THE JOURNAL OF DIGITAL FORENSICS, SECURITY AND LAW

Journal of Digital Forensics, Security and Law

Volume 11 | Number 2 Article 9

2016

An Automated Approach for Digital Forensic Analysis of Heterogeneous Big Data

Hussam Mohammed
School of Computing, Electronics and Mathematics, Plymouth

Nathan Clarke
School of Computing, Electronics and Mathematics, Plymouth; Edith Cowan University

Fudong Li
School of Computing, Electronics and Mathematics, Plymouth

Follow this and additional works at: https://commons.erau.edu/jdfsl

 Part of the Computer Engineering Commons, Computer Law Commons, Electrical and Computer Engineering Commons, Forensic Science and Technology Commons, and the Information Security Commons

Recommended Citation

Mohammed, Hussam; Clarke, Nathan; and Li, Fudong (2016) "An Automated Approach for Digital Forensic Analysis of Heterogeneous Big Data," *Journal of Digital Forensics, Security and Law*: Vol. 11 : No. 2 , Article 9.
DOI: https://doi.org/10.15394/jdfsl.2016.1384
Available at: https://commons.erau.edu/jdfsl/vol11/iss2/9

This Article is brought to you for free and open access by the Journals at Scholarly Commons. It has been accepted for inclusion in Journal of Digital Forensics, Security and Law by an authorized administrator of Scholarly Commons. For more information, please contact commons@erau.edu.

(c)ADFSL

AN AUTOMATED APPROACH FOR DIGITAL FORENSIC ANALYSIS OF HETEROGENEOUS BIG DATA

Hussam Mohammed[1], Nathan Clarke[1,2], Fudong Li[1]
[1]School of Computing, Electronics and Mathematics, Plymouth, UK
[2]Security Research Institute, Edith Cowan University, Western Australia
hussam.mohammed@plymouth.ac.uk, n.clarke@plymouth.ac.uk
fudong.li@plymouth.ac.uk

ABSTRACT

The major challenges with big data examination and analysis are volume, complex interdependence across content, and heterogeneity. The examination and analysis phases are considered essential to a digital forensics process. However, traditional techniques for the forensic investigation use one or more forensic tools to examine and analyse each resource. In addition, when multiple resources are included in one case, there is an inability to cross-correlate findings which often leads to inefficiencies in processing and identifying evidence. Furthermore, most current forensics tools cannot cope with large volumes of data. This paper develops a novel framework for digital forensic analysis of heterogeneous big data. The framework mainly focuses upon the investigations of three core issues: data volume, heterogeneous data and the investigators cognitive load in understanding the relationships between artefacts. The proposed approach focuses upon the use of metadata to solve the data volume problem, semantic web ontologies to solve the heterogeneous data sources and artificial intelligence models to support the automated identification and correlation of artefacts to reduce the burden placed upon the investigator to understand the nature and relationship of the artefacts.

Keywords: Big data, Digital forensics, Metadata, Semantic Web

1. INTRODUCTION

Cloud computing and big databases are increasingly used by governments, companies and users for processing and storing huge amounts of information. Generally, big data can be defined with three features or three Vs: Volume, Variety, and Velocity (Li and Lu, 2014). Volume refers to the amount of data, Variety refers to the number of types of data, and Velocity refers to the speed of data processing. Big data usually includes many datasets which are complicated to process using common/standard software tools, particularly within a tolerable period of time. Datasets are continuously increasing in size from a few terabytes to petabytes (Kataria and Mittal, 2014). According to IDC's annual Digital Universe study (2014) the overall created and copied volume of data in the digital world is set to grow 10-fold in the next six years to 2020 from around 4.4 zettabytes to 44 zettabytes. This increasing interest in the use of big data and cloud computing services presents both opportunities for cybercriminals (e.g. exploitation and hacking) and challenges for digital forensic investigations (Cheng et al., 2013).

© 2016 ADFSL

Digital forensics is the science that is concerned with the identification, collection, examination and analysis of data during an investigation (Palmer, 2001). A wide range of tools and techniques both commercially or via open source license agreement (including Encase, AccessData FTK, and Autopsy), have been developed to investigate cybercrimes (Ayers, 2009). Unfortunately, the increasing number of digital crime cases and extremely large datasets (e.g. that are found in big data sources) are difficult to process using existing software solutions, including conventional databases, statistical software, and visualization tools (Shang et al., 2013). The goal of using traditional forensic tools is to preserve, collect, examine and analyse information on a computing device to find potential evidence. In addition, each source of evidence during an investigation is examined by using one or more forensic tools to identify the relevant artefacts, which are then analysed individually. However, it is becoming increasingly common to have cases that contain many sources and those sources are frequently heterogeneous in nature (Raghavan, 2014). For instance, hard disk drives, system and application logs, memory dumps, network packet captures, and databases all might contain evidence that belong to a single case. However, the forensic examination and analysis is further complicated with the big data concept because most existing tools were designed to work with a single or small number of devices and a relatively small volume of data (e.g. a workstation or a smartphone). Indeed, tools are already struggling to deal with individual cybercrime cases that have a large size of evidence (e.g. between 200 Gigabyte and 2 Terabyte of data) (Casey, 2011); and it is common that the volume of data that need to be analysed within the big data environment can range from several terabytes up to a couple of petabytes.

