2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)

Digital Forensics Challenges to Big Data in the Cloud

Xiaohua Feng
Department of Computer Science and Technology
University of Bedfordshire
Luton, UK
Xiaohua.feng@beds.ac.uk

Yuping Zhao
Microwave Wireless Telecommunications Research Lab
School of Electronics Engineering and Computer Science
Peking University, Beijing, China
yuping.zhao@pku.edu.cn

Abstract — **As a new research area, Digital Forensics is a subject in a rapidly developing society. Cyber Security for Big Data in the Cloud is getting more attention than ever. A computing breach requires digital forensics to seize digital evidence to determine who is responsible and what has been done maliciously and the possible further consequences. In particular, for Big Data attack cases, Digital Forensics is facing even more challenge for earlier digital breach investigations.**

For the PPI (Protection of Personal Information) a GDPR (General Data Protection Regulation) law has been launched to be implemented from the 25th May 2018. This compulsory regulation will have an important impact on healthcare PPI in the cloud (ICO, 2017; Deloitte, 2014).

Nowadays, Big Data with the characteristics of three "V"s (Volume, Velocity, and Variety), are either synchronized with the Cloud, or stored in the Cloud, in order to solve the storage capacity and so on problems, which made Digital Forensics investigation even more difficult. The Big Data Digital Forensics issue for the Cloud is difficult. One of them is the need to identify which physical devices have been compromised. Data are distributed in the Cloud, so the customer or digital forensics practitioner cannot have full access control like the traditional investigation does.

Smart City are making use of ICT (information communications technology) to collect, detect, analyze and integrate the key information data of core systems in running the cities. Meanwhile, the Control Centre is making intelligent responses to different requirements that include daily livelihood, PPI security, environmental protection, public safety, industrial and commercial activities and city services. The Smart City healthcare Big Data are collected and gathered by the IoT (Internet of Things) (Liu, 2014; Qi, 2016) and applying GDPR prevent Cyberstalking and Cybercrimes.

This paper summerises our review on the trends of Digital Forensics used for Big Data. The evidence acquisition challenge is discussed. A case study of a Smart City project with IoT services collecting Big Data which are stored in the Cloud computing environment is represented. The techniques can be generalised to other Big Data in the Cloud environment.

*Keywords –**Big data, Cloud Computing, Smart City system, IoT, Cyber Security and Personally Identifiable Information (PII) GDPR and Protection of Personally Information (PPI) against Cyberstalking***

I. INTRODUCTION

Digital Forensics is important in cyber security. Evidence data is crucial for Digital Forensics. Big data evidence collecting is challenging in the cloud.

While Big Data is defined as a large collection of data sets this is too complex to be processed by handoff achievable. Instead, it needs database management tools to be handled. Big Data is a technology which analyses an amount of data in order to find out the profiles, rules, collect valuable opinions and predict complex problems. It maybe involves capture, storage, search, sharing, transfer, analysis and visualization. Big Data has Volume, Velocity, and Variety (Veracity); especially in the Cloud environment these three "V"s determine their characteristics.

Volume is the amount of data, usually when data capacity reaches up to Terabyte (1024 GB) and Petabyte (1024 TB, 1,048,576 GB). Velocity is the speed of collecting, processing and using data. Big Data Security often needs speedy processing or even real time processing of data. Velocity becomes a vital measure of big data usage, especially when real time data process is required. Variety is the type of big data involved. In the current era, Big Data usages include text, audio, video and multimedia types of data format. Currently, there is large volume of data available, which brings challenges for digital investigation. For example, challenges are in data collection, data analysis, data recovery and many more (Harshish and Feng, 2011). The problem solving is complicated and is still in progress

978-1-5386-3066-2/17 $31.00 © 2017 IEEE
DOI 10.1109/iThings-GreenCom-CPSCom-SmartData.2017.132

(Alessandro, 2015). At the Smart City project of this research, Big Data are collected by the IoT automatically; which is a system combining information and communications technology to make use of radio frequency identification and electronic product code techniques to service global applications.

