On Big Data Forensic - A Case For Forensic Cloud Environment

Oteng Tabona · Thabiso Maupong · Kopo M. Ramokapane · Thabo Semong · Banyatsang Mphago ·

Received: date / Accepted: date

Abstract Background

The high rise in electronic devices in modern-day society has resulted in crimes in cyber-related crimes as criminals resort to hacking, illegal use of these devices. This is primarily due to perceived high rewards and low chances of being apprehended. The rise in cyber crimes poses a significant challenge to forensic investigators as now they have to process huge volumes of data from a variety of sources within a limited time. This results in investigators taking longer to process cases and in some instances missing links as they deal with data from a variety of sources.

Findings

In this paper, we provide a definition of big data forensics, and then we discuss the challenges associated with digital forensics investigations when dealing with big data. We provide details on how volume, variety, and velocity all pose a huge challenge in digital forensics investigations. We then discuss how a novel solution called Forensic Cloud Environment (FCE) leverages the power of Hadoop, HBase, and MapReduce to provide a solution for big data forensic challenges.
Conclusion

In conclusion, the fact that FCE provides an environment to store huge volumes of data from a variety of sources allows for an improved processing time of data. Hence, providing an environment for big data forensics for the future.

Keywords Big Data · Big Data Forensics · Digital Forensics as a Service · Hadoop · Cloud computing · Forensics.

1 Introduction

The growing amount of digital evidence that is collected and its complexity has impact in the overall processing time in digital forensic. Therefore, investigations now concerns big data [1], hence big data forensics. According to [2], big data forensics is defined as follows.

Definition 1 Big data forensics is a special branch of digital forensics where identification, collection, organisation, and presentation processes deal with a very large-scale dataset of possible evidence to establish the fact about a crime.

While Def. 1 is widely accepted in the literature, it is not reflective of the variety aspect of evidence. Furthermore, it does not emphasise the time aspect associated with criminal investigations. For these reasons, we propose the following variation of Def. 1.

Definition 2 Big data forensic is a special branch of digital forensic where identification, collection, validation, analysis, interpretation and presentation processes are carried out on large datasets from a variety of evidence sources to establish the facts of a crime promptly.

It follows from Def. 2 that not only is big data forensics concerned with large volumes of data but that data is from a variety of evidence sources, and most importantly the importance of time factor when conducting big data forensics is also captured in the definition. It follows naturally from Def. 2 that big data forensics is characterized by 3 v’s, namely volume, variety and velocity and this is in alignment with the notion of big data proposed in [3]. Consequently, any tool or service designed for big data forensics must be able to effectively address the 3 v’s and facilitate for the timely conclusions of investigations. In this paper we make a case for Forensic Cloud Environment (FCE), proposed in [4], as a “perfect” solution for big data forensics as its design provides a platform for dealing with the 3 v’s and consequently providing a solution for conducting investigation in a timely manner.

In the following, we provided details on the 3 v’s, how they are evolving and the challenges faced by the current state of art digital forensic solution concerning each v.
1.1 Volume

Volume describes the sheer amount of data that is available from evidence sources. The Regional Computer Forensic Laboratory (RCFL) [5] statistics indicates that the size of evidence from 2006 – 2013 had grown by over 500%. The total data size that was investigated in 2006 was 916 TB compared to 5973 TB in 2013 in the USA alone. These figures have increased due to the growing number of devices per individual which is currently is at 3.96 devices. This number is expected to increase to about 9 devices by 2025 [6]. The growing number of affordable devices with large amounts of storage coupled with the rise in the number of cyber-crime [7] will inadvertently result in even huge volumes of data to investigate.