The remainder of the paper is structured as follows: section 2 presents a literature review of the existing research in big data forensics, data clustering, data reduction, heterogeneous resources, and data correlation. Section 3 describes the link between metadata in various resources and digital forensics. Section 4 proposes an automated forensic examiner and analyser framework for big, multi-sourced and heterogeneous data, followed by a comprehensive discussion in section 5. The conclusion and future works are highlighted in section 6.

2. LITERATURE REVIEW

Digital forensic investigations have faced many difficulties to overcome the problems of analysing evidence in large and big datasets. Various solutions and techniques have been suggested for dealing with big data analysis, such as triage, artificial intelligence, data mining, data clustering, and data reduction techniques. Therefore, this section presents a literature review of the existing research in big data forensics and discusses some of the open problems in the domain.

2.1 Big Data Acquisition and Analytics

Xu et al. (2013) proposed a big data acquisition engine that merges a rule engine and a finite state automaton. The rule engine was used to maintain big data collection, determine problems and discover the reason for breakdowns; while the finite state automaton was utilised to describe the state of big data acquisition. In order for the acquisition to be successful, several steps are required. Initially, the rule engine needs to be created, including rules setup and refinement. After that, data acquisition is achieved by integrating the rule engine and two data automation processes (i.e. a device interaction module and an acquisition server). The device interaction module is

 © 2016 ADFSL

employed to connect directly to a device, and the acquisition server is responsible for data collection and transmission. Then, the engine executes predefined rules; and finally, the export process generates results. Generally, this combination gives a flexible way to verify the security and correctness of the acquisition of big data.

In attempting to deal with big data analysis, Noel and Peterson (2014) acknowledged the major challenges involved in finding relevant information for digital forensic investigations, due to an increasing volume of data. They proposed the use of Latent Dirichlet Allocation (LDA) to minimize practitioners' overhead by two steps. First, LDA extracts hidden subjects from documents and provides summary details of contents with a minimum of human intervention. Secondly, it offers the possibility of isolating relevant information and documents via a keyword search. The evaluation of the LDA was carried out by using the Real Data Corpus (RDC); the performance of the LDA was also tested in comparison with current regular expression search techniques in three areas: retrieving information from important documents, arranging and subdividing the retrieved information, and analysing overlapping topics. Their results show that the LDA technique can be used to help filter noise, isolate relevant documents, and produce results with a higher relevance. However, the processing speed of the LDA is extremely slow (i.e. around 8 hours) in comparison with existing regular expression techniques (i.e. approximately one minute). Also, only a selection of keywords that were likely contained within the target document was tested.

In another effort to link deep learning applications and big data analytics, Najafabadi et al. (2015) reported that deep learning algorithms were used to extract high-level abstractions in data. They explained that, due to the nature of big data, deep learning algorithms could be used for analysis and learning from a massive amount of unsupervised data, which helped to solve specific problems in big data analytics. However, deep learning still has problems in learning from streaming data, and has difficulty in dealing with high-dimensional data, and distributed and parallel computing.

2.2 Data Clustering

Recently, data clustering has been studied and used in many areas, especially in data analysis. Regarding big data analysis, several data clustering algorithms and techniques were proposed and they will be discussed as below. Nassif and Hruschka (2011), Gholap and Maral (2013) proposed a forensic analysis approach for computer systems through the application of clustering algorithms to discover useful information in documents. Both approaches consist of two phases: a pre-processing step (which is used for reducing dimensionality) and running clustering algorithms (i.e. K-means, K-medoids, Single Link, Complete Link, and Average Link). Their approaches were evaluated by using five different datasets seized from computers in real-world investigations. According to their results, both the Average Link and Complete Link algorithms gave the best results in determining relevant or irrelevant documents; whilst K-means and K-medoids algorithms presented good results when there is suitable initialization. However, the scalability of clustering algorithms may be an issue because they are based on independent data. From a similar perspective, Beebe and Liu (2014) carried out an examination for clustering digital forensics text string search output. Their study concentrated on realistic data heterogeneity and its size. Four clustering techniques were evaluated, including K-Means, Kohonen Self-Organizing Map (SOM), LDA followed by K-Means, and LDA followed by

© 2016 ADFSL

SOM. Their experiment result shows that LDA followed by K-means obtained the best performance: more than 6,000 relevant search hits were retrieved after reviewing less than 0.5% of the search hit result. In addition, when performed individually, both K-Means and SOM algorithms, gave a poorer performance than when they were combined with LDA. However, this evaluation was achieved with only one synthetic case, which was small in size comparing with real-world cases.

