

云计算取证模型

高元照^{1,2}, 李学娟³, 李炳龙^{1,2}, 吴熙曦⁴

(1. 信息工程大学, 河南 郑州 450001;
2. 数学工程与先进计算国家重点实验室, 河南 郑州 450001;
3. 河南理工大学审计处, 河南 焦作 454010;
4. 白城兵器试验中心, 吉林 白城 137000)

摘要: 针对云取证面临的主要挑战, 分析云计算特性, 提出了一种云取证模型。提出了持续性取证准备服务的部署、基于“迭代”的多轮次证据识别策略和分布式文件系统的多层级数据定位方法、基于“数据隔离”和“按需收集”策略的证据收集方法、基于云计算资源的综合性证据分析平台的建立。最后, 结合云环境下的取证场景, 分析了该模型的有效性。

关键词: 云计算; 云取证; 取证模型; 取证策略; 取证方法

中图分类号: TP311

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2017.00194

Cloud computing forensic model

GAO Yuan-zhao^{1,2}, LI Xue-juan³, LI Bing-long^{1,2}, WU Xi-xi⁴

(1. Information Engineering University, Zhengzhou 450001, China;
2. State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 450001, China;
3. Audit Department of Henan Polytechnic University, Jiaozuo 454010, China;
4. Baicheng Weapon Testing Center, Baicheng 137000, China)

Abstract: Aiming at the main challenges facing cloud forensics, cloud characteristics were analyzed and a cloud forensic model was proposed. In the model, the deployment of persistent forensic readiness service, multi-round identification strategy based on “iteration” and multi-level evidence location method for distributed file system, “data isolation” and “on-demand collection” strategies, and the establishment of comprehensive evidence analysis platform based on cloud resources were presented. Finally, the validity of the model was analyzed combining the forensic scene in the cloud environment.

Key words: cloud computing, cloud forensics, forensic model, forensic strategy, forensic method

1 引言

云计算能够提供按需、弹性、便捷的网络访问和计算资源, 目前已成为 IT 领域的主流技术^[1,2]。但云计算的快速发展也使其成为数字犯罪新的攻击

目标。云安全提供商 Alert Logic 2015 年的云安全报告称, 受到 APP 攻击、可疑行为攻击、暴力破解攻击的云服务提供商 (CSP, cloud service provider) 分别占 70%、68% 和 56%^[3]。云计算安全已成为用户选择云服务的主要关注之一^[1,4]。

收稿日期: 2017-08-02; 修回日期: 2017-09-04; 通信作者: 李炳龙, lbl2017@163.com

基金项目: 国家高技术研究发展计划 (“863”计划) 基金资助项目 (No.2015AA016006); 国家自然科学基金资助项目 (No.60903220)

Foundation Items: The National High Technology Research and Development Program of China (863 Program) (No.2015AA016006), The National Natural Science Foundation of China (No. 60903220)

云计算取证是数字取证科学在云环境中的应用,在威慑、打击云犯罪活动和维护用户权益上发挥着重要作用。2011年6月,我国成立了中国云计算安全政策与法律工作组,并将执法取证作为云安全领域的主要研究方向^[5]。2013年,云安全联盟(CSA, Cloud Security Alliance)发布的《云计算关键领域安全指南 V3.0》^[6]也将云取证作为一项关系云计算发展的重大问题提出。但由于云计算基础设施规模庞大,服务种类多样,并具有分布性、虚拟性和共享性等特点。证据的识别、收集与分析面临巨大挑战。

云取证模型能够引导取证人员从大量用户和云服务数据中识别涉案数据的范围,定位以虚拟化或分布式方式存储数据的物理位置,采取合理可行的证据收集策略与方法,提供大数据和云平台独有数据类型的有效分析方法,对于云取证工作具有重要的指导意义。Ruan等^[7]在全球范围发起了一项关于云取证关键问题的调查,回复的257位数字取证专家中,87%认为设计云取证模型是当前云取证研究的主要方向之一。

然而,由于云计算的固有特性,传统的数字取证模型多针对单个或少量的计算机或服务器,难以适用于云环境。现有的云取证模型多针对云计算的特性和云取证面临的挑战进行理论分析,并就其中的部分问题指出某种应对方法的必要性与可行性,但没有真正结合云取证环境的复杂性和取证场景与对象的多样性,提出综合性、整体性的取证策略与实施方法,不足以指导云环境下的取证工作。

本文针对云取证面临的主要挑战和云取证模型研究的不足,分析云计算的特性以及由此带来的云服务部署模式、攻击方式、证据类型及存储方式等方面的多样性,建立了一种云取证模型,提出了不同取证阶段的任务、目标,以及实现这些目标所制定的取证策略与方法,以指导、规范、监督取证人员完成取证工作。最后,结合云环境下的案例场景,分析了该模型的有效性。

2 云取证面临的主要挑战

云计算能提供对可配置的共享资源池(如网络、服务器、存储等)快捷、按需的网络访问,

能够以极小的管理负担以及与CSP的交互实现资源的快速配置和释放^[8]。云计算具有大规模、虚拟性、分布性、共享性和高可靠性等固有特性,给云取证带来了巨大挑战,是云取证研究需要解决的最主要问题。本节对云取证面临的主要挑战进行详细介绍。

1) 取证数据量大

云计算平台规模庞大,并且易于扩展。一方面,大量的结构化、半结构化和非结构化数据并存,并且很多数据格式是云计算专有的^[9];另一方面,数字设备的数量和容量增长迅速。识别与安全事件相关的数据并定位其物理存储位置十分困难^[10]。证据分析面临巨大负担^[11],现有的取证工具在云平台独有数据格式的分析上存在很大局限性。

2) 数据存储的分布性

云环境下多采用分布式文件系统(DFS, distributed file system)作为主要的文件系统^[12]。DFS分布式的特性导致文件被分割成数据块,存储在不同的数据中心^[13],甚至跨越多个司法管辖范围^[14]。此外,一些DFS(如HDFS)将分割后的块文件以数据块编号命名,仅从文件名无法判断文件的类型和归属关系,为可疑人员的追溯与判定带来了困扰。

3) 不同用户数据存储的混杂性

虚拟性是云计算的显著特点之一。在云环境下,大量用户共享云的基础设施或应用,数据混杂存储。由于数据存储的非连续性和分布性,用户数据的分离困难,收集证据时可能会掺杂无关用户的数据,增大数据提取量和数据分析负担。此外,取证人员必须具有充足证据证明数据的所属关系^[15],同时保护无关用户数据的机密性。

