INTERNATIONAL
HELLENIC
UNIVERSITY

Big Data on the Cloud. Addressing the Digital Forensic Challenges

Charalampos Stefanidis

UNIVERSITY CENTER OF INTERNATIONAL PROGRAMMES OF STUDIES

SCHOOL OF SCIENCE AND TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in Cybersecurity

May 2021

Thessaloniki – Greece

Student Name:	Charalampos Stefanidis
SID:	3307190017
Supervisor:	Prof. Nikolaos Serketzis

I hereby declare that the work submitted is mine and that where I have made use of another's work, I have attributed the source(s) according to the Regulations set in the Student's Handbook.

May 2021

Thessaloniki - Greece

Abstract

This dissertation was written as part of the MSc in Cybersecurity at the International Hellenic University. A summary of the dissertation follows in the following paragraphs.

The main objective of this dissertation is to identify the emerging challenges in the field of digital forensics when applied to Big Data in a cloud computing environment and propose viable solutions. This task requires extensive knowledge of each field involved and of the existing forensic models. For this reason, this dissertation will focus on presenting each field, analyzing the characteristics and differences of each other. It will also extensively discuss challenges that digital forensic investigators have to face, as well as possible solutions to these challenges. Finally, case scenarios will be provided in order to better showcase the differences between the traditional forensic process and the one where cloud computing and Big Data are involved.

I would like to thank my supervisor Dr. Nikolaos Serketzis for the guidance and support he provided to me during these last months in every step of this dissertation. I would also like to thank my parents and my sister for being by my side during all these years and never stopped supporting me as I try to achieve my goals.

Keywords: Big Data, Cloud Computing, Digital Forensics, Cloud Forensics

Charalampos Stefanidis

4/5/2021

Table of Contents

Table of Figures

1. Introduction

In this day and age, where computer science makes leaps with each passing moment, new fields are explored. This dissertation will feature some of them. More specifically, Big Data and Cloud Computing. In the latest years, these technologies are being utilized by both individuals and organizations to enhance their decision making, productivity, speed and elasticity, while at the same time reducing investment costs. Large organizations, one of which is the European Network and Security Agency, or ENISA for short, have predicted a rapid assimilation of cloud technology and Big Data by not only enterprises, but also from educational and government organizations

While Big Data and cloud computing are beneficial in many ways, specific individuals will attempt to exploit and bend them into their will in order to compromise the security of innocents or delve into other malicious actions. While exploitation has always been an issue with computer science, the rise of malpractice in the recent years has made the need for digital forensics clearer than ever. ENISA, in its report "Threat Landscape 2014" [1], regards these new technologies as highly important in the near future and links this with the shift of cyber criminals towards them. That being said, these tools can be used against them by professionals in the field of cyber security. By identifying and understanding the different challenges present when dealing with cloud computing and Big data, which differ in regards to traditional digital forensics, practitioners aim to gain the upper hand in countering malicious individuals. Once again, ENISA contributes in this constant struggle by enhancing EU's cyber security, providing reports, inside which both challenges and good practices that should be followed are included [2]

Considering the above, the following dissertation aims to explore and present the challenges that are present during a digital investigation in a cloud environment while handling Big Data in the following order:

- *Chapter 1*, Introduction briefly describes the emerging threat of malicious individuals that is looming over new technologies and highlights the significance of adapting the traditional digital forensic methodology to counter them effectively.

- *Chapter 2*, Advances in Computer Science acts as an introductory chapter, presenting the two distinct fields of Cloud Computing and Big Data, along with their characteristics and peculiarities.

- *Chapter 3*, Digital Forensic Science provides information related to the existing investigation framework, as well as differences and challenges an investigator might face when dealing with large volumes of data and a cloud environment respectively.

- *Chapter 4*, Addressing the Challenges of Big Data and Cloud Forensics, analyzes solutions proposed by other researches in order to tackle challenges that have been previously referenced in the third chapter.

- *Chapter 5*, Case Studies, presents two distinct scenarios that aim to provide insight on both challenges and solutions previously analyzed, when applied in a real-life scenario.

- *Chapter 6*, Discussion and Future Work mentions issues the writer dealt with during the dissertation and provides suggestions for future research and improvements.

- *Chapter 7*, Conclusions, acts as a final chapter to this dissertation, recapping the work that has been done during the dissertation and presents the conclusions.

2. Advances in Computer Science

Every day, new leaps are being made in the domain of computer science. In this chapter, the terms of Big Data and Cloud computing will be analyzed to provide insight into their characteristics and specifications, which set them apart from other fields of computer science.

2.1. Big Data

As technology advances, so does the amount of data transferred is increased. This exponential growth in both structured and unstructured data over the years has brought forth a new term, that of Big Data. Big Data can be defined as "a set of data that is so large that it creates difficulty in storing, managing, processing and analyzing them by traditional means" [3]. It is distinguished by six characteristics which are namely volume, velocity, variety, veracity, variability and value.

Volume

As mentioned before, Big Data is characterized by a large amount of data that cannot be contained in a single machine. For this reason, there is a need for specialized tools and frameworks in order to be able to process and analyze that kind of data. As an example, social media platforms these days handle millions, if not billions, of messages each day, which in turn translates to a massive size of data that needs to be processed and stored. Therefore, there is a need for scalable solutions in terms of data storage, such as multiple servers, and a distributed approach in processing the collected data.

Velocity

This characteristic refers to the speed that data is moved around or generated. As mentioned before, applications like social media can generate data in high velocity, which in turn results in accumulated data to have a large volume in a short span of time. Also, some applications that analyze data can have deadlines that are quite strict, such as online fraud detection tools, where there is real time data analysis. For this reason specialized tools need to be deployed in order to ingest that incoming amount of data into the infrastructure and analyze it on the spot.

Variety

Variety refers to the type of the data that is processed. The data could be unstructured, structured or even semi structured, such as text data, images, audio or even data that is collected from sensors. Therefore big data systems require certain flexibility in order to cope with handling each case.

Veracity

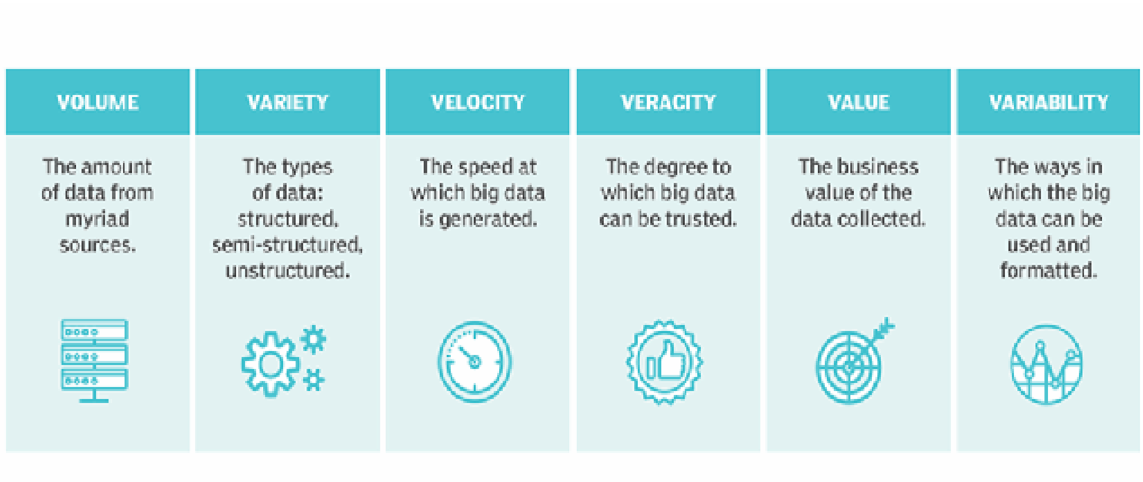
Veracity in Big Data refers on how accurate the collected data is. In order any value to be extracted, the data mass needs to be processed in such a way to remove any kind of noise or unimportant information. Applications that are data driven can only benefit from big data when that data is actually accurate and meaningful. For that, cleansing them so that any incorrect or not useful data are filtered out is quite important.

Variability

This characteristic refers to how inconsistent data can be over time, when it comes to volume, velocity and variety. Data processes can usually generate different data load at different times. Therefore, the system needs to be able to handle such variability, especially when it reaches its peak, in storing and processing the data.

Value

Last but not least, when it comes to the term of value in Big Data, it usually refers to how useful that data is for a specific purpose. Since the end goal of Big Data analytics systems is to provide as much value as possible from the data, the term is closely related to both veracity and the accuracy of the data. There are also applications in which value depends on the speed the data is processed.



VOLUME	VARIETY	VELOCITY	VERACITY	VALUE	VARIABILITY
The amount of data from myriad sources.	The types of data: structured, semi-structured, unstructured.	The speed at which big data is generated.	The degree to which big data can be trusted.	The business value of the data collected.	The ways in which the big data can be used and formatted.

Figure 1, The six attributes of Big Data [4]

2.2. Cloud Computing

The term of cloud computing refers to a model in which multiple clients share and distribute computing resources among them, in order to provide services. A more complete definition has been provided by the National Institute of Science and Technology, or NIST for short, in which "Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" [5].

The institute further analyzes the composition of this model, in which there are five distinct essential characteristics, three service models, and four deployment models, as seen in "Figure 2" below.