2.3 Data Reduction with Hash-sets

With the aim of dealing an on-growing amount of data in forensic investigations, many researchers attempted to use hash sets and data reduction techniques to solve the data size problem. Roussev and Quates (2012) attempted to use similarity digests as a practical solution for content-based forensic triage as the approach has been widely used in the identification of embedded evidence, identification of artefacts and cross target correlation. Their experiment was applied to the M57 case study, comprising 1.5 Terabyte of raw data, including disk images, RAM snapshots, network captures and USB flash media. They were able to examine and correlate all the components of the dataset in approximately 40 minutes, whereas traditional manual correlation and examination methods may require a day or more to achieve the same result. Ruback et al. (2012) developed a method for determining uninteresting data in a digital investigation by using hash sets within a data-mining application that depends on data collected from a country or geographical region. This method uses three conventional known hash databases for the files filtration. Their experimental results show that a reduction of known files of 30.69% in comparison with a conventional hash-set, although it has approximately 51.83% hash values in comparison with a conventional hash set.

Similarly, Rowe (2014) compared nine automated methods for eliminating uninteresting files during digital forensic investigations. By using a combination of file name, size, path, time of creation, and directory, a total of 8.4 million hash values of uninteresting files were created and the hashes could be used for different cases. By using an 83.8-million-file international corpus, 54.7% of files were eliminated as they were matched with two of nine methods. In addition, false negative and false positive rates of their approach were 0.1% and 19% respectively. In the same context, Dash and Campus (2014) also proposed an approach that uses five methods to eliminate unrelated files for faster processing of large forensics data during the investigation. They tested the approach with different volumes of data collected from various operating systems. After applying the signatures within the National Software Reference Library Reference Data Set (NSRL-RDS) database, an additional 2.37% and 3.4% of unrelated files were eliminated from Windows and Linux operating systems respectively by using their proposed five methods.

2.4 Heterogeneous Data and Resources

The development of information technology and the increasing use of sources that run in different environments have led to difficulties in processing and exchanging data across different platforms. However, a number of researchers have suggested potential solutions for the problem of the heterogeneity of data and resources. Zhenyou et al. (2011) studied the nature of heterogeneous databases and integration between nodes in distributed heterogeneous databases. They suggested the use of Hibernate technology and query

 © 2016 ADFSL

optimization strategy, which have the capability to link between multi-heterogeneous database systems. Furthermore, Liu et al. (2010) proposed a framework based on Middleware technology for integrating heterogeneous data resources that come from various bioinformatics databases. They explained that Middleware is independent software that works with distributed processing, where it is located on different platforms, such as heterogeneous source systems and applications. Their system used XML to solve the heterogeneity of data structure issues that describe the data from different heterogeneous resources while ontology was used to solve the semantic heterogeneity problem. The key benefit of this system is that it provides a unified application for users.

Mezghani et al. (2015) proposed a generic architecture for heterogeneous big data analysis that comes from different wearable devices, based on the Knowledge as Service (KaS) approach. This architecture extended the NIST big data model with a semantic method of generating understanding and valuable information by correlating big heterogeneous medical data (Mezghani et al., 2015). This was achieved by using Wearable Healthcare Ontology which aids the aggregation of heterogeneous data, supports the data sharing, and extracts knowledge for better decision-making. Their approach was presented with a patient-centric prototype in a diabetes scenario, and it demonstrated the ability to handle data heterogeneity. However, the research aim tended to focus on heterogeneity rather than security and privacy through data aggregation and transmission. In the context of heterogeneity Zuech et al. (2015) reviewed the available literature on intrusion detection within big heterogeneous data. Their study sought to address the challenges of heterogeneity within big data and suggested some potential solutions, such as data fusion. Data fusion is a technique of integration of data from different sources that commonly have deferent structures. Most of all they suggested that big heterogeneous data still present many challenges in the form of cyber security threats and that data fusion has not been widely used in cyber security analysis.

2.5 Data Correlation

Although there has already been some work in the data correlation of digital forensics in order to detect the relationship between evidence from multiple sources, there is a need for further research in this direction. Garfinkel (2006) proposed a new approach that uses Forensic Feature Extraction (FFE) and Cross Drive Analysis (CDA) to extract, analyse and correlate data over many disk images. FFE is used to identify and extract certain features from digital media, such as, credit card numbers and email message IDs. CDA is utilised for the analysis and correlation of datasets that span on multiple drives. Their architecture was used to analyse 750 images of devices containing confidential financial records and interesting emails. In comparison, the practical techniques of multi-drives correlation and multi-drives analysis require improvements to their performance in order to be used with large datasets.

Another experiment sought to perform forensics analysis and the correlation of computer systems, Case et al. (2008) presented two contributions to assist the investigator in "connecting the dots." First, they developed a tool called Ramparser, which is used to perform a deep analysis of Linux memory dump. The investigator uses this tool to obtain detailed output about all processes that take place in the memory. Secondly, they proposed a Forensics Automated Correlation Engine (FACE), which is used to discover evidence automatically and to make a

correlation between them. FACE provides automated parsing over five main objects, namely memory image, network traces, disk images, log files, and user accounting and configuration files. FACE was evaluated with a hypothetical scenario, and the application was successful; however, these approaches only work with the small size of data and are not tested on big and heterogeneous data. In addition, the correlation capabilities could leverage existing methods by adding statistical ones.