4) 证据易失性

云环境下数字证据的易失性主要体现在删除数据难以恢复。对于虚拟机(VM, virtual machine)实例,当用户释放VM后,VM占用的空间将被CSP收回,此时在众多的云节点中识别、定位删除数据是很困难的^[4,16]。对于其他被删除的用户数据,将随着时间的推移被新的数据覆盖,入侵者可利用这一特性销毁证据。

5) 证据监管链维护困难

证据监管链(chain of custody)在取证过程

中十分重要，它需要清晰地表明证据是如何被收集、分析和保存的，以确保证据在法庭中可以被接受^[17]。由于取证人员对云平台中数据的物理控制能力下降，取证过程对 CSP 有依赖性，但无法证实 CSP 是否提供真实可信的数据，并且云取证过程可能涉及多个司法管辖范围的取证人员，导致证据监管链的维护困难^[18,19]。

6) 时间线重建复杂

时间线的重建是重构完整犯罪过程的关键，但从不同系统中获取的时间戳不一定完全同步。在云环境下，一方面，同一用户（或云服务）的数据可能存储在不同的地理区域，这些区域处于不同的时区^[4,20]；另一方面，VM 时间戳容易受到 VM 运行状态变化等因素的影响，产生时间漂移，增加正确分析时间戳的复杂性。此外，不同文件系统时间戳的更新规则也不完全相同^[21]。

3 云计算取证模型

3.1 术语定义

CSP：云服务提供商，具有对云平台中所有数据的控制权。

用户 U ：给定 CSP 的用户合集。假设用户数量为 n ， $U = \{u_1, u_2, \dots, u_n\}$ ，其中 $u_i (1 < i < n)$ 代表使用云服务的单独个体、组织或公司。

云服务 S ：IaaS、PaaS 和 SaaS 这 3 种模式下的服务合集。设服务数量为 m ， $S = \{s_1, s_2, \dots, s_m\}$ ，其中 $s_j (1 < j < m)$ 代表单个服务。

取证人员 I ：取证过程的执行者。

司法部门 J ：正确使用有关的法律规定，判定证据的有效性，并准确认定案件事实。

损失方代表 LPR：损失方可能是 CSP 本身或用户。损失方的加入能够对取证过程进行监督。由于涉及用户可能较多，选择出若干人员作为监督代表，即 LPR。

物理节点 PN ：物理节点是云平台部署的基本单元，可用于运行多种云服务或存储用户数据^[22]。本模型以内部 IP 标识 PN ，如 PN_{IP-1} 表示内部 IP 为 $IP-1$ 的物理节点。

虚拟机 V ：虚拟机通过 PN 及其在 PN 中的编号进行标识，如 V_{PN-1} 表示某 PN 中编号为 1 的 VM。

数据对象 d ：针对用户 u_i 的个人数据，以

$d_k^{u_i} (1 < k < p)$ 表示该用户的单一数据对象， p 代表数据对象的个数。 D^{u_i} 表示 u_i 的数据对象集合；针对云服务 s_j 的数据，用 $d_k^{s_j} (1 < k < q)$ 表示该服务的单一数据对象， D^{s_j} 表示数据集合。

数据条目 E_d ：在取证人员获取的原始数据中加入部分描述信息形成数据条目。 E_D 表示所有数据条目的集合。

分析结果条目 E_a ：在取证人员分析原始数据产生的输出结果中加入描述信息形成结果条目。 E_A 表示所有结果条目的合集。

取证数据中心 FDC：集中存储收集及分析得出的数据，确保数据的完整性。

证据分析平台 EAP：利用云计算的资源搭建的专门用于证据分析的综合性平台，要求具有良好的可扩展性，能够实现多种数据分析框架和取证工具的部署、开发。

3.2 取证模型

本文提出的云取证模型如图 1 所示。该模型分为 5 个阶段：取证准备、证据识别、证据收集、证据分析和报告。其中，取证准备服务由 CSP 部署，而后 4 个阶段由取证人员完成。本节对各阶段采取的取证策略与方法进行详细说明。此外，本模型假设在取证工作开始时，发生安全事件的云服务是已知的，设为 s_j 。

3.2.1 取证准备

针对云平台数据的易失性，同时预防潜在证据在取证开始前受到攻击人员或恶意内部人员的破坏^[23]，在普通云服务正常工作的同时，本模型提出部署实时的取证准备服务，将对取证具有关键作用、易受破坏的数据持续地存储到取证数据中心，同时确保不产生过多的传输和存储开销。本模型选择如下 2 种数据：云存储元数据和 VM 管理数据。

云存储维护着大量的用户数据，是重要的证据来源。元数据记录了数据对象在云存储中的产生和存储过程，以及访问权限和时间戳等信息，具有还原事件时间线、追溯犯罪嫌疑人的作用^[24]。针对被删除文件，保留其在取证数据中心的元数据，能够作为该文件曾经存在的证据和识别定位的依据。若用户数据与 CSP 没有直接利益关系，潜在证据可能受到恶意 CSP 的删除、篡改或伪造，对元数据的实时预留能有效支持证据存在性和真

