A Study Towards Post Hoc Forensic Analysis Using Big Data Analytics

Bipin Bihari Jayasingh
Professor, CVR College of Engineering/ IT Department, Hyderabad, India.
Email: bbjayasingh9@rediffmail.com

Abstract: **The network traffic data of an enterprise is considered to be huge day by day and it gradually becomes big data. The major concern here is how to analyze the data in case of an unusual event occurs and how to draw a conclusion when data is voluminous. The post mortem analysis of traffic data and grabbing the information as evidence is said to be network forensics that can be achieved through big data analytics only. In this paper, the network forensics literature is studied along with how to use big data analytics for accurate analysis of fraud. There is a well understood discussion about the security challenges of big data including big data privacy issues, data provenance problems and visual analytics. Traffic data is used for attack analysis by considering fraud detection methods for the advanced persistent threats in order to correlate large quantities of diverse data to detect an attacker. It is also considered the forensic analysis of an attack in the traffic data to extract the evidence against an attacker. It is more focused on the techniques and associated tools for use of big data analytics for security and use of information security algorithms to protect big data. This rigorous study leads to making use of available tools and techniques of big data analytics in the development of any network forensic system.**

Index Terms**: Network Forensics, Big Data Analytics, Fraud Detection, Advanced Persistent Threats, Privacy Issues, Data Provenance Problems, Visual Analytics.**

I. Introduction

In the cyber environment the data becomes more than the size of petabyte large every day due to network flows (messages, streams, and user actions), user passwords and other information. The fundamental aspect is to analyze the data in order to mitigate the system vulnerabilities and attacks. Big data visualization and analytic techniques are useful to analyze the huge network data available in the storage hub to address cyber security [2, 11]. An other way of analyzing network data after an attack is very important to extract the evidence against the crime which must be through big data analytics tools and visualization [3, 4]. The analysis after the attack is towards a post hoc forensic analysis by using "big data analytics" in the storage hub of an enterprise.

Forensics is an art of discovering and extracting information about a crime suitable to law enforcement agencies as admissible evidence [12]. In the traditional way of expressing, the prevention and detection of network attacks is the network security model that extends to network forensics. Network forensics is the use of scientifically proved techniques to collect and analyze network packets and events for investigative purposes [7]. Network forensics is a branch of digital forensics that focuses on monitoring, capturing, recording and analysis of network traffic includes volatile and dynamic data [7].

However, forensically reconstructing a network attack along with evidence is called network forensic system. The intention behind reconstruction is to prove that the network attack that has been perpetrated by the invader, for whom the evidence is collected. Though the digital evidence is fragile, it must be preserved for future trial. In the internetworking environment, it is difficult to preserve the evidence in the same system as the network traffic is huge and might crash the traffic capture system if left unattended. Hence, the more the network traffic, the harder the network analyzing. Therefore, we need an effective and automated analyzing system for network forensics where data is big.

The remainder of the paper is organized as follows: section 2 specifically discusses the challenges of big data analytics in terms of big data privacy issues, data provenance problems and visual analytics; section 3 focuses more on attack analysis methods considering fraud detection, advanced persistent threats and forensic analysis; section 4 present the use of big data analytics for security and use of security to protect big data with forensic profiles techniques; section 5 provides the concluding remarks.

II. Challenges Of Big Data Analytics

The main challenge is to identify the network attack followed by evidence reconstruction. The information present in the packet header may be crucial evidence for the detection of network level attack. Other way of detecting the network attack is looking at the data fragments carried out by the packet payloads [8]. The detection using big data analytics has a significant promise that must be addressed by the forensic specialist. Data analytics must focus on network traffic and separate the abnormal traffic from the normal flow of traffic. It becomes easy to identify the abnormal pattern in the abnormal traffic

 doi:10.32377/cvrjst1413

to find the suspected activity. The abnormal pattern may become potential legal evidence of the attack which a network forensic system requires. Another valuable phenomenon called the network forensics, is an investigation process for the network related crimes. In specific, network forensics investigates the problems created by the miscreants in accessing the network devices. It analyzes large amount of network traffic data to find the footprints of network crimes. We consider the big data privacy issues, data provenance problems and visual analytics in this section.