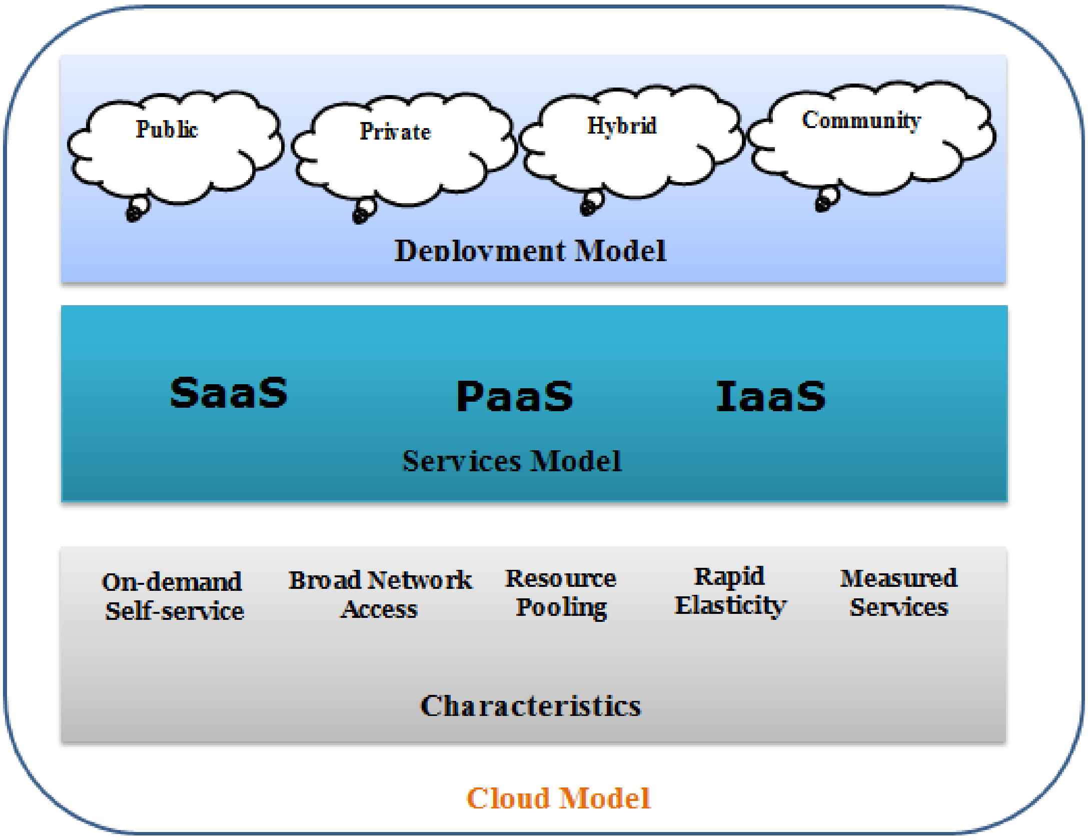


Figure 2, Cloud Computing model [6]

In order to gain a better understanding of the cloud computing model each of the characteristics and parts of it will be displayed and briefly explained.

1) Cloud Computing Characteristics

a) On-demand Self-service: This characteristic is defined by the absence of required human interaction between the provider and the consumer, when the latter requests to make use of the cloud's computing capabilities, such as network storage or server time.

b) Broad Network Access: The network's capabilities can be accessed through various applications on different platforms and devices.

c) Resource Pooling: Based on the National Institute of Science and Technology's definition "the provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand" [7]. This also creates a sense of independence, in which the consumer cannot know or control the exact physical location of the resources he is using at any moment. That being said, it is possible to pinpoint the general location, which for example could be a country or a datacenter.

d) Rapid Elasticity: The provided computing capabilities can be elastically adapted so that they meet the consumer demand in a specific timeframe.

e) Measured Services: As mentioned before, systems utilizing the cloud can automatically adjust and optimize the resource usage depending on the type of service provided, be it storage, bandwidth or processing. This also means there is a layer of transparency for both consumers and providers, since the usage is monitored, controlled and reported at any given time.

2) *Service Models*

As defined by NIST, the cloud model is composed of three service models. These are the Software as a Service model, the Platform as a Service model, and the Infrastructure as a Service model, each with distinct characteristics that differentiate one with the others.

a) Software as a Service: In this service model the provider allows the consumer to make use of the cloud's resources through the applications that are provided. Those applications are usually accessible through client devices and range from a web client, which can be accessed through a web browser, to a full program. It should be noted that in this service model the consumer may only be able to do minor configurations in the application settings, but is generally unable to manage the infrastructure of the cloud. Examples of this model are Dropbox and Salesforce.

b) Platform as a Service: This model differentiates with the one above, as the consumer is able to upload, deploy and control his own applications in the cloud. That being said, the provider still does not allow the consumer to manage the underlying infrastructure of the cloud. Examples of this model are Google App Engine, Windows Azure and Apache Stratos.

c) Infrastructure as a Service: In the Infrastructure as a Service model, the consumer is provided storage, processing and network resources by the other party, in order to be able to deploy and run his own software, which includes but is not limited to operating systems. The difference with the other two service models is that in this case, even though the provider still refuses to give control over the underlying infrastructure, he allows the consumer to manage over the storage, some networking components like a firewall, operating systems and deployed applications.

3) *Deployment Models*

a) Private cloud: In this case, the infrastructure is set to be used by members of a single organization. It can be owned, as well as managed by the organization itself, or from a third party, or even a combination of the two. Its physical layer may or may not exist within the premises of the organization.
b) Public cloud: Being the exact opposite of a private cloud model, the public cloud exists to be used by the general public. The management, operation and ownership of this infrastructure may fall under a government, a business or even an academic organization. Unlike in the previous model, the physical layer of a public cloud model can be found within the organization grounds.
c) Community cloud: Community cloud is similar to the private cloud, with the only difference that its use is not limited by a single organization, but from a community that share the same purpose and concerns, such as a shared mission or policy.
d) Hybrid cloud: As the name implies, this term is used to describe an infrastructure that is a combination of the ones above. More precisely, two or more cloud infrastructures are bound by technology that enables application and data portability, even though they remain unique entities.

3. Digital Forensic Science

As technology advances, it is expected for individuals to try and exploit it for nefarious reasons. Therefore, it is imperative that law enforcement manages to adapt and stop such actions. This chapter will showcase the ways forensic science is applied in a digital environment, including Big Data and cloud computing, as well as the challenges a digital forensic investigator will have to face during his attempt to solve a crime.

3.1. Digital Forensics

The term of digital forensic science, or most commonly known as Digital Forensics, first appeared forty years ago, starting as a minor branch of criminal investigations. Since then, it has grown into a very important part of investigations involving digital means and information. Digital forensics can be explained as the method of discovery and extraction of information related to criminal activity while ensuring that there has been no compromise to its integrity. Furthermore, it has been described by Palmer as "the use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations" [8]. The digital sources are not limited to computers, but could include any kind of digital medium, such as phones, servers, smart watches, digital cameras or any digital storage device such as an external drive.

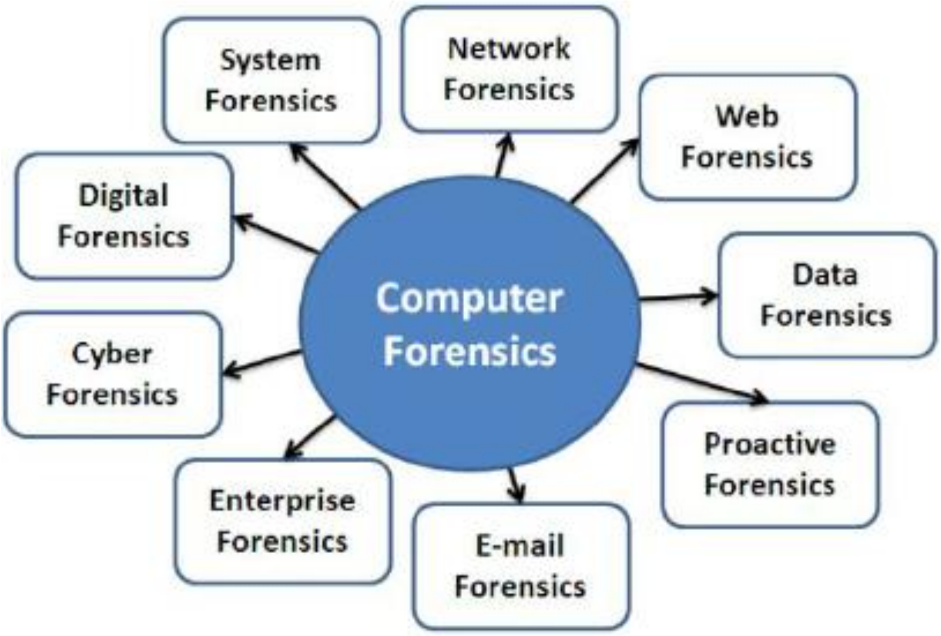


Figure 3, The Branches of Computer Forensics [9]

Over the years there have been many digital investigation frameworks and process models proposed in the academic community. From an overview of these models, Digital Forensics usually consists of the processes mentioned before by Palmer. Namely, these are preparation, preservation, collection, collection, examination, analysis and presentation. These can be seen in 'Figure 4'

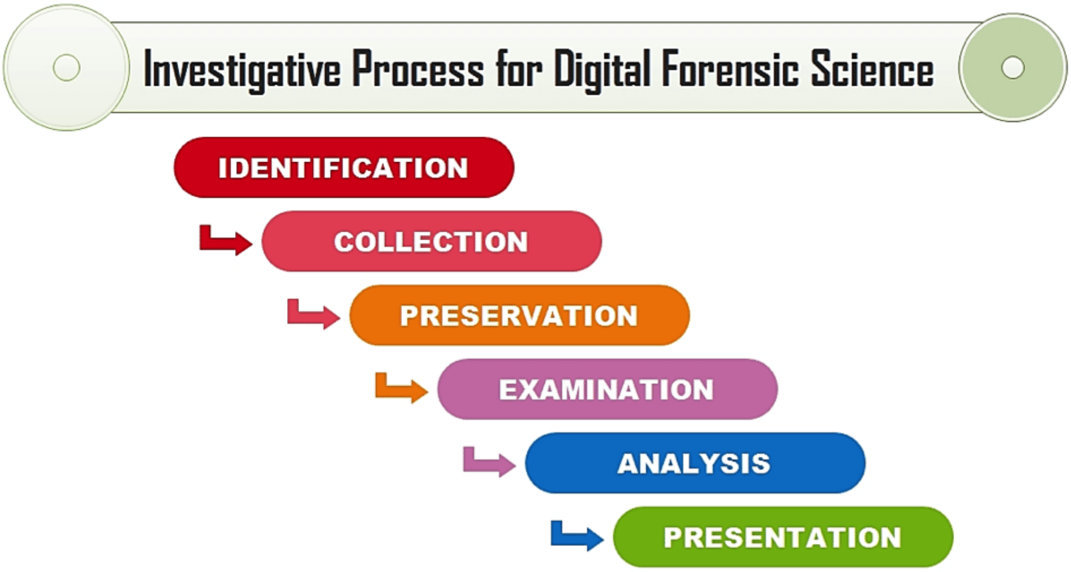


Figure 4, Investigative Process for Digital Forensics [10]

More specifically:

- *Preparation:* This process involves the prerequisite preparation of the required equipment, tools and personnel for an investigation, along with the necessary approval or authorization in order to collect data. This process exists in order to provide assurance that the digital evidence can and will be collected, when it is needed, in a manner that is both efficient and correct.
- *Preservation:* The process of preservation includes the actions of isolating and securing the state of both the physical and digital evidence present at the scene of the crime. This requires that specific actions are taken, which are necessary in order to ensure that the evidence have not been tampered or altered while the investigation is ongoing.
- *Collection:* This stage comes after the identification of the evidence by the investigator. The collection of them, in physical or digital form, in order to

support each investigation case, should be conducted by him, following the proper procedures and techniques.