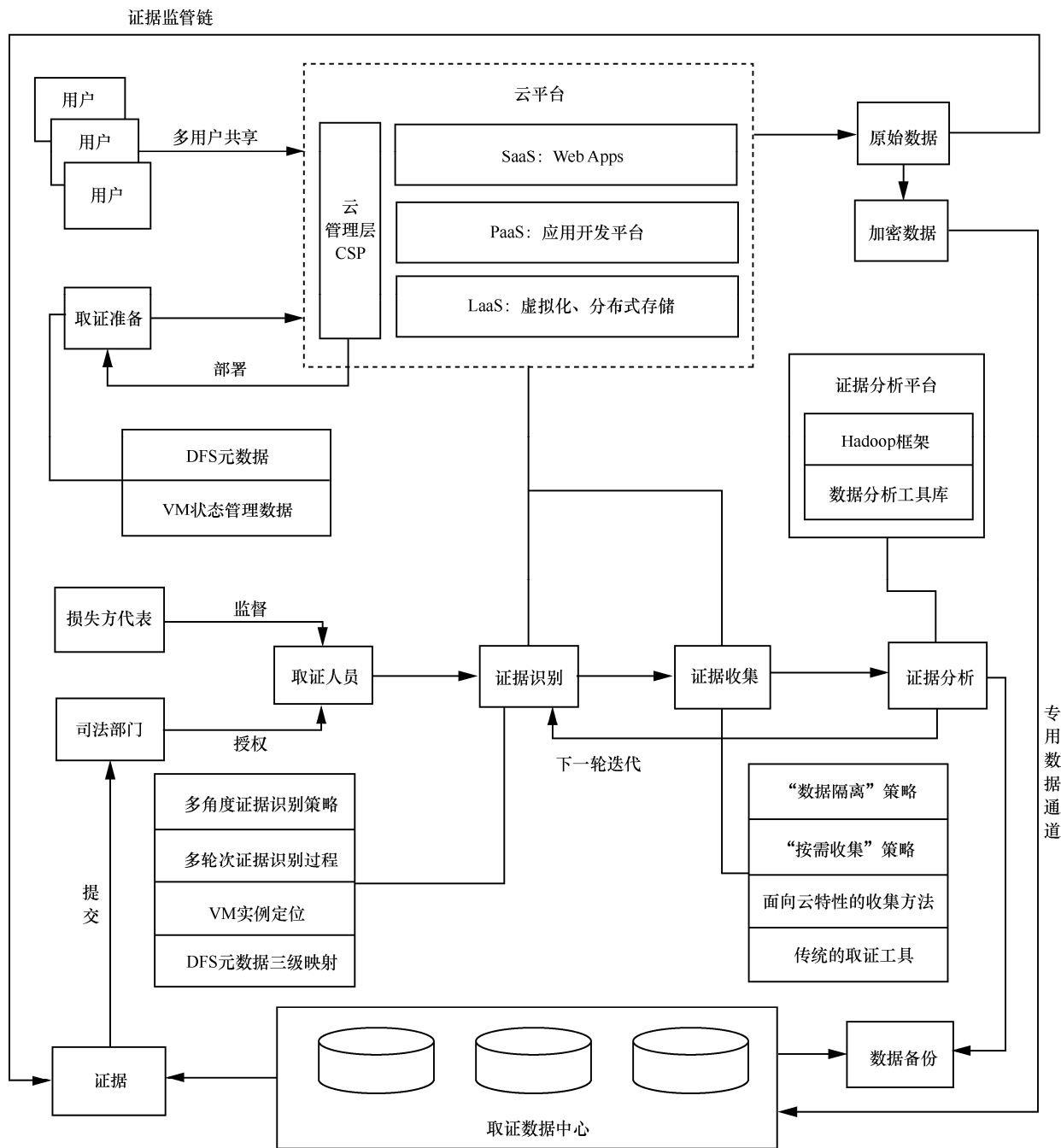


图 1 云取证模型

实性的判定。此外，一些 DFS 采用集中管理的方式存储元数据（如 HDFS、GFS、Lustre 等），相比于实际数据，元数据更易于获取，并且存储开销较小。

VM 在云环境下广泛应用,既是犯罪人员主要的攻击目标,也是其利用云资源实施犯罪行为的主要手段^[23,25,26]。VM 管理数据包括 VM 网络日志、安全日志、运行日志和配置信息等。这些数据可用于重构 VM 的行为、定位 VM 位于的物

理节点，其对于 VM 的取证作用与云存储元数据对于云存储的作用是类似的。VM 管理数据部分存储在 VM 内部，可通过虚拟机管理器（VMM，virtual machine monitor）进行获取。

3.2.2 证据识别

云取证环境的大规模、多层次性和攻击方式的多样性增大了证据识别的复杂性,同时,频繁的数据交互要求证据识别的时效性,以防止证据发生变化。因此,本模型首先提出4条识别策略,

明确不同取证场景下的证据识别范围,提高识别的完备性。针对数据存储的混杂性,对于已识别的数据,提出精确定位其物理位置的方法,确保将数据归属到特定用户。

1) 层次化的证据范围识别

依据 s_j 所属的服务模式及其建立方式确定基本的证据范围。例如,云平台中的 Web 服务可能是由 CSP 提供的 SaaS 服务,也可能是公司基于 PaaS 平台或 IaaS 基础设施建立的服务。不同模式的服务间存在内部关联,不能依据服务模式简单限定证据类型。如果 Web 服务基于 IaaS 搭建,则证据数据表现为整个 VM 实例,而不只是 Web 服务相关的数据。本模型将云平台数据从上至下分为 6 个层次: SaaS 服务的用户数据、SaaS 或 PaaS 服务自身的数据、VM 映像、VMM 数据、物理机数据、基础设施信息。依据犯罪场景的差异,取证人员可对不同层次的数据以及同层次中的不同数据对象进行选择。

2) 分析攻击方式多样性,扩展识别对象

综合考虑多种可能的攻击来源以及云环境下独特的攻击方法,扩展数据识别对象。对于同样的犯罪目的,取证人员一方面需要考虑攻击来自外部人员、恶意内部员工或由用户租用 VM 实施的可能性;另一方面需要考虑攻击者采用常规手段以及利用云平台的自身特性实施犯罪的可能性。例如,一个 VM 实例能够通过边信道攻击从与其处于同一物理机的其他 VM 中窃取信息^[26],该 VM 实例本身也可能受到了入侵。综合 2 个方面,取证人员可将取证对象扩展到云服务的管理员、同一物理节点上的其他 VM、物理机等。

3) 网络通信记录识别

识别与 s_j 所在物理节点或 VM 相关的网络通信记录。网络通信记录是攻击者溯源和事件还原的重要依据,包括云平台内部和外部 2 个方面,主要分为以下 5 类:①服务的访问记录;②不同云服务之间出于管理、协作等目的的通信记录;③单一云服务内部不同节点间的交互记录;④服务所在节点的登录记录,如 SSH 登录;⑤邮件。根据服务的工作特点和异常状态的差异,取证人员可以选择不同的数据。

4) 基于“迭代”思想的多轮次证据识别

由于云服务的大规模和攻击手段的多样性,

一次性识别所有涉案数据是不现实的。本模型采用迭代策略,将本轮次的数据分析结果用作下一轮次证据识别的线索,并权衡证据的识别范围与数量。首轮当中,主要识别 s_j 自身产生的数据,可依据 s_j 的异常状态,从磁盘、内存和网络 3 个方面限定主要的证据类型和识别范畴。分析上述数据,判断可能的攻击类型,发现可疑的用户访问,在第二轮次中,识别 s_j 中可疑用户产生的数据,并以 s_j 为中心,从 s_j 所在节点、云内部和外部节点 3 个方向识别可疑人员与 s_j 有关的数据交互痕迹,如与 s_j 的通信记录、缓存文件等。分析上述数据,在第三轮次中,进一步缩小证据的识别范围,同时扩大范围内数据的收集量,如映像整个磁盘。通过多个轮次的证据识别,取证人员能够不断提升证据识别的精确性和完备性。此外,除首轮外,其他轮次需要补充上一轮次遗漏的数据。

本模型将数据的存储位置划分为 3 个层次:数据中心、物理节点和文件系统数据块。一个 CSP 通常拥有众多数据中心。受到攻击的数据中心是已知的,但识别的相关数据可能存储在其他数据中心。此时,可通过分析网络通信记录在下一轮迭代中对相关数据中心进行定位。