Raghavan et al. (2009) also proposed a four-layer Forensic Integration Architecture (FIA) to integrate evidence from multiple sources. The first layer (i.e. the evidence storage and access layer) provides a binary abstraction of all data acquired during the investigation; whilst the second layer (i.e. the representation and interpretation layer) has the capability to support various operating systems, system logs and mobile devices. The third layer (i.e. a meta-information layer) provides interface applications to facilitate metadata extraction from files. The fourth layer (i.e. the evidence composition and visualization layer) is responsible for integrating and correlating information from multiple sources, and these combined sources can serve as comprehensive evidentiary information to be presented to a detective. As the FIA architecture was merely conceptualised via a car theft case study, further investigation would be required for the evaluation of its practicality.

2.6 Summary

As demonstrated above, existing studies have attempted to only cope with a specific issue within the big data domain, including volume, complex interdependence across content, and heterogeneity. From the perspective of the volume of big data, the current tools of digital forensics have failed to keep pace with the increase. For that reason, a number of technologies, such as data clustering and data reduction have the potential capacity to save digital investigators time and effort, were examined. Data clustering techniques have been widely used to eliminate uninteresting files and thus speed up the investigation process by determining relevant information more quickly. So far, these techniques can be applied to large volumes of data (in comparison with traditional forensic images) but are not suitable for big data.

Regarding heterogeneity, only a few studies are available; and they were mainly focused on the integration of heterogeneous databases of much smaller sizes (e.g. Hard disk, mobile, or memory dump). Integration technology based on ontology techniques offer promising prospects although it has not been tested within forensic investigations involving big data heterogeneity. Similarly, only limited research on data correlation were conducted despite the data correlation offers a potential solution to heterogeneous data issues; and these issues have yet to be resolved, particularly those related to big data. As a result, big data analytics in the context of forensics stands in need of a comprehensive framework that can handle existing issues such as volume, variety and heterogeneity of data.

3. METADATA AND DIGITAL FORENSICS

Metadata describes the attributes of any files or applications in most digital resources (Guptill, 1999). It provides accuracy, logical consistency, and coherence of files or applications that they describe. Semantic search by metadata is one of the important functions to reduce the noise during information searching (Raghavan and Raghavan, 2014). A number of metadata types exist and provide some attributes which are important in processes as shown in table 1.

 © 2016 ADFSL

These attributes belong to different types of metadata, such as from file systems, event logs, applications, and documents. For instance, file system metadata provides file summary information that describes the layout and attributes of the regular files and directories, aiding to control and retrieve that file (Buchholz and Spafford, 2004); event log metadata provides significant information that can be used for event reconstructions (Vaarandi, 2005).

Document type definition is introduced as email metadata in Extensible Markup Language (XML) which holds content-feature keywords about an email (Sharma et al., 2008). A number of research studies employ email metadata in order to facilitate dealing with email list, such as filtration, organization, and sorting based upon reading status and senders (Fisher et al., 2007). As a result, some research considers metadata as an evidentiary basis for the forensic investigation process because it describes either physical or electronic resources (Khan, 2008; Raghavan and Raghavan, 2014).

Metadata aids to identify the association artefacts that can be used to investigate and verify fraud, abuse, and many other types of cybercrimes. Indeed, Raghavan and Raghavan (2014) proposed a method to identify the relations of evidence artefacts in a digital investigation by using metadata; their method was applied to find the association of metadata from collections of image files and word processing documents.

Rowe and Garfinkel (2012) developed a tool (i.e. Dirim) to automatically determine anomalous or suspicious files in a large corpus by analysing the directory metadata of files (e.g., the filename, extensions, paths and size) via a comparison of predefined semantic groups and comparison between file clusters. Their experiment was conducted on a corpus consisting of 1,467 drive images with 8,673,012 files. The Dirim approach found 6,983 suspicious files based on their extensions and 3,962 suspicious files according to their paths. However, the main challenge with this approach is its ability to find hidden data. Also, it is not effective at detecting the use of anti-forensics.

Therefore; these studies illustrate that metadata parameters can be used by forensics tools for investigation purposes.