The main challenge for investigators when dealing with volume arises from the fact that current state-of-the-art forensic tools such as AccessData Forensic Toolkit (FTK)[8–14] and Guidance EnCase [8–10,14–17] are single workstation based. A single workstation has limited computational and storage capabilities, and it is not easily scalable. Attempts to mitigate against the aforementioned are still at research stage, and examples include digital forensics as service (DFaaS) [18–20], data reduction techniques [21–23], triage [24–29], artificial intelligence [30–33] and data mining [34–36,22]. However, they all suffer various limitations as some are designed to address specific problems, e.g., data reduction and triage. These involves selecting certain parts of the evidence to investigate than doing a full drive analysis, even though it is critical and to do a full drive analysis to ensure completeness of data concerning digital forensic processes [37]. As a result of the inadequate computational power, investigations usually take longer, see [38,18,39], and this is expected to worsen as the volume of data grows.

1.2 Variety

Variety is concerned with different sources of evidence including commonly known sources such as computers, hard drives, USB, Internet of Things (IoT) devices, network data, emails and social media [2]. Furthermore, these sources come in different formats and file systems coupled with various types of data representation or format such as text, images, video, and audio [40]. Moreover, the advent of technology, such as the IoT and Cloud Computing means that forensic investigators must be able to investigate and correlate a variety of evidence sources. However, current forensic tools are closed and not interoperability: as a result, it is not easy to extract intelligence and collaborate evidence from disparate evidence sources even if they belonging to the same case [41, 39]. In most cases, examiners resort to manual analysis to establish correlations between evidence, and this is very demanding, especially when dealing with many big data cases. To make matters worse, forensic collections are heterogeneous, in fact, 95% of the collections are unstructured. However, existing forensic tool designs are based on relational databases which are not suit-
able for storing unstructured [42,43]. Moreover, since these tools are already resource-constrained, it makes it more challenging to incorporate efficient algorithms that can analyse unstructured data [44]. Consequently, in most cases, investigators resort to the manual examination of terabytes of evidence. This is a complicated and error-prone process which may be impractical as more evidence sources of different variety are introduced in the future.

1.3 Velocity

Velocity refers to the rate at which the generated data is been processed and actionable insights identified. High-rate adoption of digital technology means that digital evidence will continuously be generated primarily from various sources. Velocity challenge in big data is an inherent problem, as it is affected by the other v’s (volume and variety). For example, the processing time of digital evidence usually rises with the increase in the size of forensic collection [45,2]. This challenge calls for efficient algorithms to effectively and efficiently analyse these data promptly [44]. Also, the variety of evidence sources and data formats delays investigations and calls for advanced techniques to process the data and extract actionable intelligence.

As previously mentioned, it is of paramount importance to find a Digital Forensics as a Service solution to address the challenges associated with the 3 v’s. One possible solution is the novel Forensic Cloud Environment (FCE) [4]. In this note, we aim to highlight how FCE addresses the 3 v’s, and the critical components for each v. We do not test the time factor associated with carrying an investigation a case in FCE in this paper as that is left for future research. Instead, we show that in conjunction to with being a solution for the 3 v’s, FCE is a suitable platform for doing digital forensic.

The remainder of this paper is as follows. Section 2 concerns details on FCE. It is then followed by Section 3 which discusses the design of FCE for the 3 v’s. In Section 4, we discuss FCE as digital forensic solution. We conclude the paper in Section 5.

2 About FCE

Forensic Cloud Environment (FCE) was first proposed in [4] as Digital Forensics as a Service precisely for big data investigation. It consists of six key components as follows.

1. **Big data solution**: The main component of FCE and also the focus of this paper. It provide a platform to host and process big data cyber-crime evidence from multiple different digital devices.

2. **Evidence correlation**: This feature provides investigators with a platform to correlate evidence from multiple different digital devices for more precise timeline analysis of events. Investigators are able visualize a sequence of events together regardless of the sources.
3. **Collaboration** - The platform facilitates for expert from different backgrounds to collaborate during an investigation.

4. **Intelligence sharing** - One of the innovative features of FCE, it facilitates in “connecting dots” between big data cases hosted in the same FCE and different FCEs in different location.