物理节点和底层数据块的识别主要面临云计算的虚拟性和分布性。VM 作为具有操作系统的运行实体,并不建立在 DFS 上,但 DFS 可以基于多个 VM 搭建。对于虚拟磁盘、内存等 VM 数据的获取,通常只涉及对其所在物理节点的定位。基于取证准备阶段存储的 VM 管理数据能够实现;对于 DFS 中的文件,本模型提出一种基于三级映射的文件定位方法,以获取文件本地数据块的地址。本文将 DFS 从上至下划分为 4 个层次:DFS 命名空间、DFS 存储空间、本地文件系统命名空间及其存储空间。本地文件系统指组成 DFS 的各存储节点自身的文件系统。由于 DFS 存储分配单元在本地存储节点中通常以文件的形式存在,DFS 文件到存储节点数据块的映射可划分为 3 个层级:DFS 命名空间与存储单元间的映射、DFS 存储单元与本地存储节点间的映射,以及本地文件系统命名空间与存储单元间的映射。存储节点可能是物理机或 VM,若为前者,则定位过程结束;若为后者,取证人员还需定位 VM 所在的物理节点。

3.2.3 证据收集

由于云服务器时刻保持着大量的用户访问,并且内存等实时数据只能在运行状态下获取,证据收集必须在云平台正常运行的状态下进行。为防止已识别的数据继续变化或被删除、篡改,首先采用“数据隔离”策略,立即对 VM 实例或其他直接存储在物理节点上的数据进行隔离,确保原始数据在授权的证据收集前不再被任何人员访问。隔离工作可利用云计算自身的数据多备份、计算节点迁移等机制,将云服务新产生的数据或无关 VM 实例重新存储到其他节点。不对已识别数据进行迁移,能够防止迁移过程中数据的遗失或损坏。

针对云平台的数据量大,并且不同用户的数据混杂存储的客观情况,本模型提出“按需收集”策略,依据识别的数据物理位置,采用细粒度的数据收集方法,防止提取过多无关数据,在提升取证效率的同时,维护无关用户的数据隐私。此外,云平台在取证过程中新产生的数据可能具有一定的取证价值^[24]。例如,引起安全事件的用户在取证过程中可能仍存在违法行为。因此,取证人员要“按需”收集迁移到其他节点中的数据。

本模型依据云计算的虚拟性将证据收集对象划分为两类:VM 和物理机。它们主要的动态数据是内存,静态数据是文件。文件系统包括单磁盘文件系统(如 NTFS)和 DFS(如 HDFS)。

物理机的内存映像可通过传统的工具进行获取,因为云计算并没有创造新的操作系统,仍然采用 Linux、Windows 等操作系统。但若物理机中运行有 VM,只获取相关 VM 的内存数据更符合“按需”策略。VM 由 VMM 进行管理,其内存映像可通过 VMM 获取。取证人员需要先通过 VM 的挂起机制冻结 VM 系统,然后进行数据收集,以阻止内存数据变化。

对于文件的收集,本模型采取绕过文件系统命名空间,直接提取磁盘底层数据块的方式,防止对文件的时间戳等逻辑特性造成影响。单磁盘文件系统命名空间与存储空间的映射相对简单;DFS 文件可以通过三级映射方法定位它的本地数据块地址。对于 VM 中的文件(包括整个的虚拟磁盘),取证人员可通过 VMM 对数据块进行收集、重组。对于物理机上的文件(包括整个 VM 映像),本模型提出通过一种数据块的只读 API

进行收集。API 的设计需要能直接从不同节点的底层数据块中进行文件的提取和重组。本模型不通过取证工具进行文件收集,一是因为它们 DFS 数据收集中的局限性,二是由于目前没有权威机构对“使用取证工具获取文件是否对文件的逻辑特性造成影响”进行证明,而直接获取整个磁盘映像违背“按需收集”策略。

证据收集过程可能涉及隶属不同机构的多个取证人员,并且对 CSP 具有一定依赖性。因此,数据的收集由 LPR 监督进行,防止恶意 CSP 或取证人员删除证据。数据获取完成后,要及时生成数据条目,记录数据的提取过程,保证数据的可认证性并维护证据监管链的完整。本模型中,假设取证人员 I 在 t 时刻、数据节点 PN_{IP-1} 上获取用户 u_i 的数据 $d_1^{u_i}$, 数据条目 E_d 可表示为

$$E_d = \langle d_1^{u_i}, ID_{u_i}, ID_I, t, PN_{IP-1}, F_{CSP} \rangle \quad (1)$$

其中, ID_{u_i} 表示用户 u_i 的身份标识, ID_I 是 I 在云平台中注册的取证专用 ID, F_{CSP} 是 CSP 是否直接参与证据提取的标记, $F_{CSP} = 1$ 表示是, $F_{CSP} = 0$ 表示否。为防止 $d_1^{u_i}$ 被篡改或伪造,使用 LPR 的私钥 P_{k-LPR} 对 $d_1^{u_i}$ 签名,生成 E_{ed} 。

$$E_{ed} = \langle P_{k-LPR}(d_1^{u_i}), ID_{u_i}, ID_I, t, PN_{IP-1}, F_{CSP} \rangle \quad (2)$$

此处不签名 E_d 的所有元素,以便于取证人员检索数据。最后,将 E_{ed} 通过专用通道传输到证据中心并进行编号,以 $ID_{E_{ed}}$ 表示。若获取服务 s_j 的数据 $d_1^{s_j}$,在 E_d 中, $d_1^{u_i}$ 和 ID_{u_i} 分别用 $d_1^{s_j}$ 和 ID_{s_j} 替换。为强化数据的保护,取证数据中心设立如下存储与访问规则。

- 1) CSP、取证人员和 LPR 都不能删除或修改已存储的数据(除非超过最大存储时限)。
- 2) 普通用户不具有对取证数据中心的任何访问权限。
- 3) 分析结果条目在存储时必须能够表明自身与已存在的数据条目之间的关联。

3.2.4 证据分析

云平台的数据量大,格式多样,并且一些数据格式是云平台独有的。通过少量常规的取证工具或人工分析不能确保数据的有效分析。本模型利用云计算的存储和计算资源,建立证据分析平台,以满足数据分析的存储和效率需求。面向不同的分析对象,分析平台主要包含两大数据处理

