Journal of Information Assurance and Security.
ISSN 1554-1010 Volume 11 (2017) pp. 124-145

Big Data Analytics Adoption for Cyber-security: A Review of Current Solutions, Requirements, Challenges and Trends

Murad A. Rassam[1,*], Mohd. Aizaini Maarof [2] and Anazida Zainal [2]

[1] Faculty of Engineering and Information Technology, Taiz University,
6803, Taiz, Yemen
murad.utm@gmail.com, corresponding author

[2]Information Assurance and Security Research Group, Faculty of Computing, Universiti Teknologi Malaysia,
81310, Skudai, Johor, Malaysia
aizaini&anazida@utm.my

***Abstract*:** With the continuous advancements of technology, cyber-criminals accordingly develop sophisticated tactics to exploit vulnerabilities in individual systems, organization networks, and nation-states. Enterprises routinely collect huge amount of security-relevant data such as log events of people, networks, and software applications for future forensic analysis. Existing traditional security analysis tools fail to work well with large scales of data and usually produces high false alarms especially when the enterprises moves to cloud architecture and collect more data. Moreover, the detection of recent and more sophisticated attacks, like advanced persistent threats (APTs), requires continuous monitoring and analysis of huge security related data, accurately and rapidly. Big Data analytics has been in active use in several fields such as financial transactions, healthcare and industrial applications among others. Recently, it has attracted the attention of information security audience due to its promised ability in correlating security related data and draw insights efficiently at unprecedented scale. In this paper, we analyze the traditional technology/systems and Security Information and Event Management (SIEM) tools and show their shortcomings in dealing with huge data scales and advanced sophisticated threats. We then explore the requirements for Big Data analytics to be successfully adopted in cyber threat intelligence and cyber-security landscape to deal with high data scales and sophisticated threats. Finally, we highlight the challenges resulted from such adoption, and suggest some recommendations to overcome adoption challenges in future research.

***Keywords*:** Big Data, Big Data Analytics, Cyber-security, Threat Intelligence, Security Information and Event Management (SIEM).

I. Introduction

The term cyber-security is defined by Merriam-Webster dictionary as "measures taken to protect a computer or computer system (as on the Internet) against unauthorized access or attack". Another comprehensive definition for cyber-security was given by the US national initiative for cyber-security careers and studies (NICCS) as "Strategy, policy, and standards regarding the security of and operations in cyberspace, and encompass[ing] the full range of threat reduction, vulnerability reduction, deterrence, international engagement, incident response, resiliency, and recovery policies and activities, including computer network operations, information assurance, law enforcement, diplomacy, military, and intelligence missions as they relate to the security and stability of the global information and communications infrastructure" [1]. According to the latest statistics revealed in Kaspersky final statistic report in 2015, the registered notifications about malware infections that tried to hack online bank accounts access to steal money were 1,966,324 [2]. This huge number of malware infections only in one economic sector shows how big is the danger on the overall economy of the world every year. It is therefore clear that the cyber-security of IT systems, organization networks and web applications is possibly inadequate to cope with the rapid evolve of cyberattacks. As a result, a question that should be asked, what must be done for better protection and detection of such cyberattacks' rapid growth?

In 2015, International Data Corporation (IDC) conducted an interview for security specialists, executives, and industry experts in three different industries, which are the US federal government, energy, and financial services. The aim of the interview was to understand the evolution of cyber threats landscape. The interviews concluded that cyber-security threats are growing rapidly and that organizations should move from a reactive approach of security to a proactive approach that involves understanding the risks before attacks can cause damage [3]. According to [4], large enterprises generate up to 100 billion events per day, based on their size. The increase of events number is affected by the increase of data sources, number of employees, deployment of new devices, or the run of additional software applications. Existing traditional analytical techniques used for cyber-security such as log events analysis, intrusion detection systems and others are inadequate and do not work well at large scales and typically produce high false alarms. The case worsen, as the enterprise adopts cloud architectures and collects more data [4]. Therefore, cyber threat intelligence is required so that a continuous real time collection and monitoring of data streams is guaranteed and therefore a suitable risk mitigation process is launched before attacks can cause severe damages.

Security Information and event management (SIEM) is defined as a platform for collecting and correlating security events, logs, and network flow data for the purpose of security analysis and operations [5]. Cyber threat intelligence integrates functions that have been used in SIEM solutions, including log management, security event correlation and network activity monitoring [6].

According to [5], most of organizations adopt SIEM for monitoring threats rather than security traditional analysis and investigation. However, SIEM has many shortcomings that make it inadequate enough for coping with the huge evolve of threats nowadays. These shortcomings according to [5] include the following: *First*, event correlation in SIEM relies on data normalized in relation to predefined schemas, therefore, it becomes difficult to adapt with new attacks that use multidimensional tactics and differ from system to system; *Second*: platforms that adopt SIEM rely on fixed storage (Schema), therefore, it becomes difficult to cope with the continuous growth of events generated through the time; *Third*: SIEMs are based on predefined context, so they are context specific and cannot be generalized to different situations unless they are defined again, which is a time consuming process; *Fourth*: SIEMs are inflexible such that adding new rules required the reconstruction of the whole approach. As a conclusion, authors in [5] concluded that enterprises need a new approach for cyber-security such that it overcomes the aforementioned shortcomings. It is reported that, the new approach can be viewed as an end-to-end relationships between security analytics-technologies based on Big Data and the information security team skills.

Big Data Analytics, a large scale information processing and analysis, has been in active use in different fields. It also attracts the interest of the information security community as it has a promising ability in analyzing and correlating security related data in efficient manner [4]. Authors in [7] pointed out that, unknown and previously unseen cyber-attacks are increasing due to shortcomings of existing security systems. It is stated that threats are advanced from leaking personal information to destruction of services to attacking large-scale systems such as critical infrastructure [7]. It is concluded that most of unknown cyber threats are missed due to that traditional security tools are based on a simple pattern matching and need to be handled in the context of Big Data analytics.

Few Big Data analytics based security solutions were found in literature such as [8-12].

Authors in [8] addressed cyber-security insurance (CI) implementations focusing on cloud-based service offerings and proposed a secure cyber incident analytics framework using Big Data. Authors claimed that their approach was designed for matching different cyber risk scenarios, which uses repository data. Simulation results have provided the theoretical proof of the adoptability and feasibility of the framework according to authors. In [9], authors were motivated by the smart grid which is an emerging technology that can fulfill the demands of renewable energy by incorporating advanced information and communications technology (ICT). As a result, authors stated that the pervasive deployment of the advanced ICT will generate big energy data in terms of volume, velocity, and variety, especially the use of smart metering. As the generated Big Data can bring huge benefits to better energy conservation and planning and efficient energy generation and distribution, new security issues are emerged which involve end users' privacy and secure operation of the critical infrastructure. The generated Big Data can bring huge benefits to better energy planning, efficient energy generation, and distribution. However, such privacy and security issues should be mitigated. Therefore, authors surveyed smart technology for sustainable energy and related Big Data security issues. In [10], authors presented a comprehensive survey on the state-of-art of security analytics in terms of its description, technology, trends, and tools. Authors aimed to inform the reader about the possible application of analytics as an unparalleled cyber-security solution in the near future.

Authors in [11] proposed a framework to protect against advanced persistent threats (APTs). The proposed framework combines different techniques based on Big Data analytics and security intelligence to support human analysts in prioritizing the hosts that are most likely to be compromised.

In [12], authors presented a Cyber Security analytics framework aimed for comprehensive cyber security monitoring by constructing cyber security correlated events with feature selection to anticipate the users' behavior based on various sensors. The proposed Big Data analytics based Cyber Security framework is a Cyber Security Analytics (CSA) architecture based on Network Log (NetL) and in-memory Process Log (PrcL) to identify an anomaly vector with assistance of comprehensive system observations.

The transform from conventional security technologies to Big Data based security technologies to support long term large scale analytics happened due to three main reasons according to [4]: *First*: handling large data quantities was not economically feasible in traditional security based technologies and SIEM; therefore, some event records were usually deleted after periods of time to release storage space for some new events. *Second*: difficulty of performing data analytics and complex queries on large and heterogeneous datasets with noisy and incomplete features. *Third*: it was expensive to manage large data warehouses, as their deployment requires strong business cases.

Therefore, the contribution of this paper is in three folds: *First*: we study the current technologies/systems that use traditional security and SIEM tools to reveal their weaknesses due to the rapid evolve of both cyberattack landscape and huge amount of data events generated by enterprises in a day-by-day process. *Second*: we highlight the requirements for adopting Big Data analytics for cyber-security, so that the weaknesses of using traditional and SIEM based systems are overcame. *Third*: we highlight the difficulties and concerns that arise by adopting Big Data analytics for cyber-security to be considered for future research directions.