5. **Knowledge sharing** - To improving the efficiency of investigators, this module assist in knowledge sharing between investigators. Hence minimising the need to redo research when faced with challenges that may have been carried out by other investigators.

6. **Security** - This module is concerned with ensuring that the data in FCE is protected and its integrity is maintained as per ACPO guideline for dealing with digital evidence [46], the Data Protection Act 1998 [47] and that investigation are carried out in a secure and audit able manner.

A typical FCE infrastructural setup consists of multiple node servers with one as the master node and the rest slaves nodes. The master node is normally used to hosts master services such as HBase master, NameNode, Application Programming Interface (API) used for:

- loading data into the FCE like ingestion of forensic images and file system passers;
- security and ensuring integrity of images and files such as hash calculators;
- carrying out an investigation like timeline/network generator, entity extractors, and intelligence share.

The slave nodes normally act as data storage and Region Servers.

### 3 FCE architecture for big data challenges

As previously mentioned, we focus on the first component of FCE, the big data solution for forensics purposes. We discuss how FCE address the 3 v’s challenge, hence making FCE a big data digital forensic solution in line with Def. 2. The solution is facilitated by the core architecture of FCE depicted in Figure 1. The architecture comprises of three layers namely the storage, processing and application. It is this layout that allows for FCE to provide a novel solution to addressing the 3 v’s challenge as we explain in the succeeding subsections 3.1, 3.2 and 3.3.

#### 3.1 FCE for volume

As outline earlier, current digital forensic tools are single workstation-based, and as a result, they cannot efficiently assist in investigating cases that involve big data. For these reasons, there is an urgent need for a scalable platform which can expand in any chosen dimension without the need to modify its architecture (structural scalability) significantly, moreover, continue performing at the required level when the demand increases (load scalability) [48].
FCE design fulfils these requirements simply because in its core architecture the storage block is based on Hadoop Distributed File System (HDFS) [49–51] and HBase [49]. HDFS and HBase are responsible for the structural scalability component of FCE as they are capable of storing large amounts of data in distributed nodes. Furthermore, HDFS can scale up to petabytes using commodity hardware [52,53] and allows for easy commissioning of new data nodes when the demand increases [54]. On the other hand, HBase is a schema-less and distributed column-oriented database built on top of HDFS [55–57]. It can host enormous sparsely populated tables on a cluster of commodity hardware. HBase can handle huge amounts of data in petabytes magnitude and support large scale operations on the data [58]. In addition to being structural scalable, HDFS provides for reliability and availability. Most importantly, it offers integrity of data in HDFS which is critical in digital forensics and has to be maintained at all times. This has been proven in [4].

To address load scalability, FCE uses MapReduce framework [59] at the processing layer to process data. MapReduce is used for parallel processing of large volumes of data stored in HDFS or HBase through multiple nodes in a cluster. An essential feature of the MapReduce model is that it facilitates for code to be executed where the data is located/stored hence reducing bottleneck associated with moving data around for processing. This in contrast to other models, it does not require data to be transferred to the computing node which usually results in network bottlenecks.
3.2 FCE for variety

Big data variety issue is due to different evidence sources and heterogeneous data formats. The increase in the variety of different evidence sources intensifies the problems faced by investigators. This is because current forensic tools are limited to one evidence source per investigation and also lacks enough storage space. Also, traditional forensic tools are based on relational databases which are not designed to store unstructured data.

To address these challenges in the storage layer, FCE leverages on the HDFS and HBase to store all evidence in one area and analyse it as a whole regardless of the source or the structure of the data. This approach brings forth all forensic disciplines including network, mobile, computer, IoT and game console forensic in one place which is in contrast to current practices where each forensic discipline is applied individually, and then findings are manually correlated. In FCE, investigators have the opportunity to visualise a complete overview of the digital landscape of a crime regardless of the evidence sources. Hence, interesting patterns can be discovered through the timeline and network analysis of all evidence sources – an opportunity that is not incorporated in current tools. Furthermore, MapReduce is used to process all kind of data including structured, semi-structured and unstructured, hence reducing the burden of find ways of processing data based on its structure.