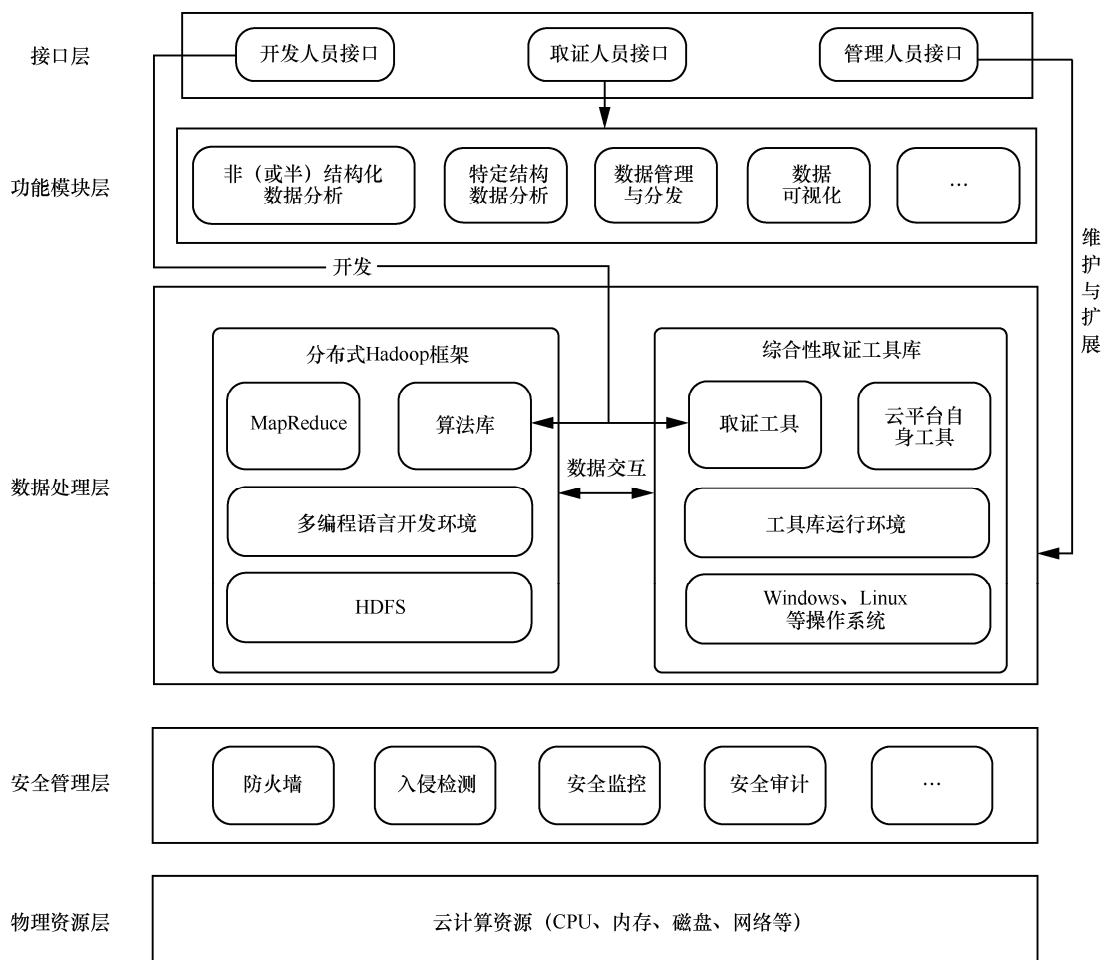


图2 证据分析平台框架

模块：Hadoop 分布式并行处理框架和综合性的数据分析工具库，如图2所示。证据分析平台的数据从取证数据中心获得，二者之间通过专门的安全通道传输数据。

面对巨大的数据量，很多研究机构已经从“分析所有数据以确保不遗漏任何内容”的取证思路转变为更多地依赖于智能方法^[27]。本模型采用Hadoop 和智能数据分析算法结合的方式，分析云服务日志等非结构化或半结构化的数据。Hadoop 的主要组件 MapReduce 是一种数据并行处理的编程模型，具体的数据分析程序由开发人员实现。生成的程序存储在图2中的算法库中，由取证人员调用。基于该模式，取证人员能够深度挖掘云服务内部各组件的行为特点、多种日志中记载的事件之间的关联和顺序、文件系统的活动特征等，同时提高数据的分析效率。

目前，本模型针对文件系统 MAC 时间戳，基于文件系统行为随机模型的思想^[21]，分析特定

文件系统 MAC 时间戳的更新规则，对文件系统活动表现出的特征进行建模，提出基于 MapReduce 对出现上述活动的文件数量及所占比例、异常活动所属的用户（或应用）等信息进行量化分析的方法。由于从不同系统中获取的时间戳可能不完全同步，在分析之前，取证人员首先要建立统一的时间标准，将获取的时间戳转化为标准时间。

针对具有复杂结构的数据，本模型从2个方面建立综合性的取证工具库。一是针对常见的磁盘、内存、网络数据包等数据格式，部署现有取证工具（如 FTK、EnCase、volatility 等）。搭建取证工具需要的运行环境，并统一管理，以便于不同工具间的协作与对比。二是针对云平台独有的数据格式，一方面采用云平台自身的数据处理工具，另一方面开发新的分析工具。利用云平台自身的工具是最高效的方式，但由于环境限制或 CSP 不提供而无法实现，开发对应的数据分析工

具是必要的。工具库可由 CSP 进行管理,对已安装的取证工具进行定期维护与更新,并对新开发的取证工具进行测试与扩展。

分析工具的输出可作为 Hadoop 的输入进行进一步的处理。例如,MAC 时间戳是从文件系统的元数据中提取的,元数据的预处理需要由分析工具完成。Hadoop 的输出也可作为分析工具的数据来源。为便于 Hadoop 框架与工具库的协作,在两者之间设置数据通道,以实现自动化的数据交互。

数据分析的输出结果可能不够直观,本模型采用数据可视化方法,借助于图形化手段,传达数据的关键方面与特征。取证人员通常只使用可视化工具的功能而不负责工具的开发。因此,综合使用开源的数据分析工具与智能化的可视化工具能更好地展示证据。本模型主要利用 Graphviz 实现可视化。Graphviz 是一种开源的基于代码生成图像的可视化工具^[28],能够通过编程批量处理同类数据,清晰展示数据集中的特殊元素,并且能够以子进程调用的方式运行,便于嵌入数据分析工具,取证人员可根据展示需求,将主要精力放在数据的处理上。Graphviz 支持 Windows、Linux、Mac 等多种操作系统,并具有 Java、Python、Perl 等多种语言的调用接口,实用性较强^[29]。