The rest of this paper is organized as follows: Section 2 explores cyber-security era and provides necessary information about the current state of cyber threat intelligence landscape. Section 3 explains Big Data concepts and fundamentals, discusses the adoption of this hot technology in different research areas. Section 4 explains in details the requirements for adopting Big Data analytics for cyber-security, navigates the current technology/solutions already exist for such adoption, and highlights the challenges

of this adoption in cyber-security landscape. Section 5 reports some recommendations for future research; while section 6 concludes this paper.

II. Cyber-security and Cyber Threat Intelligence

In this section, a fundamental background about cyber-security and cyber threat intelligence is given to establish a base for discussion in the following subsections. The current traditional cyber-security approaches are also explored and discussed.

A. Definitions

Cyber-security as defined in Section I, is a set of countermeasures, strategies, and standards that are used to prevent, detect, and defend against any type of vulnerabilities against system, organization network, or the Internet in the cyberspace.

Attack: "An attempt to gain illegitimate access to system resources, services or information, or an attempt to compromise system integrity" [1].

Advanced Persistent Threat (APT): "An adversary that possesses sophisticated levels of expertise and significant resources which allow it to create opportunities to achieve its objectives by using multiple attack vectors (e.g., cyber, physical, and deception)" [1].

Threat intelligence as defined in Gartner dictionary is "an evidence-based knowledge, that includes context, techniques, indicators, implications and actionable advice, about an existing or emerging menace or hazard to assets that can be used to inform decisions regarding the subject 's response to that menace or hazard".

Security Information and Event Management (SIEM): is defined by the Enterprise Strategy Group (ESG) as "a platform designed to collect and correlate security events, logs, and network flow data for security analysis and operations" [5].

B. Traditional Cyber-security Analysis and Management Approaches

Since the concept of cyber-security has been in place, different types of security analysis and management approaches and mechanisms have been used to defend and protect IT systems, organization networks, and the Internet from different types of cyber-security threats. According to IDC, these cyber-security threats can be categorized into 10 broad categories as shown in Figure 1. All types of threats that systems and networks face today belong to these categories or variants of them. However, distributed denial of service attacks (DDoS) and advanced persistent threats (APTs) in addition to zero-day attacks constitute the most sophisticated and long term attacks that should be detected at the suitable time and in accurate manner.

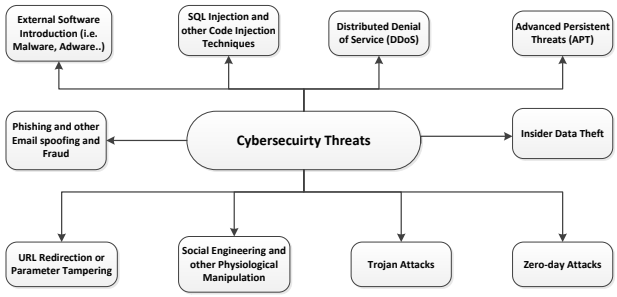

Figure 1: Cyber-security Threats

During the history of information systems or network security, various strategies and methodologies have been proposed and developed to defend and mitigate the effects of cyber-security threats. Figure 2 shows the traditional cyber-security management and analysis approaches and mechanisms that are usually seen and in process in any enterprise or some individual IT systems. A brief description of each approach is given in the subsequent paragraphs. It is necessary to say that these approaches comprise the most used among different approaches that are not the subject of this review.

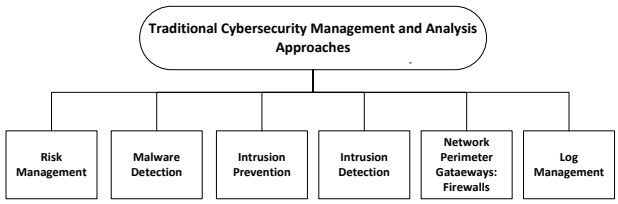

Figure 2: Traditional Cyber-security Management and Analysis Approaches

1) Risk Management

A risk in business was defined in [13] as the possibility of an event that reduce the value of the business. That event is also called an "adverse event". Authors in [13] argued that information security is information risk management as well. In addition, they stated that in order to quantify the risks and the effectiveness of security risk control measures in information security domain, some important information should be collected. These information includes the possible vulnerabilities in the information security system, information related to businesses security incidents worldwide, the direct and indirect loses caused by each incident, and the available countermeasures used to mitigate such kinds of incidents due to the vulnerabilities.

A number of risk management frameworks and approaches in information security literature have been proposed. Recently, an information security risk analysis model was proposed in [14] based on fuzzy decision theory and event tree analysis. The model identifies and evaluates the sequence of events in an incident scenario following the occurrence of potential information technology system abuse. The study in [15] proposed a hybrid information security risk assessment methodology based on the use of both quantitative and qualitative risk management approaches. The authors compared advantages and disadvantages of both analysis approaches and concluded that neither approach can achieve the best performance alone. Therefore, a hybrid approach that utilizes accurate

decision of quantitative approach can be utilized together with the qualitative approach that is based on judgments and intuitions.

Authors in [16] proposed an improved evidence theory based Information Systems Security (ISS) risk assessment framework. Authors claimed that using evidence theory can effectively model the uncertainty in the assessment process. Furthermore, a new way to define the basic belief assignment through a fuzzy measure was provided by this model, so that it can deal with fuzzy evidence in the ISS risk assessment process. Another Security Risk Analysis Model (SRAM) was proposed in [17] based on Bayesian networks and ant colony optimization algorithm. This model reduces the probabilities of occurrence and consequence of risks. After that it calculates the propagation paths of the vulnerability based on ant colony optimization algorithm to guide for developing new plans for security risk treatment. However, the treatment of uncertainty issue was not considered by SRAM model.

The above described security risk management frameworks were only given as examples of the use of this approach to mitigate cyber-security threat incidents. Many more risk management and assessment models can be found in the literature. However, we conclude the discussion on this part by presenting the five critical attributes of effective cyber-security risk management suggested in [18] as follows.

First: an effective framework that assists to design confidentiality, availability, and integrity of the information system; *Second*: it should be comprehensive in scope so that it includes all critical elements that need to be protected in the organization; *Third*: it should have a thorough risk assessment and threat modeling; *Fourth*: its incident response planning should be proactive; and finally *Fifth*: there should be sufficient resources dedicated to the effort of security risk management. According to [19], most traditional security risk management frameworks starts identifying information assets, and then identify and evaluate the potential risks with regards to those assets. This implies that this approach is not the suitable choice to mitigate real time security breaches that require online mitigation and quick response. Moreover, the problem can be further increased when the scale of information to be collected is surprisingly increased by the time.

2) Malware Detection

Malware is referred to by various names such as malicious software, malicious code and malcode. Malware have been given numerous definitions. Authors in [20] define a malware instance as a computer program with malevolent goal. McGraw and Morrisett in [21] define malware as "any code added, changed, or removed from a software system in order to intentionally cause harm or subvert the intended function of the system." In this paper, we use the malware definition given in [22] that defines malware as a term that incorporates viruses, Trojans, spywares, adware, ransomware, and other intrusive codes. Malware detection approaches can be broadly categorized into two main categories: signature-based and anomaly based. Signature-based detection depends on predefined signatures to decide the maliciousness of a suspected program. However, anomaly-based detection decides the maliciousness of a suspected program based on its prior knowledge of what establish a normal behavior. A branch of anomaly-based detection called specification-based or rule-based detection. This branch builds some specifications or set of rules to represent the normal valid behavior and uses such rules to decide the maliciousness of a suspected program. Any program violates any of these rules is usually considered malicious.

Numerous malware detection techniques have been proposed in the literature. However, most of existing used commercial tools are signature based techniques, because the anomaly based techniques exhibit high false alarms. Signature-based method is widely used in anti-malware industry for malware detection [23]. However, this approach usually fails to detect variants of known malware or new unseen malwares. That is because in signature extraction and generation process, signatures can be easily avoided [24]. According to [25], some techniques such as polymorphism, and metamorphism, may be used by malware developers to avoid the well-known signature-based detection. Commercial tools that are in use for malware detection includes antiviruses, antispyware, anti-adware, anti-Trojans, and others. Unfortunately, these signature-based techniques take time to study the signature of emerging threats and design mitigation tools to defend against them. As a consequence, the assets of the enterprise in terms of systems, networks, or even individual assets are compromised and damaged.