- *Examination:* During the examination process the evidence that were collected during the previous steps are thoroughly searched in order to identify any information relevant to the ongoing investigation.
- *Analysis*: By analyzing the data that were collected, the investigator is able to determine the significance of them which will lead him or her to conclusions.
- *Presentation:* After the analysis of the data is completed, the investigator is required to summarize and explain properly the aforementioned findings, as well as the conclusions that were drawn. This should be done in such an organized fashion, in order for it to be admissible to the court.

3.2. Big Data Forensics

This section will focus on describing the many challenges that digital forensics investigators have to surpass when dealing with big data. In order to have a clear and complete picture of the challenges, they will be presented in order, depending on the process that they derive from.

3.2.1. Challenges in Big Data Forensics

- Preparation: The main challenge that can be identified in the first stage of a digital forensics investigation is centered on the large volume of data, as well as its variety. Also, the possible diversity of devices in which the evidence can be found has to be considered. Even though it is not hard to set up in place the required procedures, policies and standards that should be followed during an investigation, training the investigator, as well as preparing the correct tools in order to deal with every possible situation can prove to be quite the challenge. As Oluwasola Mary Adedayo mentioned in her paper on Big Data and Digital Forensics [3], “ensuring that an investigator and the tools provided are fully prepared to deal with every situation, application, device, operating system, protocol, file format and encryption, as well as data being on the cloud can be considered an impossible task” [3].

- Collection: The challenge in this stage, as well as the examination and analysis stages, revolve around the growing mass, variety and variability of the data. Even though the cost of storage devices has been reduced through the years, storing large volumes of uncompressed data, the price is still significant. On the other hand, as Darren Quick and Kim-Kwang Raymond Choo have noted [11], if the data is compressed, the cost is reduced, but on the offside, it is not immediately available for analysis.

- Preservation: During the preservation stage, ensuring that the evidence remain intact, meaning that their integrity and authenticity is preserved throughout the course of an investigation, poses a challenge on its own. In addition to that, the large volume of data, along with the variety of devices in which it resides, affects the investigation process, as the time needed for the preservation becomes significantly larger which in turn leads to higher response times. Having appropriate tools to deal with each device may also pose as a separate challenge on its own.

- Examination and Analysis: The very nature of big data, with its attributes, volume, velocity, variety, veracity, variability and value pose a challenging task when trying to examine and analyze it during an investigation. Even though, according to Moore's law, computing power is increasing, the volume of the data is increasing at a much higher rate, leading to backlogs and delays [12]. The traditional methods and techniques that are deployed for digital forensic investigations, such as manual review, string analysis and decision making will not prove sufficient in this case. Therefore, organizing a dataset of this categorization in order to identify clues and facts involving criminal activity can be challenging. It is usual that special data mining techniques and appropriate tools are required in order to yield results [13] [14]. Another issue that can be identified in the examination and analysis stages is that of the false positives, which can lead to considerable longer processing times, as Nicole Beebe has addressed in her research on digital forensics [15]. These issues show the need for new techniques and more sophisticated algorithms to be developed in order to analyze the large volumes of data in a short amount of time.

The wide variety of data structures that can be encountered during an investigation involving big data can also prove challenging during the investigation and analysis stages. This happens because the existing digital forensic tools can handle a limited amount of devices and file formats. This is the reason digital forensic tools should be able to process efficiently both old and current devices and file formats, as well as the ones that will emerge in the future. This is one of the mayor challenges during a digital forensics investigation.

- Presentation: The presentation phase exists in order to gather the accumulated sum of the findings and conclusions that derive from an investigation and present them in an understandable way as evidence to a responsible audience or the court. The challenge in this phase exists in the fact that managing and providing the evidence related to big data is much a much more complex process when compared to traditional computer forensics. This happens not only because of the large volumes of data that are present, but also because the reviewing party, for example the jury, may only have some basic knowledge of computer systems, which in turn means that some technicalities that are present in the evidence, such as identifying a specific value or analyzing big data, can prove to be too complex to understand. Therefore the examiner needs to pay more attention at providing the results of his investigation in an effective, correct and acceptable way.

3.3. Cloud Forensics

Cloud forensics usually refers to anything related to digital forensic investigations involving cloud computing environment. According to the National Institute of Science and Technology, cloud forensic investigations follow the same steps that were mentioned in earlier chapters, them being preparation, collection, preservation, examination, analysis and presentation.

That being said, a forensic investigation on cloud architecture can be considered more complex and difficult than a regular digital forensic one where the investigators proceed with a traditional approach, as mentioned in previous chapters, and have to deal with devices like hard drives, flash memories etc.

Researchers have also defined cloud forensics as "the application of digital forensic science in cloud environments as a subset of network forensics" [7]. Its characteristics can be distinguished in three different aspects. Technical, Organizational and Legal. The first one refers to the hybrid approach digital investigators need to take and tools they need to use within a cloud environment. Examples of such activities are the collection of data, live forensics, segregation of the evidence and proactive measures. The Organizational dimension involves the interactions between the cloud actors, or in other words the different parties that can be associated with cloud computing. For example, cloud providers and cloud consumers, as well as cloud auditors belong in this category. Last but not least, the Legal aspect refers to the regulations and rules that are developed in order to guarantee that the forensic procedures are in line with the law [16].

3.3.1. Challenges in Cloud Forensics

This section will try to present the many challenges cloud forensics investigators have to overcome. In order for this to be accomplished in an organized fashion, they will be categorized based similarly with the process stages that have been mentioned in the earlier chapters. More specifically, the digital forensic model presented by Ameer Pichan, Mihai Lazarescu and Sie Teng Soh in their article on cloud forensics[17], an overview of which can be seen in the figure below.

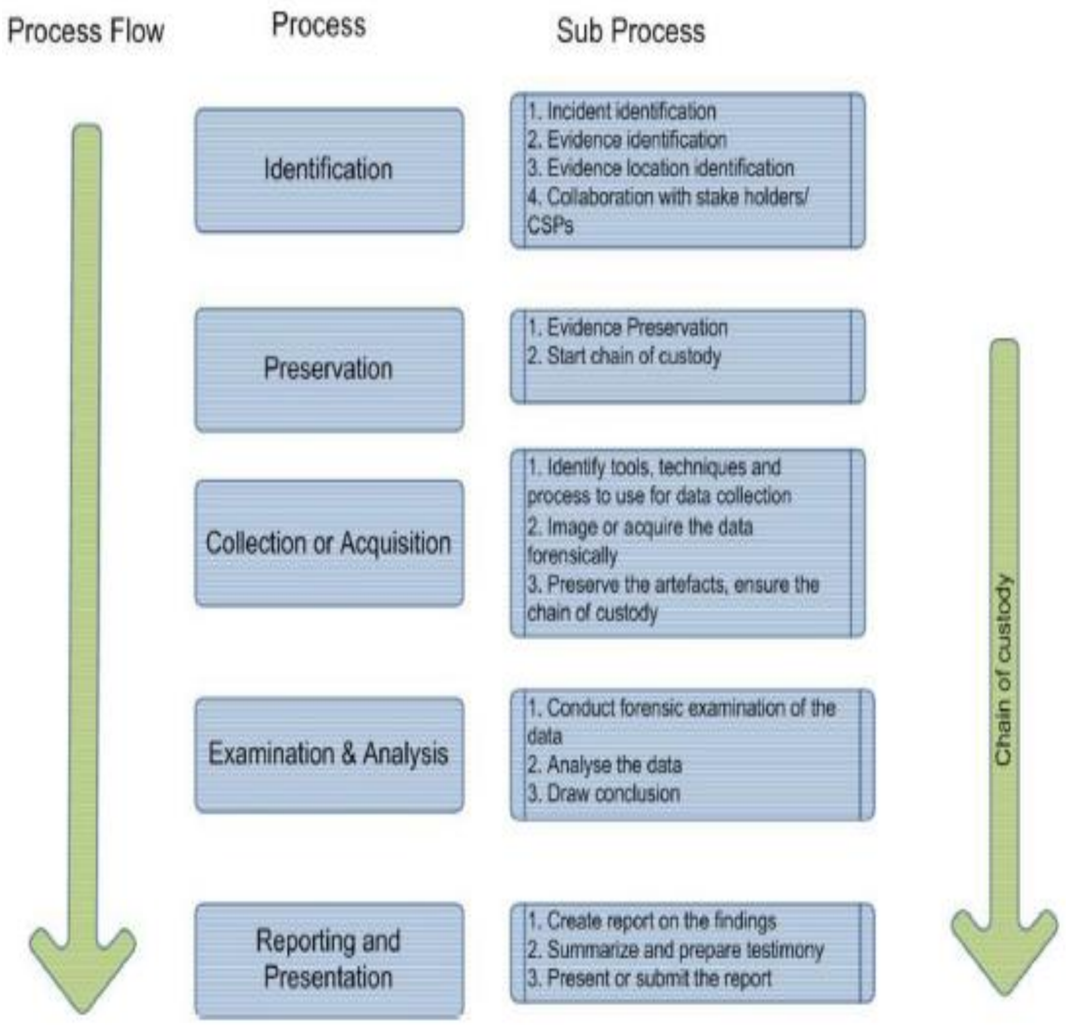

Process Flow
Process
Sub Process
Identification
1. Incident identification
2. Evidence identification
3. Evidence location identification
4. Collaboration with stake holders/ CSPs
Preservation
1. Evidence Preservation
2. Start chain of custody
Collection or Acquisition
1. Identify tools, techniques and process to use for data collection
2. Image or acquire the data forensically
3. Preserve the artefacts, ensure the chain of custody
Examination & Analysis
1. Conduct forensic examination of the data
2. Analyse the data
3. Draw conclusion
Reporting and Presentation
1. Create report on the findings
2. Summarize and prepare testimony
3. Present or submit the report
Chain of custody

Figure 5, Digital Forensics process model [17]

Preparation/Identification stage

- Physical inaccessibility. The most significant difference with the established digital forensic methodologies is that in cloud environments getting physical access to the hardware devices can be considered a difficult to impossible task. That is the case as the data is stored in distributed systems, which can fall under different jurisdictions. It should also be noted that because of this, this challenge exists in all three service models of cloud computing. This issue can also be included in the collection stage.
- Access to logs. Log files are of vital importance during an investigation. Therefore, having access to them is one of the top priorities of a digital investigator in order to identify an incident. Due to the way the system is

distributed in a cloud environment, locating the log files can be a challenging process. The service model also affects the process, since in the cases of PaaS and SaaS it is impossible to check the log files or the system status as the client has limited access. On the other hand, in the Infrastructure as a service model the logs can be accessed since it provides virtual machines which act similar to physical machines [18]. Last but not least, many cloud service providers do not offer logging services or even hide the information from their customers.