证据分析平台存储着大量的涉案数据,其安全性十分重要。本模型通过防火墙、入侵检测等安全技术防止其受到外部入侵,并通过安全监控与审计监视取证平台的运行状态,确保其正常运行。数据分析完成后,为维护证据监管链的完整,生成结果条目 E_a 。为保护分析结果及其与已存在的数据条目或分析结果条目间关联的完整性,使用 P_{k-LPR} 对 E_a 签名生成 E_{ea} 。

$$E_a = \langle D_a, ID_{E_{ED}}, ID_{E_{EA}}, ID_I, t_s, t_e \rangle \quad (3)$$

$$E_{ea} = \langle P_{k-LPR}(D_a, ID_{E_{ED}}, ID_{E_{EA}}), ID_I, t_s, t_e \rangle \quad (4)$$

D_a 表示该结果条目的具体数据集合,它由取证人员 I 分析数据集 E_{ED} 与 E_{EA} 得出。 t_s 和 t_e 是分析的开始和结束时间。 E_{ea} 通过专用通道存储到证据中心并编号,用 $ID_{E_{ea}}$ 表示。分析得出的数据,能帮助取证人员识别新的潜在证据,开始下一轮的证据识别及后续过程。

3.2.5 报告

在该阶段,证据以报告或证言的形式提交给

司法部门。考虑到司法人员通常只拥有基本的计算机知识^[4,20],取证人员需要以简单明了的方式向司法部门介绍云计算、云取证的概念以及在取证过程中运用的技术手段。

报告应当条理清晰,涵盖取证的所有阶段,详细阐述证据是如何识别、收集、存储和分析的,确保整个取证过程的可重复性,便于司法人员对存在疑问的取证操作在同样条件下进行验证;在整个过程中,证据的完整性没有遭到破坏,证据监管链能够完整地建立;以直观的方式展示证据的分析结果,能够对发生的违法行为进行证明,对犯罪人员进行追责,并重构事件的整个过程。

4 云取证模型评估

结合云环境下的案例场景,本节对所提模型各阶段的取证策略与方法的有效性进行理论分析,并与相关工作进行对比。

4.1 云环境下的案例场景

案例场景如下。假设 CSP A 部署了本模型提出的取证准备服务。公司 B 基于 A 提供的 IaaS 服务合集 S_{I1} , 将其 Web 服务 W 部署在云平台中。 $W = \{w_1, w_2, \dots, w_n\}$ 是一系列服务的集合,产生的数据 D^W 保存在 A 提供的分布式云存储中。恶意用户 u_i 基于 IaaS 服务合集 S_{I2} , 创建了若干台虚拟机 Vs , 利用 B 的服务 w_k 的漏洞,窃取了 D^W 中的数据。攻击完成后, u_i 将数据传回本地终端,并释放了 Vs 。 B 发现数据被盗后,委托取证人员 I 对事件进行调查,并委派损失方代表 LPR 进行监督。 I 在获取司法部门 J 的授权后,在 A 的协助下对云平台进行证据识别与收集,并在专门的证据分析平台下分析数据,最后将取证分析报告提交给 J 。整体的案件场景如图 3 所示。

4.2 模型有效性分析

若采用传统的数字取证模型,由于案件涉及的用户量和数据量大,数据存储具有虚拟性、共享性和分布性,并且资源的回收和再分配频繁,证据的识别和收集工作难以开展。当前的取证工具无法有效分析“大证据”、DFS 或云平台独有的数据格式^[9]。针对云取证面临的主要挑战,本模型的有效性主要体现在以下 5 个方面。

1) 实时取证准备服务的部署

针对数据的易失性,在云平台正常工作的

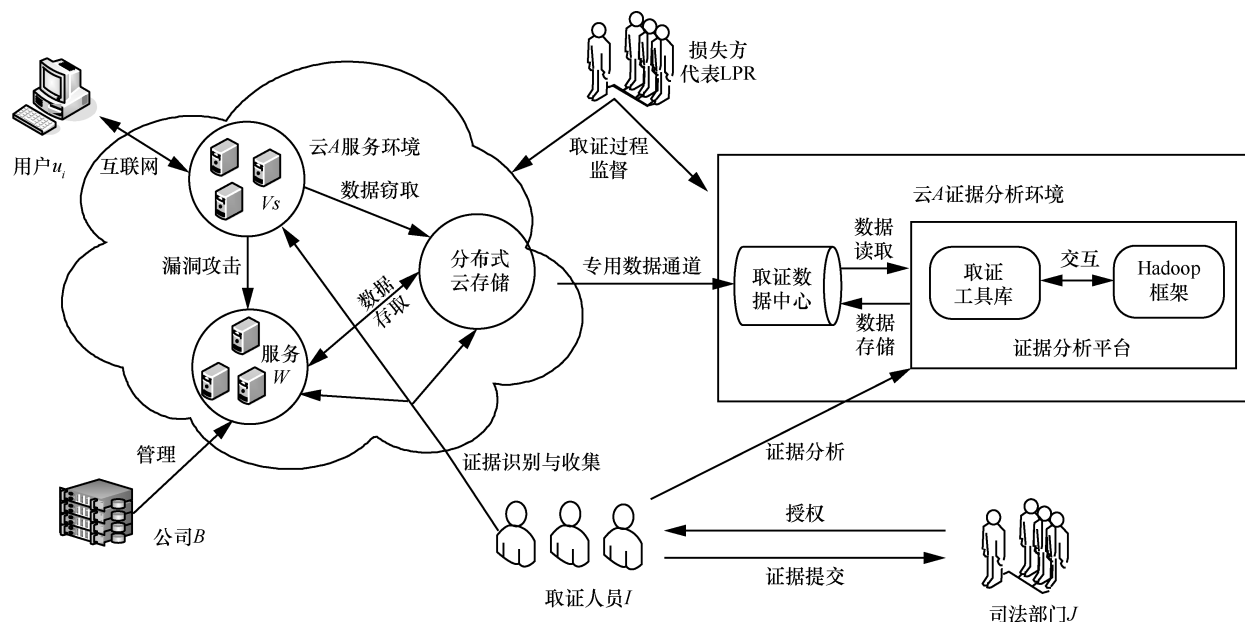


图3 云环境下的案例场景

同时，将云存储元数据和 VM 管理数据实时存储到取证数据中心。即使 u_i 释放了 V_s ， V_s 的使用和通信记录及其所在的物理节点仍然可以识别。云存储元数据提供了文件的时间戳和定位信息等关键数据，并为删除数据的恢复提供了数据基础。

CSP 通常只备份用户或服务数据，而云存储元数据和 VM 管理数据没有进行专门的预留和保护。传统数字取证模型提出的取证准备工作主要是保证取证的人力物力资源，而非关键数据的预留^[30,31]。文献[23]主要针对 VM 分析了数据易失性，但没有提出明确的数据预留对象。