In our case study, the collected Smart City Big Data are stored at the intermediate nodes, (Liu, 2014) then the clouds. This is challenging to digital forensics whenever an investigation is needed. I will discuss the details next.

II. RECENT DEVELOPMENT

With the development of technologies, our society has changed enormously. These have a big impact on conventional digital forensics investigations. With the enhancement, Big Data could provide much more in terms of quality of service than before.

II.1 Big Data Development

The recent data science development such as health record, daily sports data, and national healthcare as well as Smart City project demonstrated the future trends on Big Data applications. Big Data are difficult for Digital Forensics. Identifying the cloud security threats is largely about looking for data patterns that are out of the ordinary, whether it is an unauthorised user from an unknown IP address or a distributed denial of service (DDoS) attack. Understanding Big Data techniques allow you to analyze cloud incoming and outgoing traffic to reveal anomalies that point to a data breach (Ernst and Young, 2015). Its Volume, Velocity, and Variety characteristics has attracted majority of researchers pay more attention on social media and data science related subjects (Sremack, 2015).

II.2 The Cloud Computing Development

There are three popular cloud computing service models known as:

SaaS (Software as a Service) the best known service model
PaaS (Platform as a Service) service model and
IaaS (Infrastructure as a Service) service model.

Their pros and cons are: at SaaS, the only thing users could be get involved is the access control (AC).

While at PaaS, apart from access control, users could also get involved with applications.

And at IaaS, users could be got involved with data load and operating system as well.

These characteristics heavily affected Digital Forensics investigations. Further more, PII (Personally identifiable information) is attracting more and more attention and related to majority people's everyday activities, such as Carphone warehouse and G20 cases (BBC, 2015). That added a pressure on Digital Forensics to investigate PII breach in the cloud (Feng, 2011b; Antoine, 2013).

III. THE FORENSICS CHANLLENGES

Due to the cloud distribution feature, the conventional data acquisition regulation is not sufficient to meet the digital forensics evidence requirements. For instance, it is almost impossible to seize a physical hard drive to get all of the related forensics evidence for a case.

Many digital Forensics evidence acquisition issues depend on the CSP (Cloud Service Provider)'s support or co-operation. Zawoad (2013) has listed some updated cloud issues. The distributed data centre may cross several national borders in storage. That might make the Chain of Custody (CoC) as an important part of the documentation is extremely difficult. When Big Data in the cloud, the investigating officer has no physical control of device, the only possibility is rely on the CSP. If the related CSP is not technically competent or without forensics training, there is no guarantee that the audit trail could be put into the Chain of Custody forms and could be completed appropriately for the future testimony.

Multi-state legal issues could affected the process of obtaining permissions for authentication of evidence search as well. For instance, if the case Big Data Centre located a country related in which does not have a Data Protection Act or GDPR kind of guidance, trying to get the authorised data will be difficult.

A summary of Big Data digital forensics challenges in the cloud is reviewed and the result is shown as Table 1. Where the issues marked a yes (a "Y") shown outstanding problems; means further development work is still required currently; while issues that have a solution have been marked a no; (*i.e.*, an "N" instead).

IV. A CASE STUDY

Since 2014, a world-leading project in autonomous transportation systems and intelligent mobility (Sant, 2015) has started. The Milton Keynes Smart City project focuses on the application of the next-generation of information technology to all walks of life, thereby embedding sensors and equipment in driverless vehicles on roads and railways, bridges and tunnels, mobile communications systems, and others in every corner of the place; thereby forming the Internet of things through the Internet. This will enable us to integrate the IoT through powerful computer clusters and cloud computing. This will enable people to manage productivity and life more meticulously and in a more dynamic manner, leading to a state of global intelligence.

It is a typical Big Data collected by the IoT and stored in the computer cloud. Here, we will discuss about Big Data in Smart City project where IoT and raspberry pi gathered information; *i.e.*, Big Data in the cloud and their access security issues emerged, as well as the Digital Forensics cloud challenge solution with the impact on PII (ISO, 2014).