A. Privacy Issues

The privacy issues are viewed in a wider perspective when it relates to data mining and considers various approaches to protect sensitive information [19]. A new knowledge can be extracted through data analytics and aggregated through data processing but not lead to privacy issues of a user [14]. Data analytics has to preserve the privacy of data while extracting and correlating the data. The forensic analysis must ensure to minimize the privacy raids during analysis of data. The data can be used for the purpose for which it has been collected.

Data analytics helps to efficiently process and organize the relevant data. The forensic scientist must examine the data without encroaching the privacy of a person. The system must collect enough metadata consisting of different attributes, to confirm the identity of a person. For example, a facial analysis system should not attempt to identify a person but it must collect enough metadata about a person based on their face like the expressions of a face. The forensic analysis must respect the privacy of data and ensure the security and privacy policies. Any application tools we develop must be allied with principles and recommendations of privacy preservations [2].

B. Data Provenance Problem

Data provenance in big data has been explored recently but it has been studied in database and distributed systems communitie's. It simply refers to the ownership of data or custodian of data and the location of the data where it resides [19]. It is a process of maintaining the origin and creation of data which is useful for validating the data, determining the reliability of data and evaluating the quality of data. The quality of data depends on debugging, transformations, auditing of data objects of various sources [20]. Data provenance can also be used for performance bottleneck analysis.

Big data provenance is another challenge where it allows for expanding data sources and the data can be from any sources. However, the analysts must put proactive effort to secure the abuse of the data sources [2]. The analysts must focus the issues of data provenance discussed hereunder.

- There are various sources of data but access to all sources at a right time is difficult.
- There may be possible access of all data but data analysis probably becomes difficult.
- Generating reports using queries on each source takes longer time to run.
- It is very difficult to share the gained information and useful patterns with others.
- The trustworthiness of data source is hard to maintain to produce accurate results.

There must be some mechanism to identify and mitigate the malicious data in the data source itself like machine learning algorithms. So research is required for innovative methodologies in order to visualize and mine the data provenance [5].

C. Visual Analytics

Visual representations of data is needed for the analysts to interact with data to amplify cognition. This is called information visualization. The visualized information presents graphical models and user interfaces which interactively manipulates the large set of data [12]. The graphical models can transform raw data into proper tables, then turn into visual structure. Visualization of information always considers with small set of data and for the large data set, it will be called as visual analytics.

Human eye has a tendency to look for the summaries or overviews instead of large amounts of unfamiliar data. Analysis of data and getting a decision is a process that makes a sense of the system. This is also referred as visual analytics. The analysis looks for patterns and correlations at a different level of drill down to the data [2]. The aim is to represent the data effectively to convey information to the analyst is called visual analytics [6]. It explores and interacts with data by applying queries and finds the results and interpret the results. It helps the analyst to interpret the results and convey the useful information through effective representation [2]. The simplest example of visual analytics is discovering fraudulent transactions in credit card usage.

III. ATTACK ANALYSIS METHODS

Every day, 2.5 quintillion bytes of data [21] are created and 10 to 100 billion events are generated by a large enterprise [2]. As the events are generated from multiple and heterogeneous sources, the enterprise must deploy more devices, hire more employees and run more software for post hoc forensic analysis. As a result, the overwhelming data volume requires an effective data analysis and prediction platform to achieve the efficacy of such big data. Existing analytical techniques are not sufficient in the large-scale analysis and in the processing of big data events, so the big data analytics has attracted the interest of the security community [6]. There are three ways of doing the attack analysis either through fraud

detection methods, advanced persistent threats or through forensic analysis of the networks.

A.Fraud Detection

The law enforcement agencies attempt to detect fraud becomes harder still due to various factors that lead to fraud. It becomes easier through 'big data analytics' tool which help to anticipate and quickly detect fraud to take immediate action [18]. In order to detect fraud, we must use data mining tools that can search and find patterns on behalf of the analyst. Some such data mining tools are recorded here for the better understanding of the researchers as follows:

- Decision trees
    - Boosting trees
    - Classification trees
    - CHAID
    - Random Forests
- Machine Learning
- Association Rules
- Cluster Analysis
- Neural Networks

There are predictive models which estimate the probability of amount of fraud and focus most efficiently to prevent or recuperate fraud losses.There are some issues which law enforcement agencies must look.