- Client Side Identification. It is possible to find evidence from both the providers' side as well as the clients'. For example, the web browser which the client uses to communicate with the provider's services can leave traces for the investigator to follow. For this reason the data from the customer's environment should be included in the investigation, which makes the process even more arduous and time consuming.
- Data duplication. The act of duplicating data is one of the basic techniques employed in cloud computing. It is usually used by cloud providers to ensure that there is some level of fault tolerance, allowing the services to be kept online even if something goes wrong [17]. This feature is also beneficial for forensic reasons, since it is difficult for someone to delete all the evidence. That being said, it also poses a challenge in identifying all the possible evidence since the data is spread in different locations [17].
- Jurisdiction. As mentioned before, data can be spread in different locations. It is usual for the service providers to store the customer’s data in locations that fall outside the latter’s jurisdiction. In addition, it is to be expected that the same laws do not apply in all areas. Therefore, depending on the location of the stored data, the investigators must comply with a different set of rules. In addition to this, service providers often migrate the data from one data center to another, making the challenge that forensic investigators face even more significant [19].
- Dependence on the Cloud Service Provider. In all three service models, the investigators depend on the service providers in order to help locate all the information and evidence that exist within their infrastructure. The challenge exists in cases where the providers do not wish to provide said information,

possibly in order to protect the organization's reputation, and is present not only in the investigation stage, but in the collection and preservation ones as well [20].

- Service Level Agreement. The Service Level Agreement, or SLA for short, is a contract signed by both customer and cloud service provider, containing the terms of their arrangement. Unfortunately, on many occasions, terms related to digital forensic investigations are omitted, which can create challenges for the investigators, since there is no guarantee that the service provider follows the proper methodology in the cloud environment [7].

Preservation

- Data integrity. It is important during the investigation that the integrity of the data is not compromised. If this is not possible, then the evidence will not be accepted by the court of law. Verifying the integrity of the data is also affected by several aspects of the cloud infrastructure, which adds to the challenge. The integrity is usually verified with the use of proven hash techniques such as SHA1, SHA256 and MD5 [20].
- Data volatility. A mayor challenge during evidence preservation and collection is the volatile nature of the data in a cloud environment. Data stored in a virtual machine existing on the cloud will be deleted as soon as the machine is shut down. This means that important evidence such as temporary files, processes and even entries in the registry will not be possible to be recovered. This also means that in the case a perpetrator attacks the virtual machine; he can shut it down in order to delete the volatile data and remove his traces, if there are not active countermeasures [22]. This challenge can be also categorized in the identification stage, since in order to preserve and collect evidence, the investigator first needs to identify them, which will not be possible if they are lost.
- Chain of custody. In order for the evidence to be admissible in the court of law and for their credibility to not be questioned, it is important that the chain of custody is followed through the whole investigation process. This can be

extremely challenging when involving the cloud environment, since in many cases it entices that there are multi-jurisdictional laws that need to be followed. The involvement of the cloud service provider is also a factor since the chain of custody needs to be maintained from his side as well. Therefore, the investigators need to guarantee that the investigation follows the chain of custody by the letter, providing information that draw a clear picture on who had contact with the evidence, how they were stored and handled [23]

Collection

- Multi-Tenancy. It is usual in the Infrastructure as a Service and the Platform as a Service cloud forensic models that many customers share the same storage space inside virtual machines. This can make collecting evidence challenging, since it is important that the privacy of each customer is maintained as well as ensuring that only the data involving a specific customer are gathered. Last but not least, since the storage space is shared, it is possible that the evidence is contaminated by other individuals who have access to it.
- Multi-Jurisdiction. Acquiring evidence data is an issue in cloud forensics as the cloud's resources can be spread in various places. In cases where they reside in locations with different jurisdiction, it is important that the investigators are aware and follow the laws and regulations when retrieving the evidence. Failure to do so can lead to the evidence being invalid in front of a court [24].
- Lack of specialized tools. During an investigation is it important that every bit of data that can be considered to have value as evidence should be collected. This means that artifacts such as metadata, registry entries, file history, network logs and hypervisor logs need to be collected. The challenge exists in the fact that there is an absence of commercial tools that are forensically certified to efficiently collect these in their entirety [17].

Examination & analysis

- Encryption. It is a common method for both cloud service providers and customers to store their data using encryption methods in order to protect them [25]. Even though this is beneficial from the aspect of security, it can also pose a challenge to the investigators since criminals also use the cloud to hide illegal content, such as pornographic images [26]. Examining that kind of data will not be possible if the encryption key is not accessible. Also, the evidence can be compromised by the owner, if he is the only one who can provide the key or even in the case that the key is destroyed.

- Identity. While in traditional digital forensics it is easy to associate a machine and the data stored in it with a person, it becomes a challenge during an investigation on the cloud. The reason for this is that the cloud stores data in multiple locations, in an environment with multiple users and is usually accessed through a client interface. Therefore, distinguishing the identity of the user from a large pool of people is an intricate process [25]. Another issue that has to be considered is that a consumer might claim that his virtual machine has been compromised and is not responsible for any illegal actions.
- Volume of data. The specific challenge has been mentioned in the previous chapter regarding big data challenges. In this case it refers to the large volume of data cloud service providers are store and handle. This makes an investigator's job to locate and gather useful information difficult and time consuming.
- Reconstruction. Crime scene reconstruction can be challenging in cloud environments since the data might be spread through many locations and countries, with different time zones. As a result, placing each fact in the correct order can be troublesome [27]. Another issue that can be identified is that in the case one virtual machine is shut down, the volatile data which could be used as evidence will be lost, making it impossible for the investigator to reconstruct the scene.

Presentation

- Complexity. Similar to the challenge mentioned in the big data presentation stage, due to the limited knowledge people in the court of law might have, the investigator must be able to present the evidence in a cohesive and clear manner. They should be also able to explain the basic terms of cloud forensics and cloud computing.

- Documentation. The investigator should be able to provide proper documentation of the investigation process. By doing so they could ensure that every involved party followed approved methods and maintained the chain of custody of the collected evidence. Therefore they can be admitted to the court of law.

4. Addressing the Challenges of Big Data and Cloud Forensics

Following the identification of the challenges that appear when dealing with Big Data and Cloud forensics, this chapter will focus on presenting and analysing a variety of solutions proposed by members of the academic domain. In order to create a complete and presentable picture of the field, the chapter will be split in two, one addressing challenges that appeared through the Big Data investigation process, while the other will focus on the challenges that appear on each of the three service models present in digital forensics on cloud environment.

4.1. Academic review of Big Data Solutions

Over the years many scholars have researched the field of Big Data forensics, trying to tackle the challenges that appear as the volumes of data keep increasing. Many of them came up with propositions and techniques that can be applied in the different phases of the existing process model that was mentioned in earlier chapters of the dissertation. Therefore, it should be considered wise to present them in an orderly manner, starting from the first phase of the process, which is the preparation stage, and finish with solutions that appeal to challenges met during the presentation stage.

Preparation

During the preparation stage, the biggest challenge that is present is that the examiner needs to be prepared to deal with both the large volume of possible evidence, as well as the variety of different devices, file formats, operating systems and even the law that exists in that particular jurisdiction. In order to overcome this hurdle, an approach was suggested by Marcus K. Rogers, James Goldman, Rick Mislan, Timothy Wedge and Steve Debrota in their Paper on the Computer Forensics process model [28], was that of the use of a matrix which contains all the possible scenarios of the crime scene, the evidence, the suspect and the skills and qualifications required by each member of the investigation team. The reasoning behind this is to identify the different unknowns in the investigation case in order to determine the aim of the investigation. By knowing and understanding the aim and the levels of expertise available at that moment, the lead investigator is able to prepare a plan of attack, determine who else needs to be in the team and what evidence needs to be sought and used in order to complete this investigation successfully.

Collection

The next step of the investigation process that is going to be analyzed is the collection stage. The main challenge that is immediately identified is that of the large volume of data and possible evidence that need to be collected. Various Techniques have been proposed by researchers to tackle this problem, such as sampling [29] [30], triage [28] [29] and selective and intelligent acquisition [21] [31], all of which aim at reducing the data that needs to be collected.

One of the techniques applied in order to manage the huge data volume is that of sampling of the media, so that only data residing in specific sections are gathered. By utilizing statistical techniques, the investigator can gain insight on the contents of the media as well as reduce the amount of data collected, which renders this process and as a result the examination and analysis stages to be completed at a faster rate [29]. There is always a small chance that pieces of evidence may be missed with this technique, but this can be solved by collecting more sections or even increase the size of the blocks that are collected, diminishing the probability of it [30].

Another method that has been proposed is that of triage. In the context of digital forensics, this process can be explained as "a process in which things are ranked in terms of importance or priority" [28]. Evidence, volatile data and other possible sources of evidence are ranked so that tasks are prioritized over others during an investigation. For example, volatile data, which have a short life span, are prioritized during the process.

A third technique that has been suggested is that of intelligent and selective acquisition. This method has been proposed by various academic researchers as a way of overcoming the data volume challenge [21] [31]. It revolves around selecting and collecting data relevant to a specific investigation using logic.

Preservation

The main challenge that can be identified in this phase is the large data volume. While techniques applied in traditional digital forensic frameworks can be used, methods mentioned in the collection stage above can also prove to be useful in this stage as well.