3) Intrusion Detection

The term "intrusion" was defined in [26] as an attempt to compromise the confidentiality, integrity and the availability (CIA), or to evade the mechanisms of computer or network security. Intrusion detection is the process of monitoring events in a computer system or organizational network, and analyzing them for signs of intrusions. The intrusion detection system (IDS) is the software or hardware form of a system that incorporates the intrusion detection process [26]. Intrusion detection methodologies are categorized into two main categories: signature-based intrusion detection and anomaly-based intrusion detection. Signature-based detection relies on comparing signatures of known attack patterns against new captured events for identifying possible intrusions. This category is sometimes called knowledge-based detection or misuse detection due to the use of the existing knowledge about specific attacks or vulnerabilities. On the other hand, anomaly-based detection compares the normal reference behavior with the current observed events to identify new significant intrusions or attacks. The normal reference behavior is derived from observing and monitoring the regular system or network activities, hosts, and users for some amount of time. Any deviation from this normal behavior is considered an attack or intrusion that should be further investigated. Intrusion detection systems are also categorized based on their scope of detection into host-based IDS (HIDS), network-based IDS (NIDS), and hybrid-based IDS (HIDS).

HIDS detects intrusions or attacks at the host computer system level by incorporating the information provided by the operating systems in these computers. It is reported in [27] that attackers always leave indications of their

misbehaving activities that facilitate the HIDS track their illegitimate access to the computer system and network resources. HIDS monitors system objects log such as file system objects. For each object, HIDS creates a database that usually stores some attributes such as object permissions, object size, and object modification date. It then creates a checksum using any encryption algorithm such as MD5 for the object contents. This database should be frequently updated according to the emergent of threats every day. Antiviruses, anti-spyware, anti-adware among others, are types of signature-based or misuse intrusion detection for host-based systems. On the other hand several NIDSs have been proposed either based on misuse detection or anomaly detection. A well-known proposed misuse (signature-based) NIDS is called Snort [28]. Snort is an open-source packet sniffer used to analyze the network packets and match their characteristics with those stored in a predefined knowledge base. Anomaly based NIDSs have been an active research area in cyber-security era. Numerous anomaly-based NIDSs have been proposed and developed based on several approaches such as statistical-based, rule-based, machine learning-based, data mining and others. A comprehensive and recent review on intrusion detection methodologies and approaches can be found in [29]. The most noticeable disadvantages of IDS systems lie in the following: signature based IDS systems are unable to detect unknown and unseen variants of intrusions before being added to the signature database. On the other hand, anomaly based IDSs suffer from high false alarms due to many reasons related to the availability of learning the normal behavior and the ability to adapt to any normal changes that may happen to the environment. Detecting attacks in real time and adapting to system changes are general shortcomings that characterize the evolution of IDS systems. Moreover, large scale systems and dealing with unstructured data types, were also among the difficulties that face the success of IDSs in defending enterprises against cyber intrusions.

4) Intrusion Prevention

As discussed above, NIDS passively monitors network traffic and raises alarms when illegitimate traffic is suspected. To go one-step further, network based intrusion prevention systems (NIPS) are designed to prevent the suspected activity detected by NIDS to succeed. The mechanism of NIPS is achieved by placing the NIPS device so that the monitored traffic passes through. According to the approach used in detection either signature-based or anomaly-based, every network packet is checked and only delivered if it matches a predefined signature or anomaly threshold. Illegitimate packets are rejected and an alert is raised to the network or system administrator [30]. The advantage of NIPS over the NIDS relies on its ability to intervene and stop predefined attacks after detecting them. Nonetheless, according to [30], most of NIDS drawbacks and limitations are inherited by NIPS such as dependency on fixed and static attacks signatures, inability to deal with encrypted traffic, and difficulties with high speed networks. Furthermore, the fact that NIPS may discard suspected traffic even though it is not malicious, increases the false alarm rates generated by such system, significantly. As a result, these high false alarms may have critical and significant consequences on the functionality of business or mission critical systems. As with IDS, IPS approaches can be categorized into host-based IPS (HIPS) and network-based IPS (NIPS).

According to [30], HIDS, HIPS, NIDS, or NIPS are effective only if their respected signature databases or anomaly detection thresholds are effective and updated. Otherwise, it can be a point of weakness for both computer systems and organizational networks. A serious disadvantage of HIPS is that it incurs unacceptable high processing overhead and high system resource utilization. However, the most useful advantage of HIPS is that it can inspect encrypted files on each system, so that it can complement the function of the NIPS and NIDS to be effective in detecting encrypted intrusions. In a summary, real time detection, scaling to high volume data streaming, long-term detection, and variety of data sources are among other challenges that affect the efficacy of detecting cyber-attacks.

5) Network Perimeter Gateways (Firewalls)

Network firewalls, as defined in [31], are an important components for preserving a secure environment and are always considered the first line of defense against intrusions. They are responsible to control and limit the access to devices such as computers, networks, and servers. They are usually located in a place between secure environment (i.e. organizational network) and insecure environment (i.e. Internet). Firewalls can be categorized according to [32] into three general classes: packet filters, stateful firewalls, and application layer firewalls. Every class type offers specific security service in the context of a network layer model such as open system interconnection (OSI) or TCP/IP models. Based on their placement in the network or their intended scope, firewalls can also be classified into host-based or network-based [32]. Host firewalls usually protect one computer system. They are implemented as software resides on the computer they are intended to protect. Network firewalls, in contrast, are usually standalone devices that are situated at the network gateway(s) such as the point of Internet connection. Their mission is to protect all the inside computers in the internal network from being attacked from outside networks through the Internet. To maintain efficient network performance, a network firewall should be able to deal with high bandwidth, and process the incoming packets quickly. According to [32], although a network firewall helps administrators to control and mange security through one point, it could be a single failure point. Irrespective of the firewall implementation, location, or design considerations, constant observation by humans is still required. The effectiveness of firewalls depends on their policy that requires a good understanding of the network topology and required services. This policy should be updated according to the dynamic changes in the network topology or the series provided to its customer.

6) Log Management

A log is a record of all events, where each entry in the log contains information about events occurred within an

organization‘s computer system or network. Usually, the organization logs contain events related to security. These logs are generated by different sources including antivirus software, intrusion detection and prevention systems, firewalls, network applications, and the operating systems [33]. The process of log management, is necessary due to the rapid increase of generated events, and incorporates generating, transmitting, storing, and analyzing computer security log data. Enterprises are advised to have routine log analysis in order to identify security breaches, frauds, and policy violation. Firewall systems benefited from log management in order to ensure their continued operations. To prevent and recover form firewall failures, logged events related to firewalls ‘s operational status are necessary [34]. Historical log records are helpful for intrusion detection process such that it can be used to determine how an intrusion might have occurred. Several efforts have been devoted to utilize log analysis for further investigation about external or insider attacks such as [35]. In [35], authors discussed the issue of malicious insider detection that exploits internal organizational web servers. The objective of the study was to investigate the utilization of network monitoring and enterprise log management for insider threat activities detection through standardized tools and a common event expression framework. The research in [35] concluded the following: *First*, the detection of significant insider attacks increases by incorporating more events and devices from the analysis such as from proxy, file and print servers; however, this adds more complexity to the log analysis process. *Second*, the compatibility issue should be considered when combining multiple web server packages. The study further recommended that, to increase the efficiency and accuracy of log analysis and correlation process, a common ground between these multiple web server packages should be found. Actually, the utilization of logs opened up the road towards more security analysis and investigation. Most of security analysis methodologies such as IDS, IPS, Firewalls, and SIEM (next section) benefited from the log data in direct or in indirect way. Therefore, this log data should be mined carefully to get insights about what happened in the past or what will be happened in the future.

7) Security Information and Event Management

The term SIEM that stands for security information and event management combines two terms are security information management (SIM) and security event management (SEM). Both terms (SEM and SIM) focus on the collection and analysis of security related data. The emphasize of SEM is on the aggregation of data events into a manageable parts of information in immediate time, whereas the focus of SIM is on the analysis of historical data information for the sake of enhancing the long term effectiveness of information security infrastructure [36]. According to [37], The focus of SIEM is on the following tasks: *First*: the collected data events (sensory data) are normalized in a common format; *Second*: a rapid access should be given to the reported events; *Third*: the scattered reported events from different sources are efficiently analyzed; *Finally*: a correlation of events is performed. The last two tasks are considered the key success factors for SIEM that involve the analysis and correlation processes of security related information and events to extract insights about security incidents.