3.3 FCE for velocity

Large forensic collection and the complexity of data make investigations time consuming resulting in backlogs [60,29,61]. Moreover, the disparate of evidences sources, the limited of storage and computation power results in slow processing time of data. Hence, investigators cannot process evidence data and get insight full information promptly.

Through the use of NoSQL database (HBase) in FCE, data can be processed at high-speed. The use of MapReduce also significantly improves the speed of data processing of vast amounts of data. Hence, FCE improves the processing of a large amount of data significantly. Another feature of FCE is based on the fact that HDFS, HBase, MapReduce are all scalable concerning storage and computational power, this allows for the incorporation of advance data analysis techniques, like machine learning, to facilitate in evidence discovery and correlation. In addition, the distributed nature of FCE components makes it possible to execute various applications simultaneously. This feature enhances collaboration as multiple investigators can work at the same time resulting in improved turnaround time for investigations. The fact that data from different evidence sources are stored in one common area speeds up the process of finding the correlation between evidence. Finally, the top most layer, application layer in Figure 1, allows for easy development and deployment of application which can speed up process of conducting investigations.
In the previous section 3 we outlined how FCE is a good platform for big data digital forensics as defined in Def. 2. In this section we qualify FCE as a digital forensic solution based on a number of validation metrics.

First we outline “bad” digital forensic practice that can lead to misinterpretation of results of an investigation [62].

- Incompleteness - failure to recover or find all the data from evidence source.
- Inaccuracy
  - existence: do all artefacts presented actually exist?
  - alteration: do methods applied alter the data?
- Inaccurate: do methods applied alter the data?
- Migration: does the solution detect and compensate for missing and corrupted data
- Misinterpretation - is the data presented in its original form?

To mitigate against the above, FCE is designed in accordance with the Scientific Working Group on Digital Evidence (SWGDE) [63] best practices in digital forensics.

- Case isolation - to ensure data between cases are not commingled.
- Data integrity - digital evidence should be maintained in such a way that the integrity of the data is preserved
- Security - prevent the contamination of data between cases and unauthorised access to the evidence
- Hashing - use hashes to verify the integrity of evidence

Following data integrity element of best practices by SWGDE, during forensic imaging, [4] uses existing data acquisition tools. When imaging each evidence source, hash values are calculated, and then the images are ingested into FCE. After ingestion, the hash values are recalculated. The hash values before and after loading data into FCE are compared to verify if the original image were not corrupt during ingestion. The results of [4] demonstrated that FCE maintains the integrity of the images during ingestion.

Also, in [4], FCE’s ability to preserve the integrity of data and files is demonstrated in the following manner. File system parsers were used to extract data from the images in FCE. Following that, the hash values of the extracted data or files were computed. The output was compared to the hash values of data or files from FTK, and the results were matching, which highlights the capability of FCE to maintain data and file integrity. The authors also verified the location of the extracted data in the evidence using FTK and original drive and their tests satisfied incompleteness, inaccuracy, misinterpretation, data integrity and hashing objectives specified by [62,63].

In order to ensure case isolation, FCE store each case separately in HDFS and HBase. That is, each case has its own folder to store case-related files and HBase tables which are prefixed with the case number. Part of security requirement is addressed through case isolation, as for unauthorised action
authentication measures are used to allow only authorised investigators to execute applications on the platform. Also, audit trails are generated for every action that is performed on the evidence.

5 Conclusion

In this paper, we defined big data forensics as a digital forensic investigation that involves all characteristics of big data, more specifically volume, variety and velocity. We further outline how big data has negatively affected digital forensic investigations. We also present how FCE address each of the big data challenges. Furthermore, we qualify FCE as a forensic solution based on general forensic principles and best practices suggested by the SWGDE.