Examination

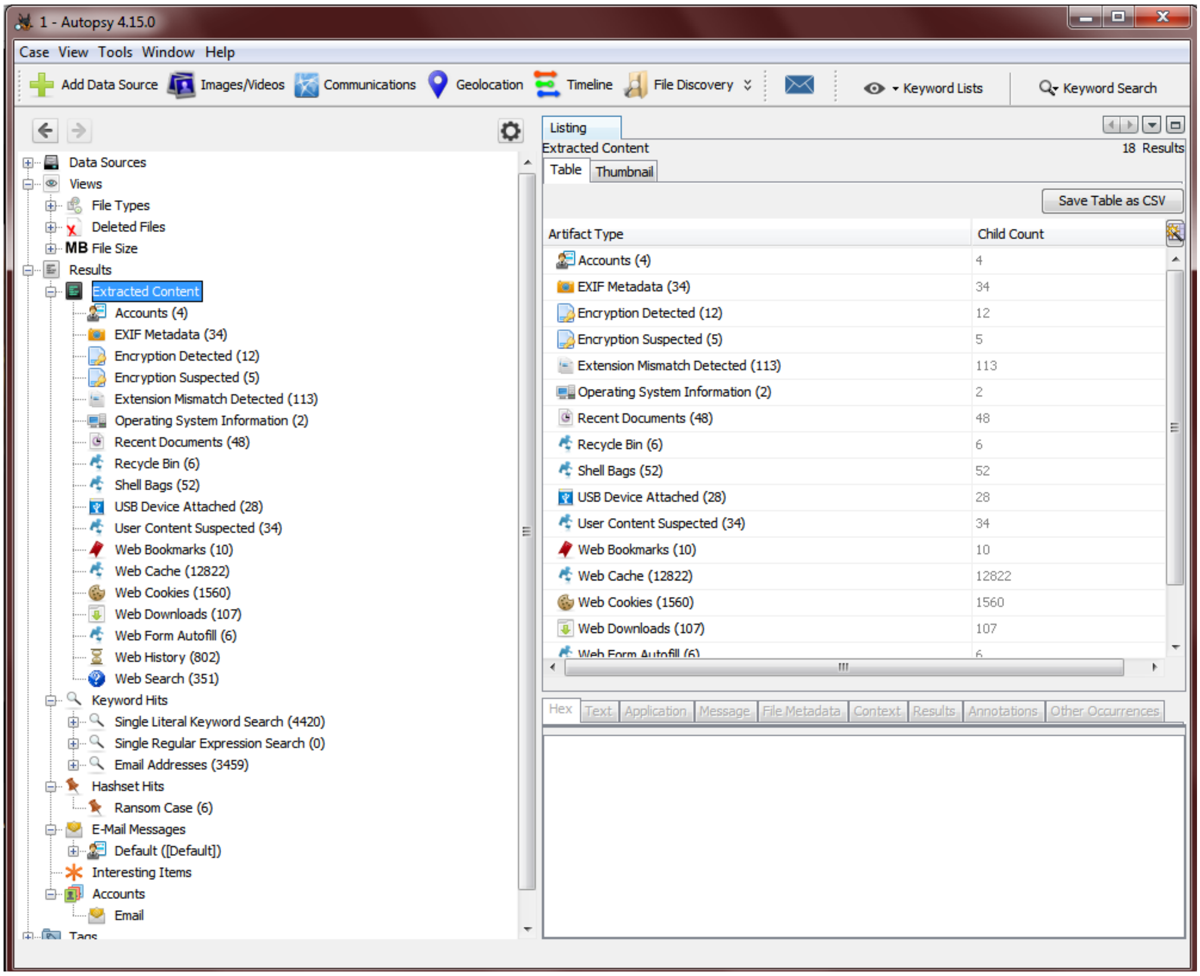
The next step in the process involves the examination of the collected data for useful evidence. The main challenge that has been identified in this stage is that of the time required to inspect the data due to its ever growing size. Therefore intelligent algorithms are required that aim at the reduction of the retrieval overhead in order to achieve faster completion times. Many academics have proposed techniques in respect to this issue, such as the use of cluster analysis [32] and data visualization [29] [32] [33]. Another possible method is through the use of outlier analysis [29] [32] [33] and cross-drive analysis [34].

The further reduction of the number of results that require to be analyzed later can be achieved by utilizing neural networks to cluster the search results [32]. Nicole Lang Beebe and Jan Guynes Clark have proposed an approach based on that concept, which involves using appropriate clustering algorithms on data sets to produce thematically clustered search string results [32].

On the other hand, with the approach of data visualization it Is possible to provide a visual interpretation of the data that will that will aid the investigator while searching for irregularities. Some tools that can provide such results are Forensic Toolkit by Access Data [56] and Autopsy [51], an example of which can be seen in "Figure 6" below.

Another suggested method that could prove useful in the examination process is that of outlier analysis. This technique is explained as an automatic way to examine the collected data in relation to selected evidence [33]. These could be existing evidence, or data that are determined as hidden, suspicious or just unlike other data. The specific technique may also be used to reduce the data that needs to be analyzed.

Finally, one more technique that has been proposed by researchers is cross-drive analysis [34]. This method utilizes techniques in order to correlate data residing in different drives and images. It can be used to further identify crucial information regarding an incident, promote connections between the evidence and even verify the authenticity of found evidence.



Artifact Type	Child Count
Accounts (4)	4
EXIF Metadata (34)	34
Encryption Detected (12)	12
Encryption Suspected (5)	5
Extension Mismatch Detected (113)	113
Operating System Information (2)	2
Recent Documents (48)	48
Recycle Bin (6)	6
Shell Bags (52)	52
USB Device Attached (28)	28
User Content Suspected (34)	34
Web Bookmarks (10)	10
Web Cache (12822)	12822
Web Cookies (1560)	1560
Web Downloads (107)	107
Web Form Autofill (6)	6

Figure 6, Example of Autopsy software.

Analysis

As it has been mentioned in previous chapters, the stage of analysis in digital forensics is comprised of processes and methods that the investigators deploy in order to analyze the evidence, determine the significance of them and reach to a conclusion. It is easy to understand that the specific phase is affected the most by the ever increasing characteristics of Big Data, them being the data volume, variety, variability and the diversity of the devices, in an environment which dictates that the digital forensic investigations are completed in a swift and effective manner. In order to tackle these issues, the academic community has proposed several solutions, such as data mining, distributed computing, the use of artificial intelligence and graphical processing units as well as other techniques that aim at improving the existing processing power.[7][15][17]

The method that utilizes artificial intelligence is commonly referred to by the title of intelligent forensics. Another name that is attributed to this technique is intelligent analysis. As the name implies, it employs artificial intelligence technology, as well as social media analysis and computational modeling in order to reduce the overall analysis time during a digital forensics investigation.[13][23][35]

Another technique that has been suggested by researchers in order to reduce the time required for the analysis of the evidence during a digital forensics investigation is that of data mining. This specific method can be explained as a combination of other distinguishable techniques such as statistical modeling, artificial intelligence, data visualization and other techniques that are utilized with the aim of going through big volumes of data in order to isolate any piece of information that can be deemed important.

Other approaches that have been suggested involve the use of graphical processing units as well as distributed computing in order to further enhance the output of existing digital forensic tools which lowers the required time during the analysis stage even more [20][36]. Even though there has been very limited research related to these methods, their potential is already well known and should be looked on even further.

Presentation

During the last stage of a digital investigation that involves Big Data, the challenges that are distinct revolve around the need to present the evidence in a structured and detailed manner, while they can be understood by the judging party. Therefore, the methods proposed are similar to a traditional digital forensics investigation. A supplementary step that has been suggested in this case is that of providing extensive documentation of the processes used during the investigation in order to further impose a sense of effectiveness and credibility [6].

4.2. Academic review of Cloud Forensic Solutions

As the field of Cloud digital forensics progresses, more challenges are brought to light, many of which have been discussed in previous chapters. In regards to this, many academic researchers have devoted time and resources to develop and propose techniques and methods that could prove useful during an investigation. This chapter will focus on exploring these solutions similarly to the previous chapter involving big data. It is going to differ at some point though due to the three service models that characterize cloud computing, as each one of them brings its own set of challenges to the table.

Identification

In the starting phase of a digital forensics investigation on the cloud, the first challenge that an investigator has to deal with is that of the physical location of the data. As it has been mentioned in previous chapters, the physical location of the virtual instances is usually unknown to the consumer. This creates a difficulty in locating and identifying digital artifacts such as logs and system files during an investigation. Due to the nature of cloud computing, the data may also be inaccessible to the investigator, due to regional and issues, as it may reside outside of his jurisdiction. In his research, Brian Hay [19] suggested that customers use tags on their resources as a mean of indicating where said resources may be located and if they can be transferred in other locations. With this method the Cloud Service Provider could make decisions based on the tags, in favor of minimizing legal complications.

Resource tagging could also be used when dealing with duplicated data, which is a basic characteristic of cloud computing. Even though this service provided by the CSPs is beneficial in terms of forensics, since it is difficult for the data to be completely removed, it still makes identifying the evidence troublesome because of all the different locations. Therefore, resource tagging could be used to follow a file's logical chain and locate it, even if it is deleted [19].

Another issue investigators face is having limited access to log files, information for which are rarely provided by the CSPs, in conjunction with the decentralized nature of data processing in cloud computing, makes it difficult for the investigators to identify and secure the digital artifacts related to the investigation. Researchers have proposed solutions for each service model. For example, Ting Sang [37] has proposed a log model, in which in a Software as a Service environment the client should keep its own log locally and simultaneously, in order to be able to use it and check the activities without the need of the CSP. The model will utilize information such as timestamps, unique identification numbers and hash codes to monitor the log files for changes.

In Platform as a Service model, the service provider could supply the party that employs the platform with a log module in order to create their log module with customized specifications for both client and cloud provider [37].

The last issue that can be identified during the identification phase is that the investigators depend on the CSPs to locate and provide digital evidence from their end. Researchers have suggested that the service providers should start providing forensic services and tools, which should also be included in the SLA [17]. For example, Amazon offers services such as copies of memory dumps and the AWS Cloudtrail logging application [38]. Another suggestion that has been proposed is that of frameworks that promote accountability and build the trust between the customers and the service providers, such as Eucalyptus, with can be applied on the IaaS Cloud service model [39].

Preservation

In the second phase of the investigative process, academics have proposed methods to deal with the challenges, some them being the volatility and integrity of the data in the cloud and the chain of custody.

Having been discussed earlier, data in the cloud is quite volatile, which creates issues during the collection and the preservation of the evidence. Academics who have delved in this territory have proposed the implementation of a persistent storage to try and solve this issue. For example, Birk and Wegener [35] have proposed that a persistent storage should be kept and synchronized with the virtual machines as a mean to reduce data volatility. Even though compromised data cannot be reduced, it is easier to find the digital footprints of the perpetrator through the persistent storage. That being said, CSPs rarely provide such services as it also nullifies the very core of Cloud computing, which is characterized by the low cost and elastic nature.

Through the investigation process, it is vital that the chain of custody is maintained in order for the evidence to be admissible in the court of law. Researchers have proposed solutions to this challenge, such as guidelines for handling digital evidence and keeping an audit trail, provided by the Association of Chief Police Officers of the United Kingdom [40]. Another proposed method is that of utilization of RSA signatures. By following this technique in the cloud, it allows to store a sealed version of the evidence in the cloud, greatly helping the collection and preservation process [41]. It could also be used to verify the integrity of the evidence.

As it has been said before, maintaining the integrity of the data on the cloud is an integral part of the forensic process. The most proposed technique for ensuring this is utilizing known hashing methods such as MD5, SHA1 and SHA256 [17]. Another approach has been proposed by Birg [35], which involves the Trusted Platform Module standard or TPM for short. This module maintains the integrity by allowing a safe and reliable storage while also being able to spot modifications in past configurations.