- Fraud detection is a predictive modelling problem that can anticipate a rare event. The model can capture the fraud if the historical data is available along with previously classified verified data. The suggestion for the rare events can be predicted with a best predictor and a validated model that provide a greatest lift to maximize the likelihood of observations associated with fraud.
- Fraud detection is an anomaly detection problem that identifies outliers in multivariate space. This works well even when there is no good training dataset where known fraudulent and non-fraudulent observations are clearly identified. Using Big Data techniques, to analyze and even predict, security incidents [13] are successful.

B. Advanced Persistent Threat (APT)

The advanced persistent threat (APT) is a new breed of insidious threats where hackers can retain control over target systems unnoticed for long periods of time. The adversaries use multiple attack techniques and vectors in order to avoid detection [4]. The traditional defense mechanisms are not sufficient to detect such attacks. The focus should be on developing a depth strategy of defense only after a clear understanding of these attacks. The aim must be constantly to monitor the networks and security control systems to grab possible footsteps of such attacks for post hoc forensic analysis.

APTs occur in such a way where a victim stays oblivious to the happening of insidious crime. It operates itself while it resides in the system and executes for an extended period of time. The detection of such threats requires data from various internal and external data sources to correlate to find posterior information, lead to forensics [2]. In such cases big data tools are useful though it correlates huge amounts of diverse data and performs long term correlation that becomes fundamental for advanced persistent threat (APT) detection and forensics [4, 5]. The techniques may be a pattern analysis or feature extraction that may detect the threat at an early stage.

C. Forensic Analysis of Networks

The analysis of network logs, network flows and system events for an enterprise in order to detect an attack is a major problem in the security and forensic community due to inadequate analytics technology [15]. There are several reasons in practice where voluminous data storage was economically not feasible and retention period for variety of unstructured data is fixed. The traditional management of large data warehouses has become expensive and obsolete. However, the big data tools and techniques along with Hadoop deployment frameworks are suitable for the large scale maintenance of data that are most preferred to process and analyze the data [2].

There are applications that use big data technologies to clean, prepare and query data in heterogeneous, incomplete and noisy formats efficiently. There is also security management software that helps the analyst to do so called clean, prepare and query the data.

There is proposed literature based on Shannon entropy and machine learning techniques for attack detection in network forensics systems [8]. There is also a proposed fuzzy logic based expert system for forensic analysis of computer networks that can analyze the threat and store the digital evidence automatically [9]. There is a possibility of compromise of the system to the network attack. Now, finding the evidence against the attack needs the help of forensic experts where it may lose some useful instant evidence at the time of investigation. So, there must be integrated analysis of the log and audit system and network traffic that can make to an efficient navigation of the traffic. This work discusses the frameworks of distributed agent-based real-time network intrusion forensics system which is deployed in local area network environment [10].

The network forensics objective is to gather evidence of criminal acts in the network infrastructure includes about the perpetrator from any place of the globe. Therefore, software tools, and techniques, that can help with these digital investigations are in great demand. The self-organizing maps (SOM) are presented to analyzing and visualizing network traffic data [11]. The self-organizing map has been widely used in clustering tasks that can enable network clusters to be created and visualized in a manner that makes them immediately more intuitive and understandable.

There is an architecture called trusted Internet forensics (TIF) that collects data from the network for forensics purpose. The architecture consists of network appliances that rely on a trusted computing platform. The architecture allows for the verification of the computational chain so that the data collected in the network could be used as evidence in court [12].

It is more difficult to collect the information for attack analysis though the attackers are trying to remove attack traces such as system logs and related information on the victim systems. Therefore, most of them are focusing on gathering the network packets. There is a network forensics system, Cyber Black box, which is focused on the traffic analysis [13]. Recently, several machine learning techniques have been proposed to automate and develop intelligent network forensics systems. Others adapt recent researches in semantic-web, information architecture, and ontology engineering to design method ontology for network forensics analysis [14].

IV. Security Analysis

Traditional computing methods of security tools and techniques used to process the big data are inadequate due to the involvement of data from different machines and applications. To mine the traffic data and prevent cyber security must be more reactive than proactive. It also creates a large number of false positives, inefficient and distracting from actual threats. Data analytics has to be used to analyze the network traffic, log files and financial transactions. It may be feasible that the data correlations from multiple information sources are represented in a coherent view and possibly identify suspicious activities and anomalies. Data analytics can predict potential cyber security breaches that help to stop cyber-attacks and facilitate post breach cyber forensic analysis.