Acknowledgements This work is part of the Botswana International University of Science and Technology (BIUST) Cyber Security Research Group (CSRG).

Conflict of interest

The authors declare that they have no conflict of interest.

Ethics approval

The authors declare that they have no need for ethics approval.

Consent to participate

The authors declare that there is no personal data in use.

Availability of data and material

The authors declare that there is no data and material used at this point.

Funding

The authors declare that they have no funding.

Authors contributions

Oteng and Thabiso were responsible for the conceptualisation, initial research, the drafting and writing of the manuscript. Kopo, Thabo and Banyatsang were responsible for additional research and writing of the manuscript.
References

1. K. Truong, Sophos Naked Security (2013)
2. S. Zawoad, R. Hasan, in 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 (IEEE, 2015), pp. 1320–1325
3. D. Laney, Gartner. Retrieved 21, 2014 (2012)
4. O. Tabona, A. Blyth, in 2016 SAI Computing Conference (SAI) (IEEE, 2016), pp. 579–584
5. R.C.F.L. (RCFL). Fiscal year 2011 - 2013. https://www.rcfl.gov/ (2020)
6. H. Safaei, A.M.H. Monazzah, M.B. Bafroei, A. Ejlali, in 2017 2nd International Conference on System Reliability and Safety (ICRS) (IEEE, 2017), pp. 207–212
7. H.S. Lallie, L.A. Shepherd, J.R. Nurse, A. Erola, G. Epiphanioiu, C. Maple, X. Bellekens, arXiv preprint arXiv:2006.11929 (2020)
8. D. Manson, A. Carlin, S. Ramos, A. Gyger, M. Kaufman, J. Treichelt, in 2007 40th Annual Hawaii International Conference on System Sciences (HICSS’07) (IEEE, 2007), pp. 260b–260b
9. D. Ayers, S34-S42 (2009)
10. J. Dykstra, A.T. Sherman, Digital Investigation 9, S90 (2012)
11. R.M. Saidi, S.A. Ahmad, N.M. Noor, R. Yunus, in 2013 The International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAECEE) (IEEE, 2013), pp. 132–136
12. P.K. Kshobragade, I.G. Malik, in 2014 Fourth International Conference on Communication Systems and Network Technologies (IEEE, 2014), pp. 458–462
13. C. Smith, G. Dietrich, K.K.R. Choo, in International Conference on Security and Privacy in Communication Systems (Springer, 2017), pp. 85–103
14. J. Wagner, A. Rasin, K. Heart, R. Jacob, J. Grier, Digital Investigation 29, S42 (2019)
15. S. Widup, Computer forensics and digital investigation with EnCase Forensic v7 (McGraw-Hill Education Group, 2014)
16. S. Bunting, W. Wei, EnCase Computer Forensics: The Official EnCase Certified Examiner Study Guide (John Wiley & Sons, 2006)
17. H. Kim, N. Bruce, S. Park, H. Lee, in 2016 18th International Conference on Advanced Communication Technology (ICACT) (IEEE, 2016), pp. 722–725
18. R. Van Baar, H. Van Beek, E. Van Eijk, Digital Investigation 11, S54 (2014)
19. H. Van Beek, E. van Eijk, R. van Baar, M. Ugen, J. Bodde, A. Siemelink, Digital Investigation 15, 20 (2015)
20. C. Stelly, V. Roussev, Digital Investigation 22, S39 (2017)
21. M. Scanlon, in 2016 Sixth International Conference on Innovative Computing Technology (INTECH) (IEEE, 2016), pp. 10–14
22. D. Quick, K.K.R. Choo, Trends & issues in crime and criminal justice 480, 1 (2014)