Some examples of SIEM tools and products that are already in use for cyber threat intelligence and analysis can be briefed in the following paragraphs.

a) ArcSight

Hewlett Packard Enterprise (HPE) ‘s ArcSight SIEM solution [38] is a comprehensive platform designed for threat detection and compliance management of events. It is claimed that the ArcSight architecture is flexible, so that it allows enterprise utilizing their existing deployments.

b) RSA enVision

The RSA *enVision* platform [39] helps enterprises to reduce the burden in their compliance program with standardized, easy to create reports and alerts. It helps to decrease the necessary time and effort to gather and organize data by automating these tasks. It facilitates providing internal and external auditors with direct access to the reports they need at any time. The most important services that RSA *enVision* can provide for security analysis lie in the following two main tasks: *First*: it alerts whenever any deviations from baseline activities is occurred, and detect any potential cyber breach through multiple different devices; *Second*: it achieves forensic analysis on huge records of log data for security events.

c) NetForensics

NetForensics [40] is defined as a security information management (SIM) solution that was considered as a central point for security information gathered by different network devices. It is reported that this solution has the scalability to manage variety of security countermeasures in order to increase the whole security picture. It provides a rule-based correlation technique that analyzes and correlates security information across different types of devices. It provides a way for customizing alerts and reports, in order to enhance the management of security information flow within enterprises. NetForensics helps to make the policy compliance audit an easy task, by using a common framework for different alerting and reporting services.

d) OSSIM

OSSIM [41] stands for Open Source Security Information and Event Management (SIEM), is an open source SIEM system that provides event collection, normalization and correlation. It was proposed by some security engineers to leverage the lack of available open source security analysis products. The Open Source SIEM (OSSIM) provides one platform that supports many of the security processes such as asset discovery, vulnerability assessment, intrusion detection, behavioral monitoring, and SIEM.

e) LookWise

LookWise SIEM platform [42] provides a way to store events happened in the organization network and allows analyst to perform forensic analysis in order to test an

agreement with regulations and meet audit requirements. LookWise provides a way to correlate the information across distributed systems, and makes use of sense and context about the data in order to enhance the capability to extract useful knowledge necessary to detect cyber threats. It allows a real-time receiving of alerts, monitoring of assets and managing of events. As stated in [42], LookWise facilitates events detection which individual systems cannot cope with, and achieves forensic analysis right after the detection of any cyber threat.

f) LogLogic

LogLogic [43] is an information technology company that produces specialized data security management, compliance reporting, and other information technology products. It developed the first platform product for log management platform to collect and correlate user activity and event data. The products of LogLogic are used by many enterprises for the identifications and alert on policy violations and breaches, cyber intrusions, and insider threats.

Common challenges face the feasibility of current SIEM systems and can be summarized in the following points [44]; *First*: current SIEM systems depend highly on the configurations of the deployment of multiple sensors over the network. Therefore, there is a strong need to combine the knowledge extracted from these multiple data sources in efficient way; *Second*: the need for human (operator) intervention to support the process of correlation engines for aggregating related alerts and to select the most suitable countermeasure for a specific type of attacks; *Third*: there is an urgent need for real time analysis of security events. According to [44], Some efforts in this regards have been done by the design of intelligent IDS based on autonomous learning to adapt with emerging security breaches in real time, or by providing self-learning and adaptation in the event correlation process itself. In order to overcome the aforementioned challenges, authors in [44] proposed an enhanced correlation engine, using genetic programming, to automatically learn and generate correlation rules by considering the context for different types of attacks.

SIEM systems are used increasingly against security breaches in critical structure. More recently, authors in [45] proposed a novel SIEM system to enhance cyber-security in hydroelectric dam. The proposed system was intended to resolve security policies conflicts in existing SIEMs systems, detect illegitimate network paths and reconfigure network devices according to them, and propose intrusion detection and fault tolerant storage system that is able to ensure the integrity of stored events.

In general, some obstacles to improving the maturity of security in organizations are identified [46] and summarized into the following: *First*: cyber threats continue to evolve exponentially either in the volume or in the emerging of new threats, such as the most sophisticated APTs attacks. Therefore, in order to detect and remediate such kinds of attacks, some additional requirements for traditional security management systems and SIEM are needed in order to enhance their ability to detect incidents and response promptly. *Second*: the rapid changes of IT such that it supports virtualization, cloud computing, internet of things (IoT), and bring your own device (BYOD) programs makes an additional difficulty, and adds uncertainty to the traditional security management systems or SIEM to be engaged for such rapid changes. *Third*: a sharp security skills shortage adds another difficulty to deal with new emerging and evolving cyber threats.

In particular, existing security systems that are based on firewalls and other network perimeter devices in addition to signature based systems are not anymore sufficient against insider threat landscape because of the following reasons [46]: *First*: current security analytics tools such as legacy security information and SIEM platforms cannot handle the vast amount of collected data and processing requirements; *Second*: existing traditional security analysis tools provide monitoring and investigation against known and explicit kinds of threats such as malware threats, network threats and others, in specific location of the organization. Therefore, it lacks the ability to provide an aggregated view of the whole enterprise instead of reporting and analyzing scattered events throughout the enterprise. *Third*: existing traditional security tools depend on human intelligence, which requires continuous training on new trends in technology or infrastructure. Machine intelligence can solve that problem if adopted, so that the machine needs lower supervision as it can learn the new circumstances and changes automatically; Fourth: Incident response is not considered automatically. Instead, security analytics tools still independent from the security response systems.

To conclude, a new paradigm to security analysis is really needed such that it solves the previous shortcomings in traditional security analytics tools mentioned above. The new paradigm should allow the processing of high volumes of data streaming in real time, handle the changes in environments and learn the new circumstances automatically, deals with different data structures at the same time with less human intervention. This new paradigm, Big Data analytics, was adopted to various fields and also attract the audience of cyber-security scientists. The rest of this paper discusses and investigates the applicability of adopting Big Data analytics to cyber-security. We start the investigation by giving an introduction to Big Data analytics principles, application areas, and general challenges in Section 3.

III. Big Data Analytics

In this section, fundamentals, and applications of Big Data analytics are explored. We start by giving in brief the fundamentals. Then, some important selected applications of Big Data are described briefly.

A. Fundamentals of Big Data

Big Data has been one of the most important current and futuristic research frontier. According to Gartner [47], turning Big Data into everywhere intelligence was one of the three constitutes the tope 10 strategic technology trends for 2015. Big Data is defined in Gartner [48] as “high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making

and process automation". Another definition was given in [49] as a very huge and diverse data sets collection that is difficult to process using traditional and new data processing platforms. Figure 3 shows these three dimensions of Big Data and their perspectives. Some other researches such as in [50] added the fourth "V" to the definition and refers to the value of data.

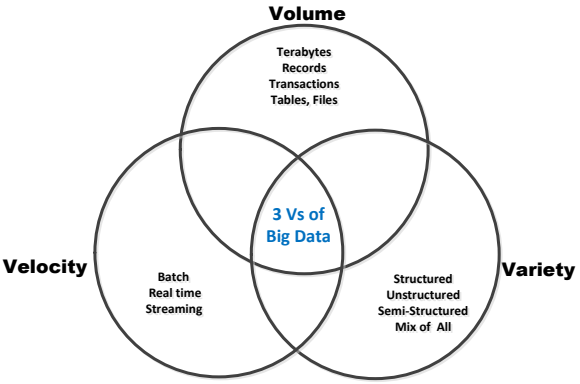

Figure 3: The 3 "Vs" dimensions of Big Data

The first "V" refers to volume indicates that terabytes of data records from transactions, tables, and files are generated. According to Massachusetts, Facebook generates more than 500 TB of data daily. It was reported that Facebook's data warehouse has a capacity of 300 PB of data, while the incoming daily rate of data is about 600 TB in 2014 [51]. The second "V" refers to velocity, specifies that data has a nature of real time, streaming which are generated at too high rate. According to Massachusetts, Big Data concerns of analyzing 2 million records daily to identify why some types of data loses is happened. The third "V" which refers to variety, shows how different data types from various resources are considered in Big Data such as structured (i.e. employee records in an enterprise), unstructured (i.e. images, audio, video, sensor data, etc), semi-structured, or a mix of all of these types at the same time.

1) Big Data Applications

Different application areas constitute data sources for Big Data analytics. Such applications include but are not limited to: education, sciences, retail, history, entertainment, government, healthcare/ medical, social networking, finance, and transportation. Figure 4 shows these different types of application areas. A brief description of the adoption of Big Data analytics for selected key application areas is given in following paragraphs. The detailed description of all application areas is not the scope of this current review.