Security researchers must keep exploring novel ways to find sophisticated attackers though bigdata is not a panacea. The technologies like network monitoring and network forensics become the landscape of security. Keeping in mind the privacy preservation, all must work together to develop some tools for the data analytics [2]. So far, the dominant strategy has been to optimize existing analytical environments that often rely on scripting languages like Python, Spark and R. Therefore, it needs to increase the effort to educate new generation computer scientists and engineers on the value of privacy.

We can use other popular algorithms which are typically available in statistics and machine learning libraries. The typical algorithms include hierarchical clustering and principal component analysis computation and projection methods. We concentrate on using big data analytics for security and forensics, also on protecting the big data through security mechanism by applying suitable advanced data mining algorithms.

A. Big Data Analytics for Security

The data processing using traditional data management tools and techniques to massive, heterogeneous and often unstructured digital content is difficult. The big data refers to the complexity and variety of data and data types whereas real time data collection and processing can be obtained by smart data analytics. Big data analytics in the trade sector provides a better understanding of customer behavior and preferences. It extracts a valuable insight that enables to increase customer satisfaction [3]. In many business and scientific applications, the use of standard advanced data mining tools and techniques that helps to extract information from large and complex datasets to make proper decisions.

The main focus is to analyze the data and find the loopholes of system security then improve the security mechanism to protect the misuse of the data. As the data analytics tools continue to be deployed in enterprise systems, it needs to improve the systems security by introducing new tools, such as Apache's Accumulo, to deal with the unique security problems in big data management [2]. The conventional security mechanisms like integrating Transport Layer Security within Hadoop are not sufficient to justify the security levels. In particular, a better monitoring of macro and micro level data security is essential which help decision makers to catch commercial opportunities i.e. anticipating recessions [4].

B. Security to Protect Big Data

In a security aspect, there are two distinct issues: securing the organization and its customers' information in a Big Data context that means protecting the data as well as data centers. Without the right security and encryption solution in place, however, big data can mean big problems. Big Data breaches will be big too, with the potential for even more serious reputational damage and legal repercussions. Techniques such as attribute based encryption may be necessary to protect sensitive data and apply access controls.

Physically data resides at different locations in a distributed manner when it becomes large volumes using big data technologies. The users have never possessed storage of their data physically. In order to perform data mining, some users hire third party data miners to process their data efficiently and effectively. In the realm of big data, protection of private data focuses on excluding third party access to the original data directly. The solutions rely on some privacy preserving approaches or encryption mechanisms to the source of data. Sometimes data accesses from servers by using virtual disks are avoided due to compromising the user privacy.

To secure the data in the big data store, there are vendors who provide encryption capabilities. It may not secure all the log files and configuration information associated with the big data environment. But the security specialist must

contend for key fragmentation that may helpful for administration of big data stores [5].

One of the solutions for data security is to use vormetric encryption schemes to the diverse data store of an organization. The encryption mechanism follows in two ways:

- Vormetric Transparent Encryption**:** These encryption mechanisms directly handle the file systems to encrypt the data and also control the access to the file system itself. It is very easy to deploy and it will not effect to any changes that are made to any kind of applications.
- Vormetric Application Encryption**:** These encryption mechanisms are encrypting the attributes of the database before it store permanently. It is encrypting the columns of the specific applications before writes to the database. There will be some sensitive field that will remain unreadable due to the column encryption scheme. Also, the same problem remains even after it is imported into, and processed within, the big data environment.

C. Forensic profiles

Data mining techniques are used to discover relevant patterns from large quantities of data. Using those patterns generating forensic profiles is good approach in the big data environment. The forensic profile can be matched to any new data coming to the data source to find the misused data. The evidence can be extracted in the case of misuse or anomaly of data. There are literatures available in the area of extracting and analyzing digital evidence from physical devices such as hard disks. Probably less effort put in the portable storage devices [15]. The forensics on portable devices using data mining is difficult due to the structure and the storage of the data.

There are good number of work that has been done to incorporate the use of decision trees as well as naive Bayesian, a priori, and neural network techniques [16, 19]. Also there are similarity-based machine learning techniques deliver more robust performance than other techniques such as decision trees, ensemble methods and regression-based methods that are to achieve practical efficiency for similarity-based techniques is to sparsely the similarity matrix [1]. Recently proposed architectures also incorporate mechanisms for monitoring process behavior, analyzing trends, and optimizing plant performance [17].