23. S. Neuner, M. Schmiedelcker, E. Weippl, Security and Communication Networks 9(15), 2876 (2016)
24. R.P. Mislan, E. Casey, G.C. Kessler, Digital Investigation 6(3-4), 112 (2010)
25. V. Roussev, C. Quates, R. Martell, Digital Investigation 10(2), 158 (2013)
26. S.L. Garfinkel, Computers & Security 32, 56 (2013)
27. B. Hitchcock, N.A. Le-Khac, M. Scanlon, Digital investigation 16, S75 (2016)
28. E. Gentry, M. Soltys, Procedia Computer Science 159, 1652 (2019)
29. R. Montasari, R. Hill, in 2019 IEEE 12th International Conference on Global Security, Safety and Sustainability (ICGSS) (IEEE, 2019), pp. 205–212
30. F. Mitchell, Digital Evidence & Elec. Signature L. Rev. 7, 35 (2010)
31. A. Irons, H.S. Lallie, Future Internet 6(3), 584 (2014)
32. H. Mohammed, N. Clarke, F. Li, (2016)
33. S. Costantini, G. De Gasperis, R. Olivieri, Annals of Mathematics and Artificial Intelligence 86(1-3), 193 (2019)
34. N.L. Beebe, J.G. Clark, Digital investigation 4, 49 (2007)
35. K. Sindhu, B. Meshram, (2012)
36. A.J. Tallón-Ballesteros, J.C. Riquelme, in *Computational Intelligence in Digital Forensics: Forensic Investigation and Applications* (Springer, 2014), pp. 413–428
37. D. Quick, K.K.R. Choo, Digital Investigation 11(4), 273 (2014)
38. S. Gabbianelli, digital investigation 7, S64 (2010)
39. S. Kaisler, F. Armour, J.A. Espinosa, W. Money, in 2013 46th Hawaii International Conference on System Sciences (IEEE, 2013), pp. 995–1004
40. E. Casey, G.J. Stellatos, ACM SIGOPS Operating Systems Review 42(3), 93 (2008)
41. J. Haggerty, A.J. Karran, D.J. Lamb, M. Taylor, International Journal of Digital Crime and Forensics (IJDCF) 3(3), 1 (2011)
42. M. Qi, in 2014 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD) (IEEE, 2014), pp. 734–739
43. I. Richard, V. Roussev, in *Digital crime and forensic science in cyberspace* (IGI Global, 2006), pp. 75–90
44. V. Roussev, L. Wang, G. Richard, L. Marziale, in *IFIP International Conference on Digital Forensics* (Springer, 2009), pp. 201–214
45. Association of Chief Police Officers (ACPO), *ACPO Good Practice Guide for Digital Evidence* (2012)
46. D.P. Act, London Station Off (1998)
47. A.B. Bondi, in *Proceedings of the 2nd international workshop on Software and performance* (2000), pp. 195–203
48. T. White, *Hadoop: The Definitive Guide*, 3rd edn. (O’Reilly Media, 2012)
49. D. Borthakur, J. Gray, J.S. Sarma, K. Muthukkaruppan, N. Spiegelberg, H. Kuang, K. Ranganathan, D. Molkov, A. Menon, S. Rash, et al., in *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data* (2011), pp. 1071–1080
50. C. Lam, *Hadoop in action* (Manning Publications Co., 2010)
51. W. Ryu, ALLDATA 2017 p. 10 (2017)
52. V.D. Jogi, A. Sinha, in *2016 3rd International Conference on Recent Advances in Information Technology (RAIT)* (IEEE, 2016), pp. 586–590
53. P. Riyaz, S.M. Varghese, Journal of Computer Engineering 17(3) (2015)
54. G.S. Inspectorate. Changing policing in ireland (2015)
55. Crown. Method validation in digital forensics. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/528123/FSR_MethodValidation_in_Digital_Forensics_FSR-G-218_Issue_1.pdf (2016)
56. SWGDE. https://www.swgde.org/documents/published (2020)