操作异常。该算法检测文件数据集中的异常目录和异常用户，以找出是否存在可疑块，并计算出可疑块比例（合译）。同时，该算法监视块数据集中的操作趋势以检测异常值，并计算可疑块比例。如果这四个维度中的任何一个发出了警告，则可能发生了数据泄漏（补充）攻击，调查人员可以根据警告信息继续进行调查。

Abnormal Directory (AD): Normal Hadoop system file operations will involve fixed directories because these directories are configured by Hadoop system profiles. Hence, in collected logs, if any directory that is out of this range is found, an attack may have happened. The suspicious blocks can thus be found by analyzing records that contain these abnormal directories.

异常目录（AD）：正常的Hadoop系统文件操作将涉及固定目录，因为这些目录是由Hadoop系统文件配置的。因此，在收集的日志过程中，如果找到了超出此范围的目录（被动转主动），则可能发生了攻击。因此，可以通过分析包含这些异常目录的记录来找到可疑块。（被动转主动）

Abnormal User (AU): By studying the HProgger logs, it can be concluded that all of the Hadoop system file operations except copying files from the HDFS to a local machine will involve only one operator, i.e. the Hadoop super user.Hence, if any other user instead of the Hadoop super user is found in HProgger logs, an attack may have happened. The suspicious blocks can thus be found by analyzing records that contain these abnormal directories. On the other hand, if an attacker wants to steal data directly from the operating system, they must have root permission or Hadoop super user permission. However, because more than one user can switch to the root, how can the real user who conducted the operations in root permission be found? Fortunately, HProgger is able to keep track of all users’data as long as they log in to the system, so the HProgger logs can be analyzed to find the real user.

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复制到本地计算机外，所有Hadoop系统文件操作将只涉及一个操作员，将其定义为Hadoop超级用户（改译）。在HProgger日志中找到其他用户而不是Hadoop超级用户（被动转主动），则可能发生了攻击。因此，可以通过分析包含这些异常目录的记录来找到可疑块被动转主动。另一方面，如果攻击者想直接从操作系统窃取数据，则他们必须具有root权限或Hadoop超级用户权限。但是，由于有多个用户可以切换到root用户，如何才能找到以root用户权限进行操作的真实用户？幸运的是，HProgger能够跟踪所有用户登录到系统的数据，因此可以分析HProgger日志以找到实际用户（被动转主动）。

Abnormal Operation (AO): Every Hadoop system operation in the HDFS is composed of reading files from the HDFS and writing files into the HDFS. By studying the HProgger logs generated by all kinds of Hadoop system operations, it can be concluded that every block is written only once and read at least twice during its life cycle (other backups of this block are not considered).However, if an attacker wants to steal a block from the physical machine, they need to copy (move) it to another directory and sometimes may rename the block, so if any of these operations are found, the attack may have happened. Moreover, the attack scenario can be reconstructed based on these operations. The attacker and the lost files may also be identified based on them.

异常操作（AO）：HDFS中的每个Hadoop系统操作都由从HDFS读取文件并将文件写入HDFS组成。通过研究由各种Hadoop系统操作生成的HProgger日志，可以得出结论，每个块在其操作周期（专业术语）中仅写入一次并至少读取两次（不考虑该块的其他备份被动转主动）。 要从物理机上窃取一个块，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块，因此，如果找到这些操作中的任何一个，（被动转主动）则可能发生了攻击。此外，可以基于这些操作来重建攻击场景（被动转主动）。攻击者和丢失的文件也可以基于它们来识别。

Block Proportion (BP): Let SusBlockNum (File-ID) represent the number of suspicious blocks in a file dataset and AllBlockNumber (FileID) represent the total number of blocks in a file dataset.We define block proportion (BP) as the result of dividing SusBlockNum (FileID) by AllBlockNumber (FileID). Under normal circumstances, the value of BP should be 0 as no suspicious block should be found. The larger BP is, the greater the probability there has been an attack, so if BP is larger than 0, the algorithm will give a warning.

块比例（BP）：让SusBlockNum（文件ID）代表文件数据集中可疑块的数量，而AllBlockNumber（FileID）代表文件数据集中的总块数。我们将块比例（BP）定义为用SusBlockNum（FileID）除以AllBlockNumber（FileID）。在正常情况下，BP的值应为0，因为不应找到可疑块。 BP越大，发生攻击的可能性越大（从句简译），因此，如果BP大于0，则算法将发出警告。

The above describes the automatic detection algorithm for an OS-layer data leakage attack. Detection of an application-layer attack is much easier based on the HProgger log. For example, regarding stealing behavior using a legal user account, we are not able to find the thief based on Hadoop audit logs because necessary information is missing. However, based on HProgger, the thief can be identifi ed according to the following steps:

上面介绍了OS层数据泄漏攻击的自动检测算法，基于HProgger日志检测应用层攻击要容易得多。例如，如果（介词结构）使用合法用户帐户进行的窃取行为，由于缺少必要的信息，我们无法根据Hadoop审核日志找到窃取者（先因后果）。但是，基于HProgger，可以根据以下步骤识别窃取者：

Step 1: Find the HDFS blocks according to the fi le name and obtain their block ID from Hadoop logs or fsimage fi les.

Step 2: Filter out HProgger logs that contain“copyToLocal” or other necessary key operation types.

Step 3: Filter out the Hadoop logs that contain the block ID found in Step 1, then map these logs to the HProgger logs obtained in Step 2 based on their operation time. If such mapping records are found, the attacker name can be found in the relevant HProgger logs, and these mapping records can support each other.

步骤1：根据文件名称查找HDFS块，并从Hadoop日志或fsimage文件中获取其块ID。  
步骤2：过滤出包含“ copyToLocal”或其他必要键操作类型的HProgger日志。  
步骤3：过滤出包含在步骤1中找到的块ID的Hadoop日志，然后根据它们的运行时间将这些日志映射（专业术语）到在步骤2中获得的HProgger日志。 如果找到了此类映射记录，则可以在相关的HProgger日志中找到攻击者的名称，并且这些映射记录可以相互支持。

总结：

1. 专业词汇较多

如：images 映像 ；forensic 电子取证 ；map 映射

1. 多被动转为主动

如：To make it applicable in Hadoop forensics, performance must be improved and the number of logs must be reduced as much as possible.

为了使其适用于Hadoop电子取证，必须提高性能并且必须尽可能减少日志数量。

1. 英语为先果后因，汉语为先因后果

如：we are not able to find the thief based on Hadoop audit logs because necessary information is missing.

由于缺少必要的信息，我们无法根据Hadoop审核日志找到窃取者。

1. 较少长难句从句（长句但不难）