In the data mining perspective, there is a proposed model called big data processing model that is related to HACE theorem. The work characterizes the features of big data revolution that starts with large-volume, heterogeneous, autonomous sources with distributed and centralized control [18]. Data mining algorithms has to perform in such a way that privacy information must be preserved and security of sensitive data must not be compromised [19].

We have proposed a Network Forensic System (NFS) for network related crimes that helps the investigator in finding the attackers associated with crime. The proposed forensic system will combine the best features of existing traditionally framework models and extend them to form a focused investigatory model for networked systems. The forensic system will be classified into three phases shown in fig. 1.

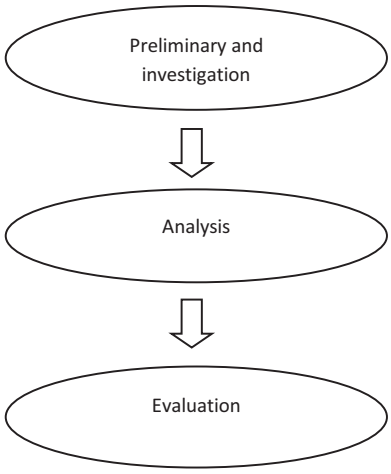

Figure 1. Model of the network forensic system

- Preliminary and investigation

The preliminary process includes the victim of the crime, cyber forensic experts and law enforcement agency. The investigation process will start after the victim's complaint against the crime to the law enforcement agency about the incident that has occurred. The law enforcement deputes the cyber forensic expert to find sufficient evidence against the crime and submit a brief report. This process mainly concerns on validation of the incident and followed by checking requirements for the investigation like human resources, etc. We combined the preliminary and investigation phase that saves the time and money. In this phase, the cyber forensic expert will identify, collect and preserve potential digital evidence that are required to proceed with the investigation. In this process the cyber forensic expert try to find the relation between the culprits and the crime. Any further evidence found will be collected and produced under the court of law or as part of any other relevant legal processes. Once investigation phase is finished, the relevant information will be stored as evidence and will be reported to next step i.e., Analysis phase.

- Analysis

The Analysis phase is the most important part of the model as the investigator needs to verify all the information found is connected to the incident so that it serves as potential legal evidence. The analysis and examination of evidence determines the width and depth of the crime.

Analyzing the suspect's information recorded in the service provider's database is the digital forensic investigation. It is useful to map the relations of the suspects profile with the profiles of other people to obtain constructive evidence. This phase determines if the evidence is genuine or if there is generated evidence to prove the crime.

- Evaluation

In this phase, the investigator will produce all the evidence that are collected about the culprit from the network related crime. This is through appropriate documentation as a report. The investigator needs to present the legal evidence to the law enforcement agency against the crime to fulfill the aim of the investigation.

V. CONCLUSIONS

The study undertakes more emphasis on advanced data mining tools and techniques relating with big data analytics. Though all the information captured or recorded are not useful for analysis, it will be decided by big data analytics. A network forensic system must be developed which can provide an analyzed information to the forensic experts and reduce the time and cost of forensic analysis in case of network attacks. There is a proposed network forensic system in general for the investigators to understand the judiciary process of cybercrime. In the future the analysis of network traffic will be carried out for an attack to show sufficient information as evidence against an unusual event using big data analytics.

REFERENCE

[1] Dorit S. Hochbaum, Philipp Baumann, Sparse Computation for Large-Scale Data Mining, IEEE Transactions on Big Data, Vol. 2, No. 2, April-June 2016, pp. 151-174.

[2] Alvaro A. Cárdenas, Pratyusa K. Manadhata, Sreeranga P. Rajan, Big Data Analytics for Security, IEEE Security & Privacy, 1540-7993/13, 2013 IEEE, pp.74-76.

[3] A. Katal, M. Wazid, R. Goudar. "Big data Issues, challenges, tools and good practices", in the Sixth International Conference on Contemporary Computing, Aug. 2013, pp. 404-409.

[4] D. F. Nettleton, "Commercial Data Mining: Processing Analysis and Modeling for Predictive Analytics Projects", 1st ed. Boston, United states, Morgan Kauffman Publishers-Elsevier, 2014.

[5] Domenico Talia, "Clouds for Scalable Big Data Analytics", IEEE Computer Society, 0018-9162/13, 2013, pp.98-101.