2) “迭代”策略和 DFS 多层次数据定位方法

针对取证环境的多层次性和攻击方式的多样性，本模型依据 W 的部署方式、可能的攻击来源以及 W 一段时间内的网络通信情况，提出了基于“迭代”策略的多轮次识别过程，以最终识别 u_i 租用的 V_s 。文献[24]只对不同服务模式下的数据类型和“迭代”识别思想进行了概念性介绍，没有提出详细的涉案数据识别策略，不足以指导具体的证据识别工作。

针对数据存储的混杂性与分布性，本模型提出了基于元数据层间映射的 DFS 数据定位方法，能够快速、精确定位 W 的数据在 DFS 中的存储节点及数据块地址。传统的数字取证模型多针对单计算机的取证环境^[9]，没有针对 DFS 数据定位

的策略。文献[32]提出通过“数据起源”识别数据的物理存储设备，但没有提出“起源数据”的生成方法。

3) “数据隔离”和“按需收集”策略

针对云环境下用户和取证人员对数据的物理控制权下降的客观状况，首先采用“数据隔离”策略，确保原始数据不再变化或受到删除、篡改等破坏。本模型采取迁移无关数据或 VM 实例的“隔离”方法，文献[33]则对涉案 VM 实例进行迁移，可能造成数据在传输过程中的破坏。

针对多用户数据存储的混杂性，提出“按需收集”的策略，以尽量避免无关数据的收集，并减少对云 A 和服务 W 的影响。针对证据类型及存储方式的多样性，综合运用 VMM、文件数据块只读 API 或传统取证工具对内存或磁盘数据进行细粒度提取，确保不改变数据内容及其逻辑特性。文献[20,24,34]提出的取证模型更侧重于对“数据隔离”等策略以及只读 API 等数据获取方法的必要性与可行性的分析，但没有结合数据存储的虚拟性与分布性提出具体的实施方法。传统的离线证据收集策略不适用于云取证。

4) 基于云资源的综合性证据分析平台的构建

针对涉案数据量大且格式多样的问题，基于云计算的资源构建由 Hadoop 框架和综合性取证工具库组成的证据分析平台，并建立了二者之间的数据交互通道，使 Hadoop 能够对 DFS 等具有

复杂结构的数据进行分割处理。在工具库的建设上,支持云平台自带的数据分析工具的动态扩展,以提高云平台独有数据格式的处理能力。

基于传统的数字取证模型和方法,只能解决单机环境与云环境所共有的证据类型的分析问题,难以对大数据或云平台独有的数据格式进行有效分析。当前运用 Hadoop 进行的文件系统分析多是文件信息统计^[35]或模式匹配^[36]等简单处理,缺乏对文件系统活动进行更深层次的取证分析。

此外,可视化是大数据分析结果展示的要点,本模型采用 Graphviz 等智能化的可视化工具对数据分析结果进行可视化,而传统的数字取证模型或现有的云取证模型一般没有提及分析结果可视化的问题^[23,24,30,31,34]。

5) 证据完整性和证据监管链的维护

在云平台中,数据的存储与传输环境开放,证据的完整性易于受到破坏。本模型在取证过程中的各阶段提出证据完整性的保护措施。

① 证据识别后到证据收集前:识别涉案数据后,立即采取“隔离”策略,防止数据在收集前继续变化或被破坏。

② 证据收集及存储阶段:证据的收集过程由 LPR 监督,获取的数据需由 LPR 签名,防止数据篡改或伪造。此外,取证数据中心禁止对已存储数据的修改。

③ 证据分析及分析结果存储阶段:证据的分析过程由 LPR 监督,分析结果由 LPR 签名,并且必须与已存在的数据条目建立关联,确保数据分析过程可重复验证,防止恶意取证人员伪造分析结果。

④ 取证数据的传输:取证过程中收集的数据和分析结果在网络中传输时,都是经过 LPR 签名的,能够对其完整性进行检验。

在证据监管链的维护方面,假设本案例识别到的用户 u_i 的数据集合为 D^u_i ,取证人员 I 收集数据后,生成数据条目集合 E^u_D 并由 P_{k-LPR} 签名生成 $E^{u_i}_{ED}$,将 $E^{u_i}_{ED}$ 存储在证据中心。分析 $E^{u_i}_{ED}$ 后,生成的结果条目集合 E^u_A 同样用 P_{k-LPR} 签名形成 $E^{u_i}_{EA}$ 后,存储到证据中心。数据是由谁、何时、如何、在什么位置被收集、分析、存储的等,信息通过 $E^{u_i}_{ED}$ 和 $E^{u_i}_{EA}$ 进行了记录。因此,在整个取证过程中,本模型能够维护证据监管链的完整。

5 结束语

云计算的大规模、虚拟性和分布性等固有特性导致传统的数字取证模型和方法在云环境下存在很大的局限性。证据的识别、收集和分析面临巨大挑战,证据监管链难以维护。本文针对云取证面临的主要挑战和当前云取证模型研究的不足,分析云计算特性,提出了一种包含 5 个阶段的云取证模型。该模型综合运用面向云特性的取证策略与方法、传统的取证手段、CSP 的数据控制权限和云计算的资源,以指导云环境下的取证工作。最后,本文结合云环境下的案例场景,分析了模型的有效性。

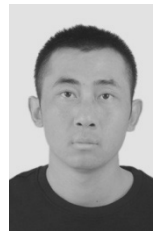
在未来工作中,将联合云服务提供商和司法部门,在真实的云环境下检验本文模型的可行性与有效性,并对模型进一步完善。

参考文献:

- [1] PASQUALE L, HANVEY S, MCGLOIN M, et al. Adaptive evidence collection in the cloud using attack scenarios[J]. Computers & Security, 2016, 59(6): 236-254.
- [2] 林闯, 苏文博, 孟坤, 等. 云计算安全:架构、机制与模型评价[J]. 计算机学报, 2013, 36(9): 1765-1784.
- [3] LIN C, SU W B, MENG K, et al. Cloud computing security: architecture, mechanism, and modeling[J]. Chinese Journal of Computer, 2013, 36(9): 1765-1784.
- [3] SALINAS S, LEE M, COTY S, et al. Cloud security report 2015[R]. Houston: Alter Logic, 2015.
- [4] SIMOUS S, KALLONIATIS C, KAVAKLI E, et al. Cloud forensics: identifying the major issues and challenges[J]. Lecture Notes in Computer Science, 2014, 8484:271-284.
- [5] 中国云计算安全政策与法律蓝皮书(2014)[EB/OL]. <http://www.infsecclaw.net/news/html/841.html>.
- [5] Cloud computing security policies and laws blue book of China (2014)[EB/OL]. <http://www.infsecclaw.net/news/html/841.html>.
- [6] ARCHER J, CULLIANCE D, PUHLMANN N, et al. Security guidance for critical areas of focus in cloud computing v3.0[R]. St Petersburg: Cloud Security Alliance, 2011.
- [7] RUAN K, CARTHY J, KECHADI T, et al. Cloud forensics definitions and critical criteria for cloud forensic capability: an overview of survey results[J]. Digital Investigation, 2013, 10(1): 34-43.
- [8] MELL P, GRANCE T. The NIST definition of cloud computing[R]. Gaithersburg: NIST, 2011.
- [9] 杨泽明, 刘宝旭, 许榕生. 数字取证研究现状与发展态势[J]. 科研信息化技术与应用, 2015, 6(1):3-11.
- [9] YANG Z M, LIU B X, XU R S. Current situation and trend of digital forensics research[J]. e-Science Technology & Application, 2015, 6(1):3-11.
- [10] PICHAN A, LAZARESCU M, SOH S T. Cloud forensics: technical challenges, solutions and comparative analysis[J]. Digital Investi-