In this case study, the IoT terminal devices to collect customer's information data, the data volume of which is consistently increasing; *i.e*, Big Data formation. Then the substantial various data are transmitted to the cloud service provider to come up as Big Data in the cloud. The Smart City Control Centre according to these data to provide local information and suggest the best service, in order to make the customer have a very enjoyable stay at Milton Keynes. Therefore, a large amount of information and interactive data

generated. And in order to get Cyber Security, data acquisition, processing and storage are consequently vital.

In the case study, the IoT terminal device to collect customer information, the data volume of which is consistently increasing; *i.e*, Big Data formed. Then the substantial data of various kinds are transmitted to the cloud service provider to come up as Big Data in the cloud. The Smart City Control Centre uses these data to provide local information and suggest the best service, in order to make the customer have a very enjoyable stay at Milton Keynes.

However, if these Big Data are breached, the provided information could be false, which maybe leads to serious crime. Then the digital forensics practitioner will walk in and acquire evidence to locate the suspect.

Nevertheless, in the cloud, Big Data of the Smart City are not the only data to exist in that cloud. It might be distributed at several locations and at different sectors of the storage. If these data are hacked, the driverless vehicle is crashed. It would be extremely difficult to get a warrant, visit the crime scene, acquire the possible evidence, image the Big Data, fill-in the CoC forms, preserve the acquired evidence (as Table 1 shown). Then examine recovered data and work out the analysis to report to the law enforcement.

The investigating officer needs to bear in mind he does not own the cloud utility, no matter SaaS, PaaS or IaaS as the following table shown.

Many tasks need to be done by the CSP's help. Even under the CSP's support not all the investigation works could be complete as an individual digital/computer system experienced (Feng, 2015). The SaaS is the best developed service. However, apart from access control, authorized investigation officer cannot control the applications, data load, operating systems, servers and network system. IaaS is the closest to hardware scenario when a digital forensics cloud investigation is carried out. However, an IaaS user cannot get control of the servers and network system of the IaaS fully for the authorized investigation officer.

V. DISCUSSIONS

Due to the volume issue, the investigation officer requires an adequate bandwidth to image the virtual machine of the Big Data in the cloud. Ruan (2015) did a survey on the cloud forensic capacity, the results showed that the majority lies on CSPs. With support from CSP, a read only API provided by the case relevant CSP for network, process and access logs to the representative of the customer to acquire some data by read only. Marty (2011) proposed an Ajax library for logging checking or imply log management in the Cloud by the CSPs.

Hegarty and Feng (2011) suggested for each time uploading or downloading, checking data integrity as one of the problem solving solution; if anything suspicious going on, isolate the cloud immediately. Nevertheless it seems to complicate the cloud activities.

Another problem is potential multi location raised multi state-laws. In order to specify CSPs' responsibility in the cloud cyber security as well as their role in a digital forensic investigation, an updated Service Level Agreement (SLA) needs to be published and put into operation. Globally collaboration is required, all the states laws should apply.

Digital Forensics Challenges to Big Data in the Cloud	Cloud Services			Note
	SaaS	PaaS	IaaS	
Physical Access control/accessibility	Y	Y	Y	
CSP dependent	Y	Y	Y	Logs
Volatile data	Y	N	N	
Trustiness	Y	Y	Y	
Bandwidth	Y	N	N	
Multi-tenancy	Y	Y	N	
Distributed logs	Y	Y	Y	
Volatility of logs	Y	N	N	
Logs in multiple tiers and layers	Y	Y	Y	
Logs accessibility	Y	Y	Y	
Logs lack of critical information	Y	Y	Y	
Chain of Custody	Y	Y	Y	AC issues
Issues on existing forensic tools	Y	Y	Y	
Crime scene recognition	Y	Y	Y	
Crime scene reconstruction	Y	Y	N	
Multi-state laws	Y	Y	Y	
Report	Y	Y	Y	
Compliance	Y	Y	N	
Presentation	Y	Y	Y	
Combination issue	Y	Y	Y	
Integrity	Y	Y	Y	
Warrant	Y	Y	Y	
Localization	Y	Y	Y	locate
GDPR	Y	Y	Y	