[6] R. Nambiar, R. Bharadwaj, A. Sethi and R. Vargheese."A Look At Challenge And Opportunities In Big Data Analytics In Health Care", In IEEE International Conference In Big Data, Oct. 2013, pp. 17-22.

[7] J. Buric, INsig2 d.o.o., Zagreb, Croatia, D. Delija, Challenges in Network Forensics, IEEE 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), DOI: 10.1109/MIPRO.2015.7160490, 25-29 May 2015, pp. 1382 – 1386.

[8] Khoa Nguyen, Dat Tran , Wanli Ma, Dharmendra Sharma, An Approach to Detect Network Attacks Applied for Network Forensics, IEEE 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), DOI:10.1109/FSKD.2014.6980912 19-21 Aug. 2014, pp. 655 – 660.

[9] Jung-Sun Kim, Chonnam Nat., Dong-Geun Kim, Bong-Nam Noh, A Fuzzy Logic Based Expert System as a Network Forensics, IEEE International Conference on Fuzzy Systems DOI: 10.1109/FUZZY.2004.1375521, Vol. 2, 25-29 July 2004, pp. 879 – 884.

[10] Wei Ren, Wuhan, China, Hai Jin, Distributed Agent-Based Real Time Network Intrusion Forensics System Architecture Design, 19th IEEE International Conference on Advanced Information Networking and Applications (AINA'05), DOI. 10.1109/AINA.2005.164, vol. 1, 28-30 March 2005, pp. 177 – 182.

[11] E. J. Palomo, J. North; D. Elizondo; R. M. Luque, VisualizationOf Network Forensics Traffic Data With A Self-Organizing Map For Qualitative Features, IEEE International Joint Conference on Neural Networks (IJCNN), The 2011, DOI: 10.1109/IJCNN.2011.6033434, July 31 -Aug. 5 2011, pp. 1740 – 1747.

[12] D. Bruschi, M. Monga; E. Rosti, Trusted Internet Forensics: Design of a Network Forensics Appliance, IEEE Workshop of the 1st International Conference on Security and Privacy for Emerging Areas in Communication Networks, 2005, DOI: 10.1109/SECCMW.2005.1588292, 5th -9th Sept. 2005, pp. 33 – 35.

[13] Yangseo Choi, Joo-Young Lee;Sunoh Choi ; Jong-Hyun Kim, Introduction to a Network Forensics System For Cyber Incidents Analysis,18th IEEE International Conference on Advanced Communication Technology(ICACT),DOI: 1109/ICACT.2016.7423270 Jan. 31 2016-Feb. 3 2016, pp. 50 – 55

[14] SherifSaad, IssaTraore, Method Ontology for Intelligent Network Forensics Analysis, Eighth IEEE Annual International Conference on Privacy Security and Trust (PST), DOI: 10.1109/PST.2010.5593235, 17-19 Aug 2010, pp. 7 – 14.

[15] V.H. Bhat, "A Novel Data Generation Approach for Digital Forensic Application in Data Mining," Proc. 2nd Int'l Conf. on Machine Learning and Computing (ICMLC 10) IEEE, 2010, pp. 86-90.

[16] F. Camastra, A. Ciaramella, and A. Staiano, "Machine Learning and Soft Computing for ICT Security: An Overview of Current Trends," J. Ambient Intelligence and Humanized Computing, Oct. 2011; doi:10.1007/s12652-011-0073-z.

[17] T. Kilpatrick et al., "An Architecture for SCADA Network Forensics," Proc. IFIP Int'l Conf. Digital Forensics (IFIP 06), Nat'l Center for Forensic Science, 2006, pp. 273-285.

[18] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, Data Mining with Big Data, IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. 1, January 2014, pp. 97-107.

[19] Lei Xu, Chunxiao Jiang, Jian Wang, Jian Yuan,Yong Ren, Information Security In Big Data:Privacy And Data Mining, DOI 10.1109/Access.2014.2362522, IEEE ACCESS, October 20, 2014, Pp. 1149-1176.

[20] Y. L. Simmhan, B. Plale, and D. Gannon, A survey of data provenancein e-science," ACM Sigmod Rec., vol. 34, no. 3, 2005, pp. 31-36.
[21] "IBM What Is Big Data: Bring Big Data to the Enterprise,"http://www-1.ibm.com/software/data/bigdata/, IBM, 2012.