© 2016 ADFSL

Table 1
Some of Input Metadata Parameters for Forensics Tools

Source of Input	S/No	Input Parameter	Data Type
Log File Entries (Security Logs)	1	Event ID	Integer
	2	User Name	String
	3	Date Generated	Date
	4	Time Generated	Time
	5	Machine (Computer Name)	String
File System Metadata Structures (NTFS)	6	Modification Time	Date & Time
	7	Access Time	Date & Time
	8	Creation Time	Date & Time
	9	File Size	Integer
	10	Directory flag (File/Folder)	Boolean
	11	Filename	String
	12	File Type	String
	13	Path	String
	14	File Status (Active, Hidden etc.)	Enumeration
	15	File Links	Integer
Registry Information	16	Key Name	String
	17	Key Path	String
	18	Key Type	String
	19	Key Data / Value	String or integer
Application Logs	20	Name	string
	21	Version no.	Integer
	22	Timestamp	Date & Time
Network packet	23	Packet length	Integer
	24	Class	String
	25	Priority	String
	26	Source IP	Integer
	27	Destination IP	Integer

4. SEMANTIC WEB-BASED FRAMEWORK FOR METADATA FORENSIC EXAMINATION AND ANALYSIS

This proposed system seeks to provide an automated forensic examination and analysis framework for big, multi-sourced heterogeneous data. An overview of the proposed framework is illustrated in Figure 1, with three layers of data acquisition, data examination, and data analysis.

4.1 Data Acquisition and Preservation

This layer is used to image all available suspect resources within the same case. It includes creating a bit-by-bit perfect copy for some of the digital media resource (e.g. hard

 © 2016 ADFSL

disk and mobile) and stores images in a secure storage. The preservation process is used to ensure that original artefacts will be preserved in a reliable, complete, accurate, and verifiable way. This preservation is achieved by using hash functions that can be used to verify the integrity of all evidence.

4.2 Data Examination

The examination phase is the core of proposal system. In this layer, a number of techniques are employed to achieve the examination goal, including data filtering and reduction, metadata extraction, the creation of XML files to represent the metadata files, and semantic web technology to work with heterogeneous data. Details of these techniques are presented below:

4.2.1 Data Reduction

The data reduction step has been used widely with a variety of digital forensic approaches and has provided for a significant reduction in data storage and archive requirements (Quick and Choo, 2014). In addition, most the forensic tools and investigating agencies use the hashing library to compare the files which are examined in the suspect cases against the known files, separating relevant files from benign files; as a result, the uninteresting files will not be examined and investigator's time and effort will be saved.

4.2.2 Metadata Extraction

The functionality of this step provides a metadata extractor according to the source type, including hard disk, logs file, network packet, and email. In order to extract metadata, the metadata extractor determines the suitable metadata which should be extracted. For example, it extracts information that holds usual meaning from hard disk files, attributes of logs file and network packets can be used as metadata to reconstruct the events. The output is structured information which makes data in the digital resources easier for retrieving.

4.2.3 XML File Creation

XML is a method to describe structure information which makes them more readable. After metadata are extracted, the XML file will be utilised to create metadata for each file. As a result, the output of this step is a metadata forensic image based on an XML file. The use of metadata helps to solve the big data issues as the size of XML file, which represents metadata of the original file, is much smaller than original files.

4.2.4 Metadata Repository

The repository has the capability to support data from multiple sources of digital evidence that related to one case (metadata forensics images from various resources). It will be used as a warehouse for the next step that feeds the semantic web technology with data.

4.2.5 Ontology-Based Forensic Examination

The major aim of this step is to integrate heterogeneous resources that are located in the metadata repository by using techniques based upon semantic web technology. The semantic web enables a machine to extract the meaning of web contents and retrieve meaningful results without/less ambiguity. Additionally, semantic web technology has not only contributed to developing web applications, but also a typical way to share and reuse data in various sectors (Alzaabi et al., 2013). It also provides a solution for various difficult tasks such as searching and locating specific information in big heterogeneous resources (Fensel et al., 2002). However, the strength of the semantic web relies on the ability to interpreting of relationships between heterogeneous resources.

The core of semantic web is a model known as an ontology that is built to share a command understanding knowledge and reuse

them by linked many resources together in order to retrieve them semantically (Allemang and Hendler, 2011). Some modelling languages are introduced by the semantic web standards such as Resource Description Framework Schema (RDFS) and the Web Ontology Language (OWL). The first layer of the semantic web is based on XML schema to make sure that a common syntax and a resource metadata exist in a structured format, which is already achieved at the previous step in this proposal system.

The second layer is a use of RDF for representing information about resources in a graphical form. RDF relies on triples subject-predicate-object that forms a graph of data. In addition, each triple consists of three elements, namely subject, predicate, and object. The standard syntax for serializing RDF is XML in the RDF/XML form. The third layer is OWL that has found to solve some limitations of RDF/RDFS because it is derived from description logics, and offers more constructs over RDFS. In the other word, a use of RDF/XML and OWL forms ease the search about relative artefacts semantically from multiple resources.