Collection

In the third phase of the investigation process, researchers have struggled to find ways to counter challenges such as multi tenancy in the cloud, the jurisdiction issues that arise from cloud resources being located in different geographical and jurisdictional areas and the need for specialized tools to assist the investigation process in a cloud environment.

First and foremost, multi tenancy is one of the major points of cloud computing, where a physical machine can host multiple virtual machines, which in turn can be the hosts to numerous tenants. While this is beneficial in terms of cost, it becomes a serious challenge in the forensic field, as it is difficult to collect the whole physical machine without compromising the privacy of the other customers. In order to overcome this problem, researchers have come up with ways that ensure that the CSP and the investigators will be able to collect the related data while at the same time not breaking any regulations and protecting the privacy of the other tenants.

One of the techniques proposed by Dykstra and Sherman [36] is to use the management plane in order to acquire the data. More specifically, they managed to get evidence from Amazon's EC2 Cloud [42] by utilizing known forensic tools such as Forensic Toolkit (FTK) [43]. Other methods that have been proposed revolve around isolating the virtual cloud instance in question by using either methods such as instance and address relocation [44], or even creating a sandboxed version of the virtual machine [44] and continue with the investigation after ensuring the security of the evidence.

Academic researchers have also made propositions to deal with the issues that arise from cloud instances existing in different jurisdictional areas. The most prominent one is that of adding specific clauses in the Service Level Agreement, which would enable the customers to specify the locations their data may be stored or transferred [26].

Finally, the last challenge investigators face during the collection of the evidence in a cloud forensic investigation is the lack of commercially available specialized tools that are able to collect all the possible forensic artifacts in a cloud environment. That being said, researchers such as Dykstra and Sherman have managed to prove that one can acquire data remotely from a user account that is active [36]. They also designed a toolkit on the management plane and actualized it in an Infrastructure as a Service environment, more specifically in a private instance on the OpenStack cloud platform. This specific toolkit is comprised of three tools which are able to accurately collect forensic data such as the API and firewall logs, as well as virtual drives. Another suggestion in regard to this issue was brought on by Federici [45], who presented a software called Cloud Data Imager. This software is able to log communications with the cloud at the application level, as well as acquisition of data remotely from the cloud while retaining the integrity of the digital evidence. This is achieved by applying a read only access to the data [45].

Examination and Analysis

Having collected all the possible evidence, the investigator's work continues to the next stage, that of the examination and analysis of the data. Other challenges are distinguishable there, such as the lack of a standardized log framework, the fact that data can be encrypted, the large volume of data that has to be analyzed as soon as possible and finally the reconstruction of the crime scene by using the evidence.

The lack of a standardized log framework which can be applied in a cloud environment creates challenges in creating a correct time line of events. Once again, Dykstra and Sherman [36] have proposed a framework and offered extensive directions on exactly what, when and where to log in order to assist the investigator's work. Another suggestion that has also been mentioned before is the use of Amazon's CloudTrail service [38].

Another issue that has been discussed in earlier chapters of this dissertation is the fact that encryption can be used for both the benefit of the consumer, but also by the malicious individuals. For this reason, it can be deduced that encryption of data may pose a challenge during an investigation In Cloud Security Alliance's report it is proposed that digital evidence first responders could be able to get access to the decryption key via an option in the key management infrastructure [46]

The challenge of the large volume of data also exists in investigations related to the cloud. Researchers have proposed similar solutions to counter this and expedite the investigation process as the ones analyzed in the chapter regarding big data challenges and solutions. For example, triaging [29] [30] has been suggested as a valid option to reduce the volume that will be analyzed. Another proposition is that the evidence should be put in public cloud storage [12], although this creates new security and legal problems.

Last but not least, reconstructing the crime scene can be highly challenging for investigators. In 2014 the National Institute of Standards and Technology highlighted the need for an appropriate framework containing algorithms, software and guidelines to support the reconstruction process [47]. This draft was finalized and publicized in

2020 [48]. At the moment of writing this dissertation, one can also find by visiting their website that they have the "Guiding Principles for Crime Scene Investigation and Reconstruction (draft OSAC Proposed Standard)." under development [49].

5. Case Studies

Following the previous chapters, where the analysis of the challenges, and possible solutions, took place, this chapter will focus on presenting scenarios of investigations where Big Data in the Cloud are involved. The scenarios will follow the investigation process, highlighting the challenges that exist as well as providing a possible solution. Each case will focus on a different service model of cloud computing, them being SaaS and IaaS, starting with the Software as a Service model.

5.1.Case Study: Software as a Service

In the first fictional scenario, a civilian, who for convenience will be named Alice, suspects that her Dropbox account may be compromised, and someone has been altering her uploaded documents to include pornographic imagery as well as uploading illegal pornographic material in it and most likely sharing it using the link feature that the service provides. Alice has reported it to the authorities, who have charged an investigation team lead by a digital forensics expert named Bob to examine the case, cease the illegal actions and find the perpetrator.

Having been briefed on the situation, Bob begins the investigation by identifying the event's scope and creating a list of what needs to be collected. Realizing that this case is not a traditional digital forensic investigation, where one could need to collect only physical hardware related to the victim, but instead, the crime involves the use of cloud technology, he has to verify that the cloud service provider and the location of the cloud storage fall under his jurisdiction. Should the CSP and the cloud storage be located outside his authority, there will be a need for international cooperation. For the sake of this scenario, both Bob and Alice are located in the United States, where

Dropbox and its servers are located. Therefore, he can continue knowing the possibility of involving extraterritorial jurisdiction is low.

After questioning Alice, she admits to be connecting to Dropbox only via her personal computer's web browser. Therefore, Bob decides that he should get a forensic image of Alice's computer system since it is possible that the intruder first gained access to her Dropbox contents from there. The contents of the Dropbox account should also be considered possible evidence and should be collected. Finally, he requests from the Cloud Service Provider to provide him with the cloud provider access logs, the NetFlow logs and a copy of the virtual machines that the data and duplicates reside. At this point, the Cloud Service Provider suggests that they conduct their own investigation and seem reluctant to provide said data. The provider also points out that giving such information could endanger the privacy of its other customers and refers to the SLA paragraph regarding third-party requests [50] and the absence of a valid search warrant or other legal documents where the end-user, in this case Alice, gives consent.

After getting Alice’s written approval, forensic investigator Bob needs to write in a detailed manner the type of data and information he believes are needed and the reason they are needed for the investigation and requests for a search warrant to be issued. Typically, a search warrant should also include specific information regarding the location of the data that need to be collected [51]. In the case of a cloud environment, though, this can be difficult as the actual physical location is not known and data is usually duplicated. For this reason, the Cloud Service Provider should be handed the warrant and be trusted to act accordingly [51].

Since everything that needs to be gathered has been identified, the collection phase begins. The Cloud Service Provider, Dropbox, after receiving the warrant, instructs an employee to carry it out and securely collect the required data. That means creating a copy of the requested logs as well as isolated instances of the virtual machines containing files related to Alice’s account in order to counter the multi-tenancy issue. The employee also provides the SHA256 hashes of the extracted data in order to verify the integrity. Everything is copied to an external drive and delivered to the investigation team lead by Bob. At this point, it should be noted that the investigation team relies on the competence of the technicians of the cloud provider to identify and gather all the needed information in a correct and complete way. This has been identified as a challenge in cloud forensics in chapter 3.3.1.

Meanwhile, lead investigator Bob tasks another team member to collect an image of Alice's computer system. This can be done by following traditional forensic techniques and utilizing software such as FTK Imager [52], and verifying the integrity by providing the hashes. Last but not least, a copy of all the contents of Alice's Dropbox account are collected in a clean, secure and isolated virtual machine in order to prevent any infection from malicious software and ensure that the data is not tampered or modified during and after the copying process. This can be done by establishing a secure active user account connection via web browser and download its contents. Then use the desktop application to synchronize the local storage, effectively creating another copy of the files. From the beginning of this process, a monitor capturing software should be recording, and a packet capturing tool such as Wireshark [53] should be activated to monitor the activity between the virtual machine and the cloud server. After the process is complete, the virtual machine should be paused, closed and secured. This method has been suggested by Darren Quick and Kim-Kwang Raymond Choo in their article regarding collecting evidence from cloud storage [54].

During every step of the collection process, extensive documentation and proof such as timestamps and hashes, as mentioned before, are kept in order to maintain the chain of custody. By doing so, it will be easier to prove the authenticity of the findings so they can be admitted in the court of law.

Before the examination step is commenced, the investigation team creates a forensic copy of all the evidence, documents it and stores the original evidence in a secure location. In this case, the team creates snapshots of the virtual machines using known virtualization software, which will be the subjects of the examination. The aim of this action is to conduct the examination while safeguarding the integrity of the original collected evidence. After everything is done, the team is ready to begin the examination.

In the examination phase of this digital investigation case, lead investigator Bob has split the team and tasked each member with analysing a different part of the evidence. For example, one member got tasked with performing a forensic investigation on the copied image of Alice's computer, while another member will go through the snapshotted version of the virtual machine that the downloaded contents of the account are stored. Other members have to perform examination upon the data supplied by the cloud provider and go through the digital artifacts that exist there.

The virtual instances can be examined using known forensic software such as Access Data FTK Imager [52], which was already used in a previous phase, and Autopsy [55]. By utilizing these programs' capabilities, the investigators can look for digital remnants present in the snapshots related to the incident, such as cookies, registry keys, deleted files, proof of modified files, and more, along with timestamps that help place each event in a timeline. More precise, searching through the isolated image provided by the Cloud Service Provider, the investigators can reveal information regarding the creation, modification, and the date that the files were accessed, as well as determine which files may also be malicious and the date they were modified or uploaded by the perpetrator. On the other hand, examining Alice's computer system can help the team find evidence that could lead on the way the perpetrator managed to get access to Alice's Dropbox account. These could be tracks left by the criminal, such as malicious software still running in the system, or remnants of his intrusion. Last but not least, the hashes produced by Autopsy and FTK Imager can also be used to confirm that there has been no alteration in the forensic images.

The forensic investigation team also devotes resources and workforce to examining and analyzing the logs provided by the CSP, the logs created from capturing traffic in the collection phase and event logs present in the copy of Alice's system. Each set of logs can provide insight from different angles and bringing the puzzle closer to completion. For example, the investigator examining event logs originating from Alice's system can uncover traces of the perpetrator such as suspicious connections being established, along with their timestamp, that could ultimately lead to tracking the criminal’s IP address. Furthermore, examining the logs from the captured packets during the collection process using Autopsy or Wireshark, the team follows the procedure and highlights any suspicious packet. Finally, logs provided by the cloud provider give information to Bob and the team about login attempts, account access dates and the IPs they were made from, as well as actions made in the account during these sessions.