Table 1 Digital Forensics Challenges Big Data in the Cloud

VI. EVALUATION AND CONCLUSIONS

In this research, we have explored the main Digital Forensics challenges about the Big Data with cloud Computing PII in the cloud services. (Rezendes, 2015; ISO, 2014). Updated impact has been discussed and problem solving solutions have been explored and critically analyzed. A series of design and implementation phases steps have been carried out and there is still new development going on at the NCCR (National Centre of Cyberstalking Research) University of Bedfordshire.

When the recent development finishes at the end of June, we can make some decision, according to further experimental

results comparison and analysis (Feng, 2016), to recommend a more appropriate approach to solve the Big Data at the cloud service environment issues for digital forensics investigations.

These digital forensics challenges with Big Data are crucial. Almost all the traditional way of investigation is not appropriate any longer, as the cloud users have lost control on many aspects. Here, a problem solving research with Big Data in the cloud environment is significant for the upcoming cyber security applications in data science, such as the Smart City project. There is a consequence impact on develop technology in the future cloud information governance, risk management, and compliance, trustworthiness (Liu, 2014), as well as human factor in Big Data Cyber Security.

In particular, on Smart City projects, PPI could be a cutting-edge element for citizen's trust (Petit, 2015; Hoppe, 2008) and leads the consequence on the development and applications (Feng, 2015). Since being created in 2004, European Network and Information Security Agency (ENISA) had done many works for the EU states. The Big Data and cloud development Big Data & Smart Sustainable Society Workshop -2016 pushes ENISA to produce more guideline and regulations to protect Big Data in the cloud. To date, we have thoroughly investigated the outstanding issues, but there is still no perfect solution yet. There are more innovative work needed.

VII. SUGGESTED FURTURE DEVELOPMENTS

For some of the Big Data digital forensics challenges we have summarized today, there are plenty of development needing to be carried out to obtain the best trade-off between cloud application services and digital forensics in the Big Data cyber security; such as, analyze social media Big Data for cyber psychology issues; or analyze Big Data in the cloud for national citizen's health or sports activity record, data service for centralized national healthcare as well.

To date, there is still neither an appropriate professional Digital Forensic toolkit to carry out computing cloud investigations, nor for the Big Data cases. Guidance Software and the Digital Forensic vendors need to speed up their developments.

Nowadays application research on cyber security field is getting more and more attention. Cyber psychology is one of the new explorations. Cyber psychology is research on any aspect of human psychology related to Cyber technology (Feng, 2016; Deekue, 2013).

As a new cross-subject area, its aim and objectives are using the conventional psychology view to analyze the newly developed cyber techniques. Then provide user guide to make use of cyber resource and worked out efficient network management. Currently University of Bedfordshire NCCR is working on social media chat messages, which is another Big Data application. Its security is not only of the technical side, but also in social psychology category to follow the GDPR and ISO 27018 for PPI against Cyberstalking. So, using digital forensics technique to acquire social network messages and analyze the content becomes another Big Data challenge. With Big Data and the cloud technology development, digital forensics will face more and more challenges in the near future. There are plenty of explorations for us to discover and work out the reasonable problem solving solutions.

REFERENCES

Alessandro Guarino (2015), "*Digital forensics as a Big Data challenge"*, eBook – 24 Aug 2015

Antoine Olivier, Mitchell Chris and Phillips John (2013), "*ISO/IEC 27018: The future standard for Personal Data protection in public cloud*", EBRC.

Zafarullah, Z.; Anwar, F. and Anwar Z. (2011), "*Digital forensics for eucalyptus*," in Frontiers of Information Technology (FIT). IEEE, 2011, pp. 110–116.

BBC (2015): "*Carphone warehouse customer data breach investigated*", BBC, UK, http://www.bbc.co.uk/news/uk-33840327, (Accessed: 08/08/2015).