 © 2016 ADFSL

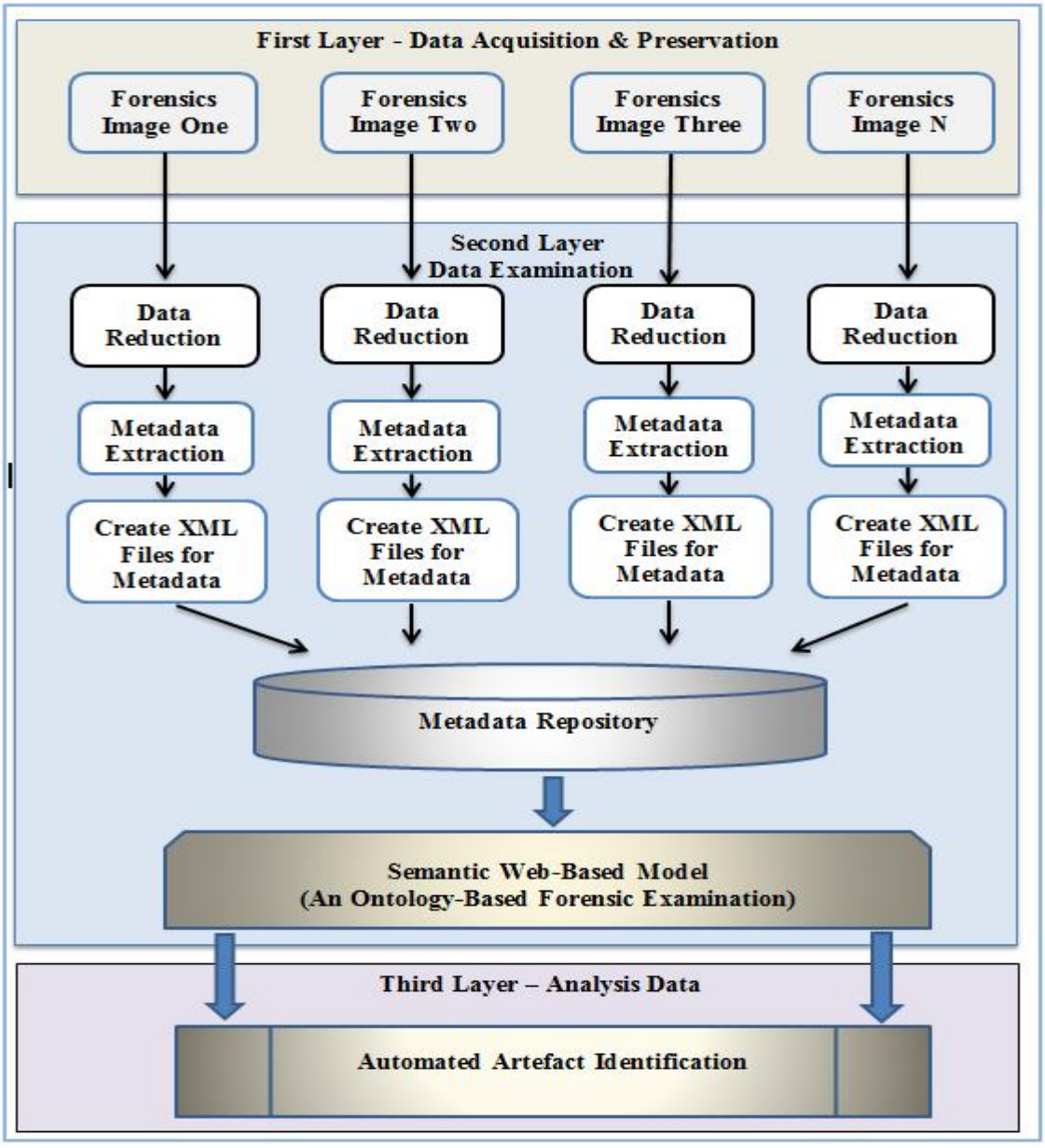

Figure 1. Semantic Web-Based Framework for Metadata Forensic Examination and Analysis

Also, the purpose of this layer is to extract concepts from different files and determine to which class this concept belongs to base on the ontology layer. Examples of such concepts that can be obtained from an email are: an email address and a word attachment belong to the contact class and the document class respectively. For example, each email holds the sender and receiver information, and timestamp. This information can be linked to an email contact instance and becomes the contact class gives two instances with same timestamp.

After the ontology has built and the semantic web technology is ready to use, a search engine will be designed based on the

© 2016 ADFSL

ontology. The search engine is used to retrieve related artefacts based on the investigator's query.

4.3 Data Analysis

The major aim of this layer is to answer the essential questions in an investigation: what, who, why, how, when and where. In addition, find the relation between the artefacts in order to construct the event.

4.3.1 Automated Artefacts Identification

The output from the previous layer is many artefacts related to the incident form heterogeneous resources. During this layer, artificial intelligent techniques will be used in order to find the association between these artefacts and reconstruct the event based on that. In addition, in order to determine the evidential artefacts, all the output artefacts from the previous layer should be analyzed that were created, accessed or modified closer to the time of a reported incident that is being investigated; however, to generate a timeline across multiple resources, it poses various challenges (e.g. timestamp interpretation). As a result, it is necessary to generate a unified timeline between heterogeneous resources; therefore, some tools will be used to cope with these issues. Of course, the answers will be provided to the questions that will be raised during forensics analysis. Preliminary research undertaken by Al Fahdi (2016) has shown that automated artefact identification is an extremely promising approach during the forensics analysis phase.