Having finished examining the data, the team performs analysis methods and cross-references the results from each data source to empower their claims. In traditional forensic cases, this should be enough to recreate the crime scene if the evidence supported it. In this case, though, since it involves cloud forensics, Bob decides to request the investigation results from Dropbox’s end. This way, he can correlate the findings with his team’s and either promote their case or dismiss it and start over. He also requests that they also provide documentation that includes a report of the actions taken by the technician in order to maintain the chain of custody and the integrity of the data.

In the specific scenario, each data source yielded several results. Investigation on Alice’s system and event logs uncovered remote connections between the system and another address, with the first one originating from a third-party web browser plugin, which executed lines of code every time the browser was opened. The code opened a remote backdoor and sent a signal to the other address.

The team member analysing the evidence collected from Dropbox also found malicious code that performed similarly to the plugin inside both the illegal content and Alice’s files. This was discovered after running tests and opening the files while Wireshark was in capture mode. Creation and modification timestamps of the files were also found.

A substantial amount of evidence was also reported by the team members tasked with performing examination and analysis techniques on the data provided by Dropbox’s technical team. More specifically, user account sessions were logged, where the login address was either Alice’s or the one also found in other data sources.

Additional information regarding the user's actions during each session was recorded, which include uploading the illegal content, replacing Alice's files with infected ones and deleting the originals. NetFlow logs also pointed that the infected files were downloaded by multiple addresses, most likely via the creation of a download link service of the Cloud provider.

Lead investigator Bob, having reviewed the findings correlates them with the ones delivered by the CSP, documenting any similarities and differences in the process. In this case, both investigation teams had reached to similar results. They were also able to pinpoint the location the IP address originates from, linking it with a suspect. For the sake of this scenario, the suspect is located in the same State as Alice and Bob.

Having a complete picture of the crime scene based on the evidence, the team then creates an extensive report. It includes a reconstruction of the crime, presenting each event in chronological order along with screenshots from the tool used to enhance the validity. Furthermore, documentation of the procedure standards that were followed and records of each action taken during the investigation process, supported by recording tools, is added to prove that the chain of custody has been maintained, set the evidence as legally valid and admissible to the court of law.

After everything is complete, Bob, after consulting with the legal department, chooses the appropriate court to present the case.

5.2. Case Study: Infrastructure as a Service

The second fictional scenario will follow a digital investigation team based in Greece, trying to solve a crime that takes place on the cloud. More specifically, a suspicious website has come under the radar of the team leader, George. The specific website acts as an online store, offering several illegal services and products such as drugs, stolen goods and even illegal pornography. It also provides its users with the capability to upload video content in exchange for credits, which can then be used in the store. Since conducting this sort of business is prohibited by the law, the team is tasked to provide sufficient evidence in order to identify the shady individuals involved in it, both owners and customers, cease the operation of the online store and delete the illegal media.

By performing a whois record check, the team determines that the website is hosted by Amazon EC2 [42], which provides Infrastructure as a Service model services. As such, the cloud service provider's access and responsibility are limited to only the hardware, storage, network and the servers. For this reason, the investigation team, representing law enforcement, request from the cloud provider to temporarily suspend the operation of the website and preserve the possible evidence as they work to issue a warrant. They also request information regarding the physical location of the hardware related to the illegal webpage so that they can follow the correct legal procedure.

The service provider suspends the website's operation after reviewing it and informs the investigation team that the hardware is located in Frankfurt, Europe, as requested by the customer [56]. Since both the law enforcement agency and the cloud provider are located in the European Union, the investigation team must follow the rules described in The Brussels I Regulation [57] and the 2012 recast version of it [58].

The investigation team must first fully identify the scope of the crime and what needs to be collected as possible evidence. In this case the cooperation of the cloud provider is required since the investigators have limited access to the website files. Therefore, when issuing the warrant, they include documentation instructing in detail the service provider to supply an exact copy of the evidence. These are cloud storage data, the web server's virtual machine, access and NetFlox logs from the provider, and information about the customer along with the payment information he used to rent the virtual space. Since illegal media are involved, the provider complies and instructs a knowledgeable employee to gather the requested data and verify their integrity by utilizing the files' hashes. That being said, the Greek investigators have no way of knowing if the employee collects all the data or if he used the proper tools to perform the procedure correctly. Therefore, they are heavily dependent on the CSP and have to trust him and the employee to deliver accordingly.

While waiting on the provider to complete the requested action, George, the team leader, instructs a team member to install a virtual machine with capturing software, similar to those used in the first case study, and try to collect as much data as possible from the website. This could provide some insight to the team as to whether the perpetrator uses the website to instigate other malicious actions, such as man-in-the-middle attacks. However, there cannot be any guarantee that the information extracted from this can be one hundred percent valid. It can be useful though to prove that the webpage deals in illegal content.

When the provider has completed this task, he sends the data to the team. They also provide a list of what has been delivered, which includes several terabytes of data, account information, access and NetFlow logs and virtual machine snapshots. After verifying the authenticity of the data using the hashes also provided by the employee, the team creates a forensic copy of everything and safely stores the originals as evidence.

Having gathered all the possible evidence, the team leader assigns tasks to each member to execute the examination and analysis stage. The investigator tasked with checking the stored terabytes of data quickly finds out that the perpetrator used encryption techniques. Therefore, the next logical step is to boot up the snapshotted virtual machine, examine the source files, identify the way the encryption works and find the decryption key. After this is done, the stored data may be decrypted and examined.

Since the volume of the stored data is quite large, it can be challenging and time consuming for the investigator team to go through everything. For this reason, the team may utilize methods such as triaging to rank data based on importance and start the analysis process from the top tier ones. Visualizing the data with the use of tools, similar to the first case scenario, is also going to be applied by the investigators to reduce the time needed for the analysis even more. Going through the data, it is easy to prove that illegal content exists. Therefore, the team must focus on finding clues that could lead to the perpetrator and the customers of the website. Examining the file metadata and timestamps is a step in the right direction. NetFlow and access logs are also analyzed since it is most likely that one of the IP addresses found in those is owned by the criminal.

After the analysis stage is complete, the team has managed to find evidence proving the existence of illegal content and over one hundred accounts of users linked to IP addresses, many of which have provided actual credit card information. The possible location of the website owner is also revealed through the access logs, cross-referenced with metadata of uploaded content and the NetFlow records. It should be noted at this point that every action is documented to maintain the chain of custody.

When the documentation is complete and in a presentable manner, the investigator team follows the legal route to prosecute the parties linked to illegal activity and remove any trace of unlawful content from the cloud. Unfortunately, due to the nature of the cloud and the specific case, issues may be highlighted by the defense to belittle the investigation process. For example, can the judging party trust that the data provided by the cloud provider are authentic and their integrity maintained? Are the timestamps and addresses consistent? One other question that could be asked would be which is the governing jurisdiction over the data, based on the location of the data centers and the perpetrator's location. In this case, as mentioned before, the location of the data center is inside the European Union and the address of the criminal also points to the same area. Should he be stationed in another country, such as China, international law would be involved, creating a more significant challenge in prosecuting him.

This second case study can be considered to be more challenging than the first since a larger volume of data needs to be analysed. Also, the investigators have no actual access to the source files and need to rely on the cloud provider to deliver everything correctly. Finally, the team needs to be knowledgeable with regulations and laws regarding not only their regional jurisdiction but also the European Union's and possibly international law.

6. Conclusions

In conclusion, the specific dissertation managed to provide insight into Big Data, cloud computing, and digital forensics. It also attempted to gather and present the many challenges a digital investigator might face in cases where Big Data in a cloud environment is involved. While the challenges met in each field might differ, they are closely related in most cases. It is worth noting that the investigator is confronted with a wide variety of challenges through every stage of the investigation process, starting from the preparation and identification, leading to the presentation of the evidence. Based on the research conducted, the challenges that are considered to be the most crucial and impactful are the large volume of data and the dependence on the Cloud Service Provider. That being said, there are, also, multiple other points of interest that require continuous attention by the digital forensic examiner.

Additionally, by listing and analyzing possible solutions to said challenges, it is the author's belief that investigators and researchers must continuously work towards developing new techniques and perfecting the existing ones, as each day new challenges appear and they will have to adapt in order to be one step ahead of criminal individuals.

Last but not least, the case studies in this dissertation aim to provide some real life scenarios that a digital investigator might come across, presenting the challenges that currently exist in that specific situation, as well as the steps one should take to overcome them and complete their objective. For example, the investigation team in the second case study had to deal with a crime that they had no access to the files, while in the first scenario they had limited access due to acquiring the victim's account. This highlighted the challenge of depending on the cloud service provider even more, and underlined the need for transparency and standardised procedures on the provider's side.

7. Proposals for future research

As mentioned before, this dissertation's aim was to identify the challenges that exist during a digital investigation when big data and cloud computing are involved. While many researchers have proposed solutions to tackle said challenges, there is much room for future work. For example, as technology keeps making strides forward, so should researchers find ways to apply it for the benefit of investigative purposes. Further research in the field of intelligent analysis and the use of artificial intelligence could yield bountiful benefits. Designing new, specialized tools would also help future investigators in their attempts to solve digital crimes. Last but not least, both researchers and legal parties can and should work together to create a legal framework, or update existing ones, in order to build trust between the customer and the provider, as well as propose standard methodologies that providers could follow to oblige to a legal warrant and collect the requested data in ways that could not be disputed by defending legal parties. These are only some of the possible directions research on cloud forensics could move forward since the specific domain is relatively new and has vast potential for improvement.

8. Bibliography

[1] European Network and Security Agency, “ENISA Threat Landscape 2014”. Available at, https://www.enisa.europa.eu/publications/enisa-threat-landscape-2014 [Accessed 01.02.21]

[2] European Network and Security Agency, “Exploring Cloud Incidents”. Available at, **https://www.enisa.europa.eu/publications/exploring-cloud-incidents** [Accessed 01.02.21]

[3] O. M. Adedayo, "Big data and digital forensics", 2016 IEEE International Conference on Cybercrime and Computer Forensic (ICCCF), pp. 1-7, 2016.

[4] B. Botelho, S.J. Bigelow, “Definition: Big data”, 2018. Available at, https://searchdatamanagement.techtarget.com/definition/big-data [Accessed 15.11.2020]

[5] P. Mell, T. Grance, "The NIST Definition of Cloud Computing Recommendations of the National Institute of Standards and Technology," *Nat. Inst. Stand. Technol. Inf. Technol. Lab.*, vol. 145, p. 7, 2011.