Bellare, M. and Rogaway, P. (2005): "*Introduction to modern Cryptography*". Verimag, France, pp.10. Available atwww-verimag.imag.fr/~plafourc/teaching. (Accessed: 04/10/2015).

Carrier, B. (2005): "*Filesystem forensic analysis*", Indiana: Addison Wesley Professional, 2005, ISBN: 9780321268174

Cohen, R. (2012) '*The past, the present, and the future of cloud computing*', Intel Technology Journal, 16 (4), pp.20-24.

Cole E. (2015) "*Cyber Security Expert Professional*", http://www.securityhaven.com (Accessed: 22/2/2017)

Deekue S.; Feng X. and Liu, E (2013): "*A strategic framework for Nigeria e-government security*", ARSR2013 Workshop, University of Bedfordshire and Manchester, UK

Deloitte (2014) "*Cloud Computing and PPI: Finding your bearing*",http://www.deloitteblog.co.za/cloud-computing-and-ppi-finding-your-bearing/ {Accessed: 20/3/2017)

Delport Waldo M. K. and Olivier Martin S. (2011), "*Isolating a cloud instance for a digital forensic investigation*", proceedings of the Information and Computer Security Architecture (ICSA).

Dykstra J. and Sherman A. (2011), "*Understanding issues in cloud forensics: Two hypothetical case studies*," Journal of Network Forensics, vol.b, no. 3, pp. 19–31, 2011.

Ernst and Young (2013), "*Demystifying 'Big Data' analysis*". EY. http://www.isacantx.org/Presentations/2013-05Post-DemystifyingBigDataAnalytics, (Accessed: 28/03/2016).

Fahad Abdullah Al (2015), "*Cloud computing security policy*", University of Bedfordshire, UK.

Farrell Paul (2015), "*Personal details of world leaders accidentally revealed by G20 organisers*", The Guardian, UK, March, 2015

http://www.theguardian.com/world/2015/mar/30/personal-details-of-world-leaders-accidentally-revealed-by-g20-organisers (Accessed: 31/3/2015)

Feng X. and Zhao Y. (2016) "*Cyber Security Data Mining with the Mobile Internet of Things (IoT) - Personally Identifiable Information on Smart Technologies*", proceeding of the DaMIS 2016 - International Workshop on Data Mining on Internet of Things Systems, U.K. 2016

Feng, X. and Zhang X. (2015) "*Personally identifiable information security in cloud computing*", IEEE International Conference on Computing and Technology Innovation UK.

Feng, X. (2011a) "*Computer Law in UK*", UCC Data Retriever, Digital Library Workshop, Ireland

Feng X. (2011b) "*Incidence Response Strategies*", the 7th Annual Forensics Workshop, U.K.

Han, L. and Kendall G. (2003) "*An investigation of a tabu assisted hyper-heuristic genetic algorithm*", Evolutionary Computation. CEC'03. The 2003 Congress on 3, pp2230.

Harshish M and Feng X. (2011) "*Challenges on Forensics, A Cloud investigations reference model*", STAN-2011, IEEE Symposium of Security, Technology and Networks.

Hoppe, T. Kiltz S., and Dittmann J., (2008). "*Security threats to automotive CAN networks-practical examples and selected short-term counter-measures*". Proceedings of Computer Safety, Reliability, and Security. 5 (2), pp. 235–248.

Grispos, G. Storer, T. And Glisson, W. (2012), "*Calm before the storm: the challenges of cloud computing in digital forensics"*, International Journal of Digital Crime and Forensics (IJDCF), 2012

ICO (2014), "*Find out how to request your personal information*",http://ico.org.uk/for_the_public/personal_information. (Accessed: 7/5/2015).

ICO (2017) "*Overview of the General Data Protection Regulation (GDPR)*", https://ico.org.uk/for-organisations/data-protection-reform/overview-of-the-gdpr/ (Accessed: 20/3/2017).