5. DISCUSSIONS

A number of studies in the literature section present comprehensive surveys of existing work in forensics analysis, with different types and sizes of data from various fields, showing significant increases in the volume of data and the amount of digital evidence needing to be analysed in investigations. Therefore, a requirement on the abandonment or modification of well-established tenets and processes exists. Accordingly, a number of solutions have already been suggested to cope with these issues; however, few researchers have proposed technical solutions to mitigate these challenges in a holistic way. Although data clustering and data reduction techniques are reasonable solutions to cope these challenges, there are few studies in this regard. Also, there is a growing need to optimise these solutions in a comprehensive framework so as to enable all the issues to be dealt with together. As a result, the semantic web-based framework for metadata forensic examination and analysis is proposed to identify potential evidence across multi-resources in order to conduct the analysis. In order to achieve the acquisition and preservation, various techniques will be used base on the resource type. For example, the dead acquisition can be used to collect data from certain types of resources (e.g. hard disk and mobile), but may not for others (e.g. network traffic and online databases). Therefore, it requires effective techniques to obtain required information.

The second layer goals to achieve the examination phase in a proper way by applying a number of processes. The digital forensics reduction is the first process that explained in this layer in order to reduce the volume of data for pre-processing by determining potential relevance data without significant human interaction. The metadata extraction is an essential process in this framework because all later processes will depend upon metadata that has been extracted. In addition, the metadata will be used to reconstruct the past event. Moreover, the use of the XML file to represent metadata will help during a semantic web process. Afterward, the using of metadata repository is beneficial to gathering all metadata from

 © 2016 ADFSL

heterogeneous sources. Accordingly, semantic web technology will integrate metadata that exists in the repository by using ontology in order to retrieve the associated metadata. After that, a variety of artificial intelligence and analysis methodologies will apply to obtain potential evidence in a meaningful way that can use in the court. It has therefore been decided to implement this framework to overcome the three issues of volume, heterogeneity, and cognitive burden.

6. CONCLUSION

The proposed framework aims to integrate big data from heterogeneous sources and to perform automated analysis of data – which tackles a new of key challenges that exist today.

To achieve and validate the approach requires future research. This effort will be focussed upon the development of experiments to evaluate the feasibility of utilising metadata and semantic web based ontologies to solve the problems of big data and heterogeneous data sources respectively. The use of the ontology based forensics provides a semantic-rich environment to facilitate evidence examination. Further experiments will also be conducted to further evaluate the use of machine learning in the ability to identify and correlate relevant evidence. A number of scenario-based evaluations involving a number of stakeholders will take place to evaluate the effectiveness of the proposed approach.

© 2016 ADFSL

REFERENCES

Allemang, D., & Hendler, J. (2011). Semantic web for the working ontologist: effective modeling in RDFS and OWL: Elsevier.

Alzaabi, M., Jones, A., & Martin, T. A. (2013). An ontology-based forensic analysis tool. Paper presented at the Proceedings of the Conference on Digital Forensics, Security and Law.

ALfahdi M 2016. Automated Digital Forensics & Computer Crime Profiling. Ph.D. thesis, Plymouth University.

Ayers, D. 2009. A second generation computer forensic analysis system. Digital investigation, 6**,** S34-S42.

Benredjem, D. 2007. Contributions to cyber-forensics: processes and e-mail analysis. Concordia University.

Beebe, N. L., & Liu, L. (2014). Clustering digital forensic string search output. Digital Investigation, 11(4), 314-322.

Buchholz, F., & Spafford, E. (2004). On the role of file system metadata in digital forensics. Digital Investigation, 1(4), 298-309.

Case, A., Cristina, A., Marziale, L., Richard, G. G., & Roussev, V. (2008). FACE: Automated digital evidence discovery and correlation. Digital Investigation, 5, S65-S75.

Casey, E. (2011). Digital evidence and computer crime: Forensic science, computers, and the internet: Academic press.

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. Mobile Networks and Applications, 19(2), 171-209.

Cheng, X., Hu, C., Li, Y., Lin, W., & Zuo, H. (2013). Data Evolution Analysis of Virtual DataSpace for Managing the Big Data Lifecycle. Paper presented at the Parallel and Distributed Processing Symposium Workshops & Ph.D. Forum (IPDPSW), 2013 IEEE 27th International.

da Cruz Nassif, L. F., & Hruschka, E. R. (2011). Document clustering for forensic computing: An approach for improving computer inspection. Paper presented at the Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on.

Dash, P., & Campus, C. (2014). Fast Processing of Large (Big) Forensics Data. Retrieved from http://www.idrbt.ac.in/PDFs/PT%20Reports/2014/Pritam%20Dash_Fast%20Processing%20of%20Large%20(Big)%20Forensics%20Data.pdf

Fensel, D., Bussler, C., Ding, Y., Kartseva, V., Klein, M., Korotkiy, M., . . . Siebes, R. (2002). Semantic web application areas. Paper presented at the NLDB Workshop.