- gation, 2015, 13: 38-57.
- [11] QUICK D, CHOO K K R. Impacts of increasing volume of digital forensic data: A survey and future research challenges[J]. Digital Investigation, 2014, 11(4): 273-294.
- [12] MARTINI B, CHOO K K R. Distributed filesystem forensics: XtremFS as a case study[J]. Digital Investigation, 2014, 11(4): 295-313.
- [13] SIBIYA G, VENTER H S, FOGWILL T. Digital forensics in the cloud: the state of the art[C]//2015 IST-Africa Conference. 2015: 1-9.
- [14] ALEX M E, KISHORE R. Forensics framework for cloud computing[J]. Computers & Electrical Engineering, 2017, 60(4): 193-205.
- [15] ZAWOAO S, DUTTA A K, HASAN R. SecLaaS: secure logging-as-a-service for cloud forensics[C]//The 8th ACM SIGSAC Symposium on Information, Computer and Communications Security. 2013: 219-230.
- [16] BIRK D, WEGENER C. Technical issues of forensic investigations in cloud computing environments[C]//The 2011 IEEE 6th International Workshop on Systematic Approaches to Digital Forensic Engineering. 2011: 1-10.
- [17] VACCA J R. Computer forensics: computer crime scene investigation[M]. Hingham: Charles River Media, 2013.
- [18] GRISPOS G, STORER T, GLISSON W. Calm before the storm: the challenges of cloud computing in digital forensics[J]. International Journal of Digital Crime and Forensics, 2012, 4(2): 28-48.
- [19] TAYLOR M, HAGGERTY J, GREYSTY D, et al. Digital evidence in cloud computing systems[J]. Computer Law & Security Report, 2010, 26(3): 304-308.
- [20] ALQAHTANY S, CLARKE N, FURNELL S, et al. A forensic acquisition and analysis system for IaaS[J]. Cluster Computing, 2016, 19(1): 439-453.
- [21] GRIER J. Detecting data theft using stochastic forensics [J]. Digital Investigation, 2011, 8: S71-S77.
- [22] PUTHAL D, SAHOO B P S, MISHRA S, et al. Cloud Computing Features, Issues, and Challenges: A Big Picture[C]//The International Conference on Computational Intelligence & Networks, 2015:116-123.
- [23] ZAWOAO S, HASAN R, SKJELLUM A. OCF: an open cloud forensics model for reliable digital forensics[C]//2015 IEEE 8th International Conference on Cloud Computing (CLOUD). 2015: 437-444.
- [24] MARTINI B, CHOO K K R. An integrated conceptual digital forensic framework for cloud computing[J]. Digital Investigation, 2012, 9(2): 71-80.
- [25] 谢亚龙, 丁丽萍, 林渝淇, 等. ICFF: 一种 IaaS 模式下的云取证框架[J]. 通信学报, 2013, 34(5): 200-206.
XIE Y L, DING L P, LIN Y Q, et al. ICFF: a cloud forensics framework under the IaaS model[J]. Journal on Communications, 2013, 34(5): 200-206.
- [26] RISTENPART T, TROMER E, SHACHAM H, et al. Hey, you, get off of my cloud: exploring information leakage in third-party compute clouds[C]//ACM Conference on Computer and Communications Security. 2009:199-212.
- [27] FAHDI M A, CLARKE N L, LI F, et al. A suspect-oriented intelligent and automated computer forensic analysis[J]. Digital Investigation, 2016, 18:65-76.
- [28] Using graphviz as a library (cgraph version)[EB/OL]. <http://graphviz.org/pdf/libguide.pdf>.
- [29] Documentation[EB/OL].<http://www.graphviz.org/Documentation.php>.
- [30] REITH M, CARR C, GUNSCH G. An examination of digital forensic models[J]. International Journal of Digital Evidence, 2002, 1(3):1-12.
- [31] 丁丽萍, 王永吉. 多维计算机取证模型研究[J]. 信息安全, 2005(11):5-9.
DINH L P, WANG Y J. Multi-dimension forensics model research[J]. NeTINFO Security, 2005(11):5-9.
- [32] TRENWITH P M, VENTER H S. A digital forensic model for providing better data provenance in the cloud[C]//The Information Security for South Africa. 2014: 1-6.
- [33] 周刚. 云计算环境中面向取证的现场迁移技术研究[D]. 武汉: 华中科技大学, 2011.
ZHOU G. Research on forensics-oriented site migration technology in cloud computing environment[D]. Wuhan: Huazhong University of Science and Technology, 2011.
- [34] SIMOUS S, KALLONIATIS C, MOURATIDIS H, et al. Towards the development of a cloud forensics methodology: a conceptual model[J]. Lecture Notes in Business Information Processing, 2015, 215:470-481.
- [35] LIM S H. Implementation of MapReduce model for disk forensic computing analysis[J]. Lecture Notes in Electrical Engineering, 2012, 164:577-584.
- [36] POVAR D, SAIBHARATH, GEETHAKUMARI G. Real-time digital forensic triaging for cloud data analysis using MapReduce on Hadoop framework[J]. International Journal of Electronic Security & Digital Forensics, 2014, 7(2): 119-133.

作者简介:



高元照 (1992-), 男, 河北衡水人, 信息工程大学博士生, 主要研究方向为云计算取证。



李学娟 (1974-), 女, 河南卫辉人, 河南理工大学工程师, 主要研究方向为控制理论与控制工程。

李炳龙 (1974-), 男, 河南卫辉人, 博士, 信息工程大学副教授, 主要研究方向为数字取证。

吴熙曦 (1990-), 女, 四川成都人, 白城兵器试验中心助理工程师, 主要研究方向为网络安全。