[6] S. A. Ali, S. Memon and F. Sahito, "Challenges and Solutions in Cloud Forensics", Proceedings of the 2018 2nd International Conference on Cloud and Big Data, pp. 6-10, 2018.

[7] K. Ruan, J. Carthy, T. Kechadi, M. Crosbie, “Cloud forensics: An overview.”, *Advances in Digital Forensics,* vol. 7, pp. 35–49, 2011.

[8] S. Zawoad and R. Hasan, "Digital Forensics in the Age of Big Data: Challenges, Approaches, and Opportunities", 2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems, New York, NY, pp. 1320-1325, 2015.

[9] N. Sridhar, Dr. D. L. Bhaskari, Dr. P. S. Avadhani, 18 “Plethora of Cyber Forensics.”, *International Journal o f Advanced Computer Science and Applications*, vol. 2, issue 11, pp. 110-114, 2011.

[10] D. Kostadinov, “ The Mobile Forensics Process: Steps & Types”, 2019. Available at, https://resources.infosecinstitute.com/topic/mobile-forensics-process-steps-types/ [Accessed 17.11.2020]

[11] D. Quick and K.-K. R. Choo, "Data reduction and data mining framework for digital forensic evidence: Storage, intelligence, review and archive", *Trends & Issues in Crime and Crimial Justice,* vol. 480, pp. 1-11, 2014.

[12] G. Grispos, T. Storer and W. B. Glisson, "Calm Before the Storm: The Challenges of Cloud Computing in Digital Forensics", *International Journal of Digital Crime and Forensics (IJDCF),* vol. 4, no. 2, pp. 28-48, 2012.

[13] R. Mall, R. Langone, and J. A. Suykens, “Kernel spectral clustering for big data networks”, *Entropy*, vol. 15, no. 5, pp. 1567–1586, 2013.

[14] X. Cai, F. Nie, and H. Huang, “Multi-view k-means clustering on big data,” *23rd international joint conference on Artificial Intelligence*, AAAI Press, pp. 2598–2604, 2013.

[15] N. Beebe, "Digital Forensic Research: the good, the bad and the unaddressed", *Advances in Digital Forensics*, vol. 306, pp. 17-36, 2009.

[16] S. Zawoad and R. Hasan, “Cloud Forensics: A Meta-Study of Challenges, Approaches, and Open Problems.”, *ArXiv abs/1302.6312*, 2013

[17] A. Pichan, M. Lazarescu, S.T. Soh, “Cloud forensics: Technical challenges, solutions and comparative analysis”, *Digital Investigation*, vol. 13, pp. 38-57, 2015.

[18] D. Mohsen, A. Dehghantanha, R. Mahmoud, S.B. Shamsuddin, “Forensics investigation challenges in cloud computing environments*.”, 2012 International Conference on Cyber Security, Cyber Warfare and Digital Forensic (CyberSec),* pp. 190-194, 2012

[19] B. Hay, K. Nance, M. Bishop, “Storm clouds rising: security challenges for IaaS cloud computing”, *Proceedings of the 2011 44th Hawaii International Conference on System Sciences (HICSS)*, pp. 1-7, 2011.

[20] S. Simou, C. Kalloniatis, E. Kavakli, S. Gritzalis, “Cloud Forensics: Identifying the Major Issues and Challenges.”, *Lecture Notes in Computer Science, 2014.*

[21] P. Turner, "Selective and Intelligent Imaging Using Digital Evidence Bags", *Digital Investigation,* vol. 3, pp. 59-64, 2006.

[22] B. Dominik, C. Wegener, “Technical issues of forensic investigations in cloud computing environments”, *2011 IEEE Sixth International Workshop on Systematic Approaches to Digital Forensic Engineering (SADFE)*, pp. 1-10, 2011.

[23] B. Martini, K-K.R. Choo, “An integrated conceptual digital forensic framework for cloud computing”, *Digital Investigation 9*, pp. 71–80, 2012

[24] G. Chen, Y. Du, P. Qin, J. Du, “Suggestions to digital forensics in Cloud computing ERA.”, *2012 3rd IEEE International Conference on Network Infrastructure and Digital Content (IC-NIDC)*, pp. 540-544, 2012

[25] G. Sibiya, H.S. Venter, T. Fogwill, “Digital forensic framework for a cloud environment”, *IST-Africa 2012 Conference Proceedings*, Tanzania, 2012.

[26] S. Biggs and S. Vidalis, "Cloud Computing: The impact on digital forensic investigations," *2009 International Conference for Internet Technology and Secured Transactions, (ICITST)*, London, pp. 1-6, 2009

[27] D. Mohsen, A. Dehghantanha, R. Mahmoud, S.B. Shamsuddin, “Forensics investigation challenges in cloud computing environments.”, 2012 International Conference on Cyber Security, Cyber Warfare and Digital Forensic (CyberSec), pp. 190-194, 2012.

[28] M. K. Rogers, J. Goldman, R. Mislan, T. Wedge and S. Debrota, "Computer Forensics Field Triage Process Model," *Journal of Digital Forensics, Security and Law,* vol. 1, no. 2, pp. 19-38, 2006.

[29] S. L. Garfinkel, "Digital forensics research: The next 10 years," *Digital Investigation,* vol. 7, no. Supplement, pp. S64-S73, 2010.

[30] S. L. Garfinkel, "Digital Forensics Innovation: Searching A Terabyte of Data in 10 minutes," *in DCACM 2013*, Washington DC, 2013.

[31] P. Turner, "Unification of digital evidence from disparate sources (Digital Evidence Bags)", *Digital Investigation,* vol. 2, no. 3, pp. 223-228, 2005.

[32] N. L. Beebe and J. G. Clark, "Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results", *Digital Investigation,* vol. 4, pp. 49-54, 2007.

[33] B. D. Carrier and E. H. Spafford, "Automated Digital Evidence Target Definition Using Outlier Analysis and Existing Evidence," *Proceeding of the Digital Forensics Research Workshop*, New Orleans, 2005.

[34] S. L. Garfinkel, "Forensic feature extraction and cross-drive analysis", *Digital Investigation,* vol. 3, no. Supplement, pp. 71-81, 2006.

[35] D. Birk and C. Wegener, "Technical Issues of Forensic Investigations in Cloud Computing Environments," *2011 Sixth IEEE International Workshop on Systematic Approaches to Digital Forensic Engineering*, pp. 1-10, 2011.

[36] J. Dykstra, A.T. Sherman, “Acquiring forensic evidence from infrastructure as- a-service cloud computing: exploring and evaluating tools, trust, and techniques.”, Digital Investigation, vol. 9, pp. S90-S98, 2012.

[37] T. Sang, "A Log Based Approach to Make Digital Forensics Easier on Cloud Computing", *2013 Third International Conference on Intelligent System Design and Engineering Applications*, pp. 91-94, 2013.

[38] “AWS Cloudtrail - Track user activity and API usage”. Available at, https://aws.amazon.com/cloudtrail/ [Accessed 10.04.21]

[39] D. Nurmi et al., "The Eucalyptus Open-Source Cloud-Computing System," 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid, pp. 124-131, 2009.

[40] ACPO, “ACPO good practice guide for digital evidence (Version 5.0)”, 2012. Available at, http://www.acpo.police.uk/documents/crime/2011/201110-cba-digital-evidence-v5.pdf [Accessed 03.02.21].

[41] C-H. Lin, C-Y. Lee, T-W. Wu, “A cloud-aided RSA signature scheme for sealing and storing the digital evidences in computer forensics.”, International journal of security and its Applications, vol. 6, issue. 2, pp. 241-244, 2012.

[42] “Amazon Elastic Compute Cloud”. Available at, https://docs.aws.amazon.com/ec2/index.html#amazon-ec2 [Accessed 12.04.21]

[43] “Access Data Forensic Toolkit (FTK)”. Available at, https://accessdata.com/products-services/forensic-toolkit-ftk [Accessed 12.04.21]

[44] W. Delport, M. Kohn, and M. S. Olivier, ‘‘Isolating a cloud instance for a digital forensic investigation’’, *Proceedings of the 2011 Information Security South Africa Conference (ISSA)*, pp. 1–7, 2012.

[45] C. Federici, “Cloud data imager: a unified answer to remote acquisition of cloud storage areas.”, *Digital Investigation*, vol. 11, pp 30-42, 2014.

[46] Cloud Security Alliance, “Mapping the forensic standard ISO/IEC 27037 to cloud computing”, 2013. Available at, https://downloads.cloudsecurityalliance.org/initiatives/imf/Mapping-the-Forensic-Standard-ISO-IEC-27037-to-Cloud-Computing.pdf [Accessed 10.04.21].

[47] NIST, “NIST Cloud Computing Forensic Science Challenges (Draft NISTIR 8006)”, 2014. Available at, https://csrc.nist.gov/publications/detail/nistir/8006/archive/2014-06-23 [Accessed 10.04.21]

[48] NIST, “NIST Cloud Computing Forensic Science Challenges”, 2020. Available at, https://csrc.nist.gov/publications/detail/nistir/8006/final [Accessed 10.04.21]

[49] NIST, “Crime Scene Investigation & Reconstruction Subcommittee”, Available at, https://www.nist.gov/osac/crime-scene-investigation-reconstruction-subcommittee [Accessed 10.04.21]

[50] “Dropbox Terms of Service”, Available at, https://www.dropbox.com/terms [Accessed 12.04.21]

[51] J. Dykstra, “Seizing electronic evidence from cloud computing environments.”, *Cybercrime and Cloud Forensics: Applications for Investigation Processes*, IGI Global: Hershey, PA, USA, pp. 156–185, 2013.

[52] “Access Data FTK Imager, Evidence Acquisition Tool”. Available at, https://accessdata.com/products-services/forensic-toolkit-ftk/ftkimager [Accessed 12.04.21]

[53] “Wireshark”. Available at, https://www.wireshark.org/ [Accessed 12.04.21]

[54] D. Quick, K-K. R. Choo, “Forensic collection of cloud storage data: Does the act of collection result in changes to the data or its metadata?”, Digital Investigation, vol. 10, issue 3, pp. 266-277, 2013

[55] “Autopsy Digital Forensics”. Available at, https://www.autopsy.com/ [Accessed 12.04.21]

[56] “Amazon Elastic Compute Cloud, Region and Zones”. Available at, https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-regions-availability-zones.html [Accessed 13.04.21]

[57] EUR- Lex, “Council regulation (EC) No 44/2001. 2000”. Available at, https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32001R0044:en:HTML [Accessed 13.04.21]

[58] EUR- Lex, “Regulation (EU) No 1215/2012”. Available at, https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2012:351:0001:0032:EN:PDF [Accessed 13.04.21]