ISO/IEC 27018: (2014), "*Information technology-- Security techniques -- Code of practice for protection of personally identifiable information (PII) in public clouds acting as PII processors, ISO 27018 standard*".

http://www.iso27001security.com/html/27018.html (Accessed: 24/3/2015)

Khoshgozaran A, Shirani-Mehr Houtan & Shahabi C. (2012), "*Blind evaluation of location based queries using space transformation to preserve location privacy*", Geoinformatica Journal, Volume 27, Issue 2, pp 413-427,

ISBN: 978-3-642-03510-4 Nov 2012.

Liu, E and Feng X. (2014): "*Trustworthiness in the Patient Centered Health Care System, Series: Communications in Computer and Information Science*", Volume, 426, Springer-Verlag, March, 2014

Ludwig Slusky M. D. and Parviz Partow-Navid, (2012) "*Cloud computing and computer forensics for business applications*," Journal of Technology Research, vol. 3.

Katz, J. and Lindell, Y. (2008), "*Introduction to modern cryptography*". Boca Raton: Chapman & Hall/CRC.

Marty R. (2011) "*Cloud application logging for forensics*," proceedings of ACM Symposium on Applied Computing.

Mollin, R.A. (2007), "*An introduction to cryptography*". 2nd Edition. Boca Raton, Flor. : Chapman & Hall/CRC.

Oltsik Jon, (2015): "*The Internet of Things: ACISO and network security perspective*", ESG (Enterprise strategy group) White Paper of Senior Principal Analyst.

http://www.cisco.com/web/strategy/docs/energy/network-security-perspective.pdf (Accessed: 08/08/2015).

Pollitt, M. (1995) "*Computer forensics: an approach to evidence in cyberspace*," National Information Systems Security Conference, vol. II, pp. 487-491, 1995.

Petit, J and Shladover S. (2015). "*Potential Cyberattacks on Automated Vehicles*". IEEE Transactions on Intelligent Transportation Systems. Vol. 16 (no. 2), pp.546-556

Qi, J., Yang, P., Hanneghan, M., Fan, D., Deng, Z. and Dong, F. {2016) "*Ellipse fitting model for improving the effectiveness of life-logging physical activity measures in an Internet of Things environment*". IET Networks, 5(5), pp.107.

Reilly D. Wren C. and Berry T. (2011), "*Cloud Computing: pros and cons for computer forensic investigations*", International Journal Multimedia and Image Processing (IJMIP), vol. 1, no. 1, pp. 26–34, March 2011.

Rezendes Christopher J. and Stephenson W. David: (2013), "*Cyber security in the Internet of Things*" https://hbr.org/2013/06/cyber-security-in-the-internet/ (Accessed: 08/08/2015).

Ruan K. Carthy J. Kechadi T. and Crosbie M. (2011), "*Cloud forensics: an overview*", in proceedings of the 7th IFIP International Conference on Digital Forensics, 2011.

Ruan K. Carthy J. Kechadi T. Crosbie M. & Baggli I. (2013), "*Cloud* forensics definition and critical criteria for cloud *forensics capacity: an overview of survey results*", Digital Investigation Vol. 10, Elsevier, 2013.

Sant P. (2015), "*International PhD students learn about MK:Smart*", MK Smart and University of Bedfordshire, http://www.mksmart.org/enterprise/. (Accessed: 08/08/2015).

Sremack Joe (2015) "*Big Data Forensics - Learning Hadoop Investigations*", Packt Publishing ISBN-10: 1785288105, ISBN-13: 978-1785288104 – 24 Aug 2015

Taylor M. Haggerty J. Gresty D. and Hegarty R.(2010), "*Digital evidence in cloud computing systems*," Computer Law and Security Review, vol. 26, no. 3, pp. 304–308,.

Uma Mohan and Saminu Salisu (2015) "*The use of big data in the field of digital forensics investigations*", International Journal of New Technologies in Science and Engineering, Vol. 2, Issue. 4, October 2015, ISSN 2349-0780 Available online http://www.ijntse.com 291

Zawoad Shams and Ragib H. (2013) "C*loud forensics: a meta-study of challenges, approaches, and open problems*", University of Alabama at Birmingham, USA.