Fisher, D., Brush, A., Hogan, B., Smith, M., & Jacobs, A. (2007). Using social metadata in email triage: Lessons from the field Human Interface and the Management of Information. Interacting in Information Environments (pp. 13-22): Springer.

Garfinkel, S. L. (2006). Forensic feature extraction and cross-drive analysis. Digital Investigation, 3, 71-81.

 © 2016 ADFSL

Gholap, P., & Maral, V. (2013). Information Retrieval of K-Means Clustering For Forensic Analysis. International Journal of Science and Research (IJSR).

Kataria, M., & Mittal, M. P. (2014). BIG DATA: A Review. International Journal of Computer Science and Mobile Computing, Vol.3 (Issue.7), 106-110.

Khan, M. N. A. (2008). Digital Forensics using Machine Learning Methods. PhD thesis, University of Sussex, UK.

Li, H. & Lu, X. (2014) Challenges and Trends of Big Data Analytics. P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), 2014 Ninth International Conference on 2014. IEEE, 566-567.

Liu, Y., Liu, X., & Yang, L. (2010). Analysis and design of heterogeneous bioinformatics database integration system based on middleware. Paper presented at the Information Management and Engineering (ICIME), 2010 The 2nd IEEE International Conference on.

Mezghani, E., Exposito, E., Drira, K., Da Silveira, M., & Pruski, C. (2015). A Semantic Big Data Platform for Integrating Heterogeneous Wearable Data in Healthcare. Journal of Medical Systems, 39(12), 1-8.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. Journal of Big Data, 2(1), 1-21.

Noel, G. E., & Peterson, G. L. (2014). Applicability of Latent Dirichlet Allocation to multi-disk search. Digital Investigation, 11(1), 43-56.

Palmer, G. (2001). A road map for digital forensic research. Paper presented at the First Digital Forensic Research Workshop, Utica, New York.

Patrascu, A., & Patriciu, V.-V. (2013). Beyond digital forensics. A cloud computing perspective over incident response and reporting. Paper presented at the Applied Computational Intelligence and Informatics (SACI), 2013 IEEE 8th International Symposium on.

Quick, D., & Choo, K.-K. R. (2014). Data reduction and data mining framework for digital forensic evidence: storage, intelligence, review and archive. Trends & Issues in Crime and Criminal Justice, 480, 1-11.

Raghavan, S. (2014). A framework for identifying associations in digital evidence using metadata.

Raghavan, S., Clark, A., & Mohay, G. (2009). FIA: an open forensic integration architecture for composing digital evidence Forensics in telecommunications, information and multimedia (pp. 83-94): Springer.

Raghavan, S., & Raghavan, S. (2014). Eliciting file relationships using metadata based associations for digital forensics. CSI transactions on ICT, 2(1), 49-64.

Roussev, V., & Quates, C. (2012). Content triage with similarity digests: the M57 case study. Digital Investigation, 9, S60-S68.

Rowe, N. C. (2014). Identifying forensically uninteresting files using a large corpus Digital Forensics and Cyber Crime (pp. 86-101): Springer.

Rowe, N. C., & Garfinkel, S. L. (2012). Finding anomalous and suspicious files from directory metadata on a large corpus Digital Forensics and Cyber Crime (pp. 115-130): Springer.

© 2016 ADFSL

Ruback, M., Hoelz, B., & Ralha, C. (2012). A new approach for creating forensic hashsets Advances in Digital Forensics VIII (pp. 83-97): Springer.

Shang, W., Jiang, Z. M., Hemmati, H., Adams, B., Hassan, A. E., & Martin, P. (2013). Assisting developers of big data analytics applications when deploying on hadoop clouds. Paper presented at the Proceedings of the 2013 International Conference on Software Engineering.

Sharma, A., Chaudhary, B., & Gore, M. (2008). Metadata Extraction from Semi-structured Email Documents. Paper presented at the Computing in the Global Information Technology, 2008. ICCGI'08. The Third International Multi-Conference on.

Vaarandi, R. (2005). Tools and Techniques for Event Log Analysis: Tallinn University of Technology Press.

Xu, X., YANG, Z.-q., XIU, J.-p., & Chen, L. (2013). A big data acquisition engine based on rule engine. The Journal of China Universities of Posts and Telecommunications, 20, 45-49.

Zhenyou, Z., Jingjing, Z., Shu, L., & Zhi, C. (2011). Research on the integration and query optimization for the distributed heterogeneous database. Paper presented at the Computer Science and Network Technology (ICCSNT), 2011 International Conference on.

Zuech, R., Khoshgoftaar, T. M., & Wald, R. (2015). Intrusion detection and Big Heterogeneous Data: a Survey. Journal of Big Data, 2(1), 1-41.

 © 2016 ADFSL