Big Data analytics according to the definition given in [52] is the paradigm of applying an advanced data analytic techniques on Big Data sets. Therefore, Big Data analytics combines two terms: Big Data and analytics. It further incorporates the way in how the two terms can be merged to produce the one of the most emerged paradigm in business intelligence (BI) today. The following paragraphs briefly explore some application areas of Big Data analytics in different aspects of human life. For each application, we show the need for adopting Big Data analytics paradigm.

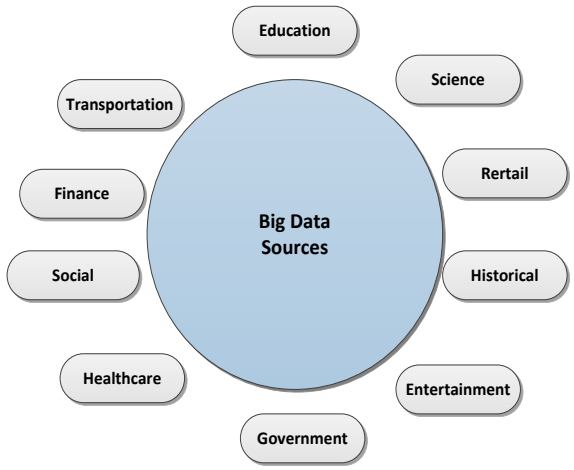

Figure 4: The different sources of Big Data

a) Big Data in Healthcare

A substantial application of Big Data analytics is in the area of human healthcare as this area exhibits largest and fastest growing datasets. It becomes very difficult to measure current size and growth rates in healthcare datasets nowadays [53]. According to [54], the size of clinical data roughly arrived 150 Exabyte in 2011, with an increase between 1.2 and 2.4 Exabyte per year. This kind of clinical data includes electronic medical records (EMRs) and imaging data. Healthcare contains different kinds of data such as personal genomic data, screening data, drug–response profiles, and drug related structures. In addition to these clinical and drug-related types of data, a kind of personal practices and preferences data is also includes such as dietary, habits, environmental factors, and financial records. By integrating all these kinds of data, Big Data analytics helps in improving personal healthcare and global well-being. Big Data allows for fast access to important data elements and facilitates the retrieval of the right data needed at suitable time. Due to the high performance processing, it further allows the real time processing of streaming data collected by sensory platforms such as patient's vital signs in emergent cases and dynamically changing environments. The promise and potential of Big Data analytics in healthcare has been recently reviewed in [55].

b) Big Data in Transportation Systems

One of the important Big Data application areas is the intelligent transportation system (ITS) where the explosive data scale of the flow of cars at intersections is the main problem that causes the ineffective traffic scheduling. Traditional data processing and analysis approaches could not deal with the huge data scale; therefore, researchers investigate more efficient and effective processing paradigms such as Big Data analytics to solve traditional processing problems. An application example of adopting Big Data analytics to ITS can be found in [56]. Authors in [56] proposed a Big Data ITS system named NeverStop, based on soft computing techniques, specifically genetic algorithms and fuzzy control methods. In these systems, sensors were deployed to automatically control the traffic lights at intersections. The fuzzy control method and genetic algorithm were used to adjust and minimize the average waiting time for the traffic lights. Rob Kitchin [57]

investigated the adoption of Big Data concepts and analytics to smart cities towards smart urbanisms concept. In his work, a number of examples and details about how cities can be instrumented with digital infrastructure to produce Big Data were drawn. Further details about the adoption of Big Data analytics for ITS can be found in [49].

c) Big Data in Finance

Financial organizations suffer a high latency in processing transactions due to the shortcoming of processing capabilities, especially the floating-point processing. This issue can be solved by using high performance computing either by adding more CPUs or memory to a single computer system or by adding computing nodes to a pool of computing systems. However, the dramatic growth of financial data shifted the difficulty towards the performance of storage systems too. As a result, the latency is not only affected by the processing capabilities but also by the frequent movement of huge data transactions [58]. This data explosion problem is also known as Big Data problem. IDC [59] reported that in the period between 2005 and 2020, the digital universe is expected to grow by a factor of 300 from 130 Exabyte to 40,000 Exabyte, or 40 trillion gigabytes (more than 5,200 gigabytes for every man, woman, and child in 2020). This means that from now to 2020, the world's data will be doubling every two years. It is also reported in [60] that in 2015, the New York stock exchange data volume reaches about 50-60 billion transactions per day (at peak) with an approximate data size of 15 TB per day (at peak). These example numbers show how the data explosion problem is serious in financial systems. The adoption of Big Data analytics for financial processing help in getting insights from data about customers, competitors, and the market in general which were not available before. These insights enhance the way of making decision and increase the commercial gain [61].

d) Big Data in Social Networks

Recent social networks such as Facebook, Twitter, and LinkedIn connect large populations of users around the world either for social or professional interactions. As a result, exabytes of information are generated daily from that interaction. The evolve of social networking and its corresponding media is considered a Big Data problem in which information system analysts in enterprises predict and study the behavior of social network users in order to enhance their marketing and sales strategies. According to the latest statistics by WeRSM social media [62] in August 2015, Facebook that has 1.4 billion active monthly users generates over 4 million posts every minute, which adds around 250 million posts per hour. In similar way, Twitter users generate 347,222 Tweets per minute or around 21 million Tweets per hour. Figure 5 shows the amount of data generated from several social media networks per minute. This data needs proper processing effectively and timely in order to draw important insights for right decision making.

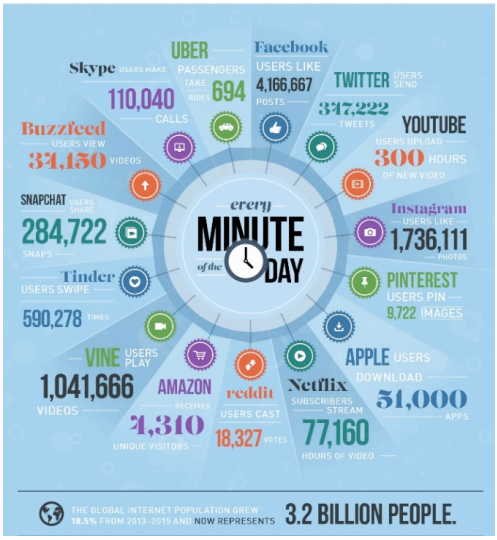

Figure 5: How much data is generated every minute? [62]

e) Big Data in Education

Data can be considered as a key for institutions to enhance their performance towards better outcomes for students by helping educators to know the reasons of negative outcomes such as the reasons behind not graduating on time, dropping or failing courses, not mastering concepts or skills, and many more. It is stated in [63] that if education institutions can detect sharp deviations in the performance and behavior of students, they can have the ability to avoid any poor assessment consequences earlier. According to the report, this predictive capability is really important to take an action in the right time and make decisions promptly for the sake of more effective learning outcomes. A survey was conducted by center of digital education (CDE) [64] in order to study the benefits of Big Data analytics in education. The results reveal that Big Data analytics plays a vital role in analyzing, tracking and predicting student performance, institutional performance, departmental performance, and educator performance. The benefits ratios differ among these sectors and varies according to the level of education either K-12 or higher education. The survey also reveals that Big Data analytics help in adjusting teaching strategies, improving student retention and graduation rate, and finding deficiencies in administrative procedures.

2. Big Data Analytics Tools and Techniques

In order to process the Big Data, a special processing structure is needed. In 2004, Google proposed MapReduce [65] as a programming model that can be implemented for generating and processing large data sets using cluster-based parallel and distributed algorithm. In MapReduce, two main processes are involved which are map and reduce processes. In map process, a set of intermediate key/value pairs is generated using a map function identified by the user. In reduce process, all intermediate value pairs associated with the same intermediate key are merged using a reduce function. MapReduce is the heart of Hadoop, which is open source software for reliable, scalable, and distributed computing. The Apache Hadoop software library is a software framework that provides distributed processing of large data sets on computer clusters using programming

models such as MapReduce. This software is designed to be scalable from single to thousands of machines, with each machine having its own local resources. The service availability of the Hadoop lies on its design to detect possible failures at the application layer instead of relying on hardware in below layers. As a result, the available service can be delivered reliably on top of cluster of computers. Hadoop project includes four main modules, which are, Hadoop common, Hadoop distributed file system, Hadoop YARN, and Hadoop MapReduce. There are other Hadoop-related projects that release some frameworks to support distributed and parallel computing and can be found in the literature such as Ambari, Avro, Cassandra, Chukwa, HBase , Hiv, Mahout, Pig, Spark, Tez, and Zookeeper. For details about these systems, the reader may refer to [49].

According to [49], existing Big Data tools focus on three processing categories, which are, batch processing, stream processing, and interactive analysis. Majority of batch processing- based tools use Apache Hadoop infrastructure, such as Mahout and Dryad. The stream processing tools are required for real time stream data applications. Storm and S4 are examples for distributed and parallel streaming data analytic platforms. The interactive analysis deals with the data interactively, by giving chance for users to process their data instantly. Each user connects to the computer and can interact directly in an online manner. The user can review, compare and analyze the data in either tabular or graphical forms or both of them at the same time.

Extraordinary processing and analysis techniques are needed for Big Data in order to analyze vast volumes of data in constrained time. According to [64], every Big Data application requires a suitable Big Data technique. For instance, online shipping enterprises such as Wal-Mart uses machine learning and statistical-based techniques to discover patterns in their data transactions to provide higher competitions in terms of pricing strategies and advertisements. A number of disciplines are involved in Big Data processing and analysis techniques due to the involvement of Big Data concept in different fields. Figure 6 depicts the most important and common disciplines for Big Data processing. These disciplines include statistical, data mining, machine learning, neuro-computing, signal processing, pattern recognition, optimization, and visualization.

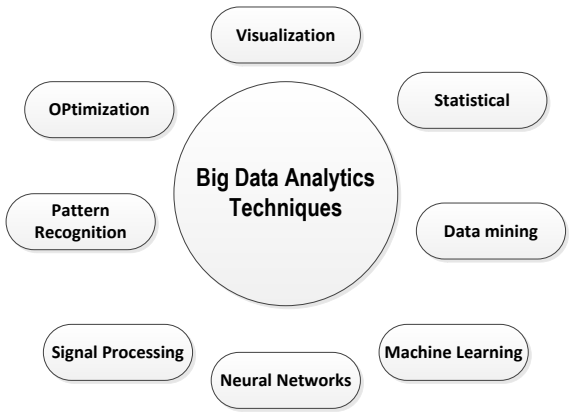

Figure 6: Big Data Techniques and Disciplines

A survey of Big Data techniques can be found in [49, 53, 68]. Many statistical techniques such as cluster analysis, factor analysis, correlation analysis, and regressive analysis have been used for traditional processing systems. They can be adopted also for Big Data processing by tuning them to suit high performance computing. Similarly, data mining techniques such as C4.5, k-means, SVM, Naïve Bayesian, Belief Bayesian among others also have been used for processing data in traditional non-Big Data systems. The key factor for Big Data processing is how to extract important information and insights from vast amount of data and draw conclusions valuable for enterprises and personnel.

B. General Challenges of Big Data

Despite the attractiveness and promise of Big Data analytics in different fields, there are still some challenges that make obstacles for their rapid spread. Some of these challenges were investigated in [66] and can be summarized in the following subsections. We categorized these challenges into three main categories related to Big Data management, visualizations, and security and privacy.

1. Challenges Related to Data Management

Four main challenges can be found related to Big Data management to differentiate it from normal and traditional data. These challenges lie in data warehousing, data diversity, data integration, and data processing and resource management. In terms of Big Data warehousing: Big Data is usually stored as huge amount of unstructured data collected from different sources. The challenge here is how to store and extract the important information from that vast amount of unstructured data in efficient manner. What is the best way of storing such data so that it can be retrieved in sufficient time frame? Does the current file system technology suitable for Big Data storing; what improvements should be given to make it suitable? What are the strategies that should be taken if Big Data should be migrated between data centers or cloud providers? How transparent are these strategies form the user of Big Data? In terms of Big Data diversity: how to deal with huge unstructured data from various data sources? How to get the useful excerpts of such huge data in quick manner? What is the best procedure for aggregation and correlation of the extracted data so that meaningful insights and conclusions can be drawn from it? In terms of Data Integration: should Big Data demand the design of new protocols and interfaces for managing data of various types and from different sources? Finally, in terms of Data processing and resource management: do we need to design new programming models to deal with streaming and multidimensional data? How to optimize resource utilization, especially energy consumption, for streaming data systems applications such as wireless sensor systems? All these challenges need to be addressed in the planning stage of Big Data analytics. All application areas share such challenges but in different degree of difficulties varying from application to another.

2. Challenges Related to Big Data Visualization

In order to provide a real time visualization of Big Data, efficient Big Data processing techniques are needed. Authors in [67] reported that many computational algorithms that have been used for Big Data analytics are complicated and

involve careful parameters tuning to fit some situations in real time. However, doing such that may be critical and time consuming. The authors concluded that some techniques should be utilized in the theme of human machine interactions in order to provide efficient and timely data visualizations. These techniques include: (1) lowering the precision of results; (2) lowering the convergence of the computational model; (3) restricting data scale; (4) data points coarsely processing; and (5) sticking to the visualization device resolution capabilities. Authors in [68] also pointed out the importance of visualization for management of computer networks and software analytics as it is relevant to large-scale infrastructure data analytics.

3. *Challenges Related to Big Data Privacy and Security*

It is good to differentiate between Big Data security and security using Big Data concepts. The former refers of ensuring that Big Data is protected from any kind of security breaches, whereas the latter refers to making use of Big Data concepts for enhancing the security against cyber threats. In this subsection, the Big Data security and privacy issues are explored briefly as one of the serious Big Data challenges, while the adoption of Big Data analytics for cyber-security will be investigated in details in Section 4. The author of [69] investigated security and privacy issues on Big Data and concluded that there is a fine line between the legal use of that data and customer privacy. Therefore, it is further stated that the biggest challenge of Big Data in terms of security point of view is the protection of user‘s privacy. The reason as stated by author is that security breaches on Big Data can have catastrophic consequences compared to normal data because it may affect much large number of people from different perspectives. The author suggested that enterprises should determine the most sensitive pieces of Big Data (Such as identities) and need to isolate them carefully to ensure compliance. In the past, large datasets were structured in a way that any sensitive information can be retrieved and queried easily. However, with Big Data that constitutes variety of structured, unstructured, and semi-structured data, querying a sensitive data is a complex and tedious process. From security point of view, it happens that various users may need to access to some particular subsets of information. To facilitate such requests, traditional encryption and access control solutions will not be sufficient anymore, and need to be modified so that it suite the new data structure demands. For example, access control policy should be designed such that users can only access the information they are authorized to.

Five major security-related considerations should be taken when dealing with Big Data as stated in [69] which are, data anonymization, data encryption, access control, security policy, monitoring, and governance schemes. The anonymization involves removing sensitive information from Big Data to address the privacy concerns. In terms of encryption, it is suggested that operations should be carried out on encrypted data instead of plaintext data. Real time threat intelligence and monitoring of Big Data is also necessary for ensuring security and privacy preservation. The policy consideration involves the identification of sensitive information within unstructured data. It further ensures that the data should not be stored as soon as it is no longer needed for any purpose. Regarding governance schemes, the newness of Big Data concept is behind the lack for procedures and policies; but some schemes are now emerging to cope with this new concept. The issues of data privacy and security in healthcare sector are life-threatening. The reason behind that lie in that clinical data handling must be dealt with not only in high compliance with laws and procedures but also in preserving privacy considerations. Some patient’s data if misplayed may cause a life consequences and may lead to death. In addition, some pharmaceutical data are considered valuable intellectual property, and its use should be protected even in restricted environments [53]. An additional source of information about security and privacy challenges in Big Data can be found in [70].

Some other challenges related to cloud computing as an infrastructure for Big Data processing such as the design of business model challenges and assisting human experts in gaining insights for better decisions can be found in [66]. The following sections explore the adoption of Big Data analytics for cyber-security, which is the main objective of this paper. In this objective, we explore the chances, requirements, and challenges that may be faced in this adoption. Some future directions remarks will be provided towards the end of this paper.

IV. Adoption of Big Data Analytics for Cyber-security

Data-oriented information security has been in use for decades in the forms of bank fraud detection and anomaly-based intrusion detection systems. Looking to the amount of data generated in fraud detection systems, these systems can be considered Big Data analytics systems as they generate millions of data instances and events daily for medium and large enterprises. Some applications areas of fraud detection includes credit card companies, healthcare, insurance, telecommunications among others. According to [71], intrusion detection from the context of data analytics has evolved through three generations as shown in Figure 7.

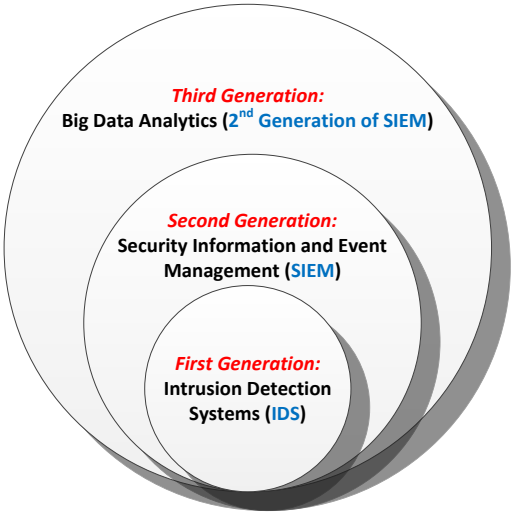

Figure 7: Evolution of Intrusion Detection in the Context of Data Analytics

First generation: Intrusion detection systems (IDS) in its

traditional form as explained in section 2, was evolved to identify security breaches that other security mechanisms such as risk assessments, malware detections, and other network perimeter security tools missed.

Second generation: Security information and event management (SIEM) which has an additional rule in aggregating and filtering alarms form different sources (i.e. IDS sensors) of the first generation.

Third generation: Big Data security analytics, which is sometimes considered as 2nd generation SIEM. It advances the efforts given in 2nd generation by reducing the time consumed in correlating and consolidating security event information. Moreover, it makes a use of context and long-term historical data for forensic analysis of cyber threats.

A. Requirements of Big Data Analytics for Cyber-security

Any Big Data analytics based cyber-security solution should satisfy some requirements that are necessary to cope with the continuous and rapid growth of sophisticated cyber threats. These requirements should consider the characteristics of Big Data itself in addition to the security demands. Based on our investigation, the following set of requirements should be considered.

1. Dealing with Multi-Sourced Data

There is a huge growth in data sources that can be considered by cyber-security systems based on Big Data analytics context. Such sources include but are not limited to, firewall logs, active directory files, operating systems events logs, SIEM data, IDS data, SQL server logs, NetFlow data, threat intelligence data, and others. Such sources of data were existed since long time, but were not utilized in the context of Big Data analytics all together. Considering the data from such multiple sources together is required for getting useful insights to detect and prevent cyberthreats much effectively.

2. Large Scale Data Management:

as a result of the growth of data sources, data volumes are increased and become an obstacle for the design of Big Data analytics based cyber-security systems. Therefore, the design of such systems should consider this requirement in order to be able to collect, process, and retrieve useful data in a timely manner effectively. Cloud computing based technologies, cluster and grid computing are platforms that should be utilized efficiently when design any Big Data based cyber-security systems to facilitate the storage, processing, retrieve necessary data, and draw useful insights about systems events when needed.

3. Dealing with Various Data Types:

the rate of data generation besides the growth of data sources increase the number of data types that can be encountered, from highly structured data to the highly unstructured datasets. In the context of security analytics, most of data that was utilized in the pats were numeric in nature and from one data type. However, nowadays, highly unstructured data can be gathered from several sources such as e-mails, blogs, social networks activities, threat feeds, and combinations of these and others sources. Therefore, a proper design of suitable Big Data analytics systems and tools is needed to consider these different data types together.

4. Data Visualization

Visualization it is a key factor in cyber threat intelligence that provides a graphical descriptive analysis of security related data. The visualized connections between devices, events, locations, signatures, and IPs, provide a way to revealing data anomalies, and intrusions. Therefore, visualization is crucial to understand these connections and extract insights about the behavior of the network systems. Keylines [72], a visualization dashboard developed by Cambridge Intelligence Corporation is a network visualization tool that is designed to visualize cyber threats to allow users to perform better and more effective data analysis. It provides a way to extract important hints from complex connected cyber data. It has four main capabilities which are: (i) analyzing software threats and vulnerabilities; (ii) detecting anomalous logins; (iii) finding pattern in data breaches; and (iv) detecting malware propagation patterns over time. The importance of such visualization applications is that they provide a way to involve users in discovering patterns and anomalies by turning raw connected data into powerful interactive charts. The need for visual analytics was also pointed out in [5] as a tool that helps security teams to understand the relations and track the historical patterns among security data elements. Visual Analytics Suite for Cyber Security (VACS) [73] is another visual analytics system designed by combining multi-criteria clustering techniques and uses three types of interactive visualizations, which are treemap, node-link, and chord diagrams. The VACS as Keylines aimed to gain insights from various threat landscapes. VACS was mainly designed as a dashboard interface that provides overview of host-based thumbnail and an interface for querying and retrieve information to investigate suspicious hosts.

5. High Performance Infrastructure Technology

The key infrastructure technologies that support Big Data analytics in general such as cloud computing, distributed computing, grid computing, stream processing, Big Data modeling, Big Data structure, and software systems should be thoroughly studied. In cyber-security, like any other Big Data applications, it is clear that the data that is utilized as evidence of attacks and security breaches is growing across the three Vs dimensions of Big Data, which are volume, velocity, and variety. As a result this growths harden the detection of such attacks using the traditional technology. For example, it is difficult to detect the most sophisticated APT attacks only by utilizing the traditional information retrieval system being used to design traditional IDS. Instead, advanced technology should be utilized such as MapReduce structure. According to [71, 74], by using MapReduce implementation, the detection systems of APT have higher chance to efficiently handle sophisticated APT unstructured data that has different formats and collected

from different sources such as system logs, IDS, NetFlow, Firewalls, and DNS systems during long period of time. Moreover, the capabilities of MapReduce in handling massive parallel processing facilitate the adoption of very sophisticated detection algorithms that traditional SQL-based data systems fail to handle. In addition, the design of MapReduce as map and reduce functions makes it easy and flexible to users to incorporate more detection algorithms. Such design makes the distributions transparent to users who work with specific data directly. It can be concluded that, the utilization of large-scale distributed systems will simplify the analysis of huge amounts of data concurrently, and hence, provide a mechanism to unveil more attacks paths and targets for detecting unknown and difficult threats like APTs.

B. *Current Big Data Analytics based Cyber-security Systems*

There exist few technology products that adopt Big Data analytics for Cyber-security threat intelligence. These products have been designed and are currently in use by some venders. Most of these products are commercial platforms whereas the research side related to these platforms still in its early stages. The following subsections explore some of these current systems in order to realize the success of Big Data analytics adoption for cyber-security intelligence. It is important to indicate that, there might be additional products that are not covered in this paper due to the lake of good documentation or the rare information about them.

1) *IBM's QRadar and InfoSphere BigInsights Solutions Platforms*

QRadar [75, 76], a product of IBM, is a Big Data platform developed to extend the benefits of 2nd generation SIEM technology. It is designed to extend the visibility into network systems, users, and applications to provide actionable intelligence against potential security breaches. According to the report published by IBM security intelligence research team [6], QRadar solutions are being used by major enterprises to gather and correlate billions of events and network traffic every day in various deployment locations. QRadar security intelligence platform integrates different security intelligence products such as SIEM, anomaly detection systems, log management systems, forensics systems, and vulnerability management and configuration systems into a unified architecture. As a result, such integration facilitates detection of APTs, provides easy to use products, and reducing the total cost of ownership. Figure 8 shows how the scalable IBM QRadar platform considers data from different range of sources and reducing it to manageable offenses based on existing and predefined rules. Figure 8 clearly shows that different types of data sources (variation dimension) such as database activity, user activity, network activity, virtual activity such as clouds based activities and others are considered all together as Big Data sources. This Big Data is then fed into event correlation process and anomaly detection.

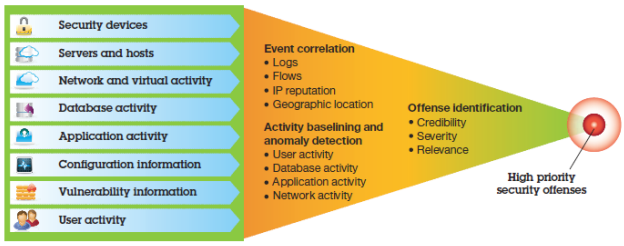

Figure 8: QRadar Security Intelligence Platform Structure [6]

The event correlation studies the relationships between different events extracted from logs, network flows, IPs and geographical locations. However, the anomaly detection engine tries to find possible deviations of behaviors from user, databases, application, and network activities. The results of event correlation and anomaly detection processes are then mapped to identify possible offenses and the degree and relevance of them to determine the priority security breaches as the final output of the platform. To extend the QRadar security intelligence platform, IBM developed InfoSphere BigInsights platform that uses adaptive analytics with Big Data capabilities. Figure 9 shows the structure of integrating InfoSphere BigInsights platform with the QRadar security intelligence platform.

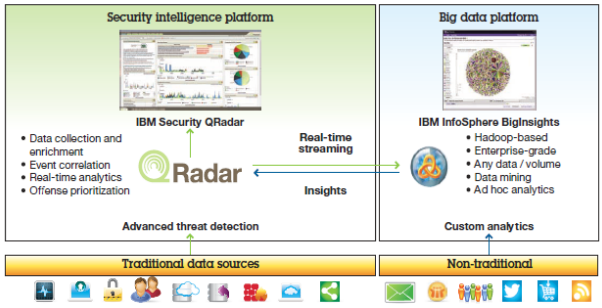

Figure 9: InfoSphere BigInsights and QRadar platforms [6]

As shown in Figure 9, QRadar utilizes the data provided by traditional data sources including user, network, and application activities. This data is used as a source to detect sophisticated threats such as APTs through enrich data collection, finding the relationships between different events, and the real data analytics through anomaly detection and data forensics. On the other hand, InfoSphere BigInsights utilizes other types of unstructured data (non-traditional) based on Big Data platforms infrastructure such as Hadoop to perform custom analytics for further enhancements in cyber threats detection in real time. InfoSphere BigInsights can utilize huge and massive amounts of structured and unstructured data sources that accommodate both volume and variety of data in order to improve the accuracy in real time and feed the extracted insights back to the QRadar. As stated in [6], such integration results in an intelligent solution that gathers, monitors, analyzes, detects, and reports any cyber-security incident that may occur in a comprehensive manner in real time.

2) *Teradata Aster AppCenter for Security*

In cyber-security intelligence, Teradata [77] helps enterprises leverage all their data to extract powerful, predictive insights against end cyber intrusions. Teradata produces a single

platform that integrates information security, cyber security, network operations, data analytics and reporting in order to enable cyber warfare specialists to proactively defend and protect data, applications and other network resources from various types of threats. According to the center of digital government ‘s report [78], as attacks become worse and enterprises respond with their greater capabilities, existing security resources are inefficient, resulting in less effective security results. The report highlights the situation that lead to the adoption of Big Data analytics which in turns yield the proposing of Teradata s’ Aster AppCenter for security. AppCenter was proposed as a web-based solution suite with a simple and user friendly GUI to design, develop, share and consume interactive Big Data applications.

The AppCenter facilitates the embedding of SQL-MapReduce and SQL-GR (SQL parallel graph) scripts, thus making the applications building, configuring, running, and sharing accelerated and simplified. Teradata further introduces an integrated data warehouse (IDW) that provides a base for innovative analysis of network activities for the goal of near real time threat detection and remediation. IDW combines structured, semi-structured, and historical data of intrusion patterns together with other compliance data in order to provide intelligence collection and event correlation. The structured data is stored and analyzed in the IDW. Teradata also integrates the IDW with another platform, that is the Teradata Aster Discovery platform and provides MapReduce functions using SQL syntax for easy access to Apache Hadoop and discovery analytics. The IDW and the discovery platform are integrated together to collect and analyze the unstructured Big Data related to cyber-attacks that was previously difficult to be processed by the conventional security analysis tools. To connect all these platforms together in a unit, Teradata introduces the unified data architecture (UDA) unit. The UDA as shown in Figure 10, provides high speed connection between structured and unstructured security Big Data and their analytics environments. It also provides a way for security analysts to query about structured and unstructured data. By discovering the characteristics between patterns, the path to data breach can be traced and the threats can be isolated in a timely manner.

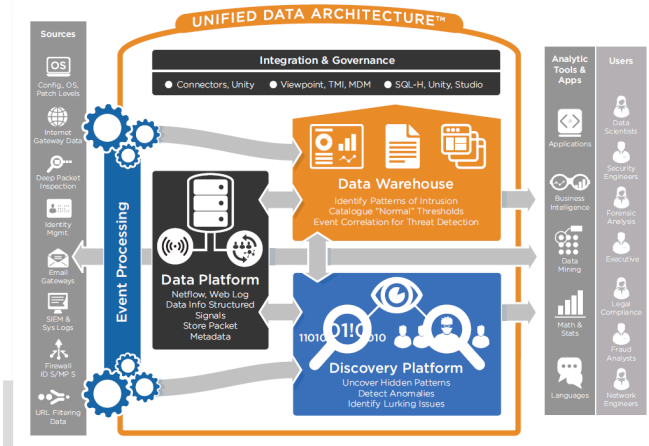

Figure 10: Teradata Unified Data Architecture unit [77]

3) BotCloud Platform

BotCloud [80], a research project lunched in 2011 in which the MapReduce was utilized to analyze massive quantities of NetFlow data in order to detect infected hosts. According to the project documentations, around 720 million records of NetFlow (77GB) were gathered per day. Therefore, it was impossible to process such amount of data through traditional security analysis tools, and hence the adoption of Big Data analytics solutions such as MapReduce for distributed computing was rationale. The mechanism of BotCloud platform depends on the use of BotTrack, an engine that test the relationships between hosts on the botnet using PageRank and clustering algorithms. Then the detection process of infected hosts has three steps, which are the creation of dependency graph, the PageRank algorithm, and the DBScan clustering algorithm. The BotCloud platform starts by constructing the dependency graph from the Netflow records by representing each host either in the source or destination of the record process by a node. PageRank algorithm then finds the patterns in the constructed graph that have similar characteristics. The clustering algorithm is finally used to group together hosts that have same patterns resulted from the PageRank step. The PageRank algorithm only resides in the MapReduce platform due its high computational cost incurred by finding the patterns. Figure 11 shows the structure of the BotCloud Big Data analytics framework for detecting botnets.

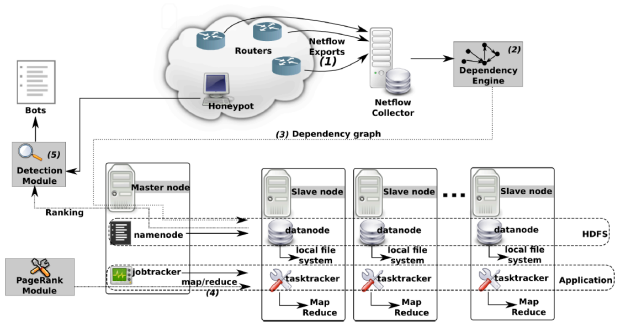

Figure 11: BotCloud Framework [80]

According to [80], BotCloud utilizes Hadoop cluster of 12 nodes (one master + 11 slaves), 6 nodes has Intel Core 2 Duo 2.13 GHz with 4 GB memory each, and 6 nodes has Intel Pentium 4.3 GHz with 2GB memory each. The size of dataset collected was 16 million hosts and 720 million NetFlow records, which resulted in a dependency graph of 57 million edges. It is clearly, that the main factor of computational complexity is the number of edges in the graph. The results of detection accuracy of botnets and the time needed for analyzing the 57 million edges are reported in [81]. The results revealed that a factor of seven was achieved in reducing the time for analyzing the 57 million edges with high accuracy rates. This example of Platform shows how the Big Data analytics play major role for detecting malicious incidents in a huge datasets which was previously impossible to be performed in considerable amount of time.

4) Beehive framework

Beehive [81], a Big Data analytics based framework developed by RSA to account APTs sophisticated attacks. The name Beehive was inspired by the multiple weak components which are sensors that cooperate together to detect APTs attacks like bees with different roles that maintain a hive. Beehive aimed to extract knowledge and insights from dirty log data produced by various security products in large enterprises. Therefore, it fulfills the requirements of Big Data paradigm in which log dirty data is generated massively from various products. The mechanism of Beehive depends on identifying suspicious host behavior to detect security breaches. These breaches are then analyzed and investigated further by breach response team to determine whether a security policy is violated or attack has been lunched. Beehive detects any behavioral deviation through anomaly sensors, where one sensor examines an aspect of host activity within the enterprise. The outcome of the alerts from different sensors indicates that a malicious behavior is suspected and therefore a further investigation by response team is needed. The combination rules for each pattern sets of security incidents is set by human analysts in advance based on the expected behavior change by such incidents. Results reported in [81] showed that the Beehive has the ability to process around billion log messages of day data in an hour. It further showed that the Beehive has successfully detected policy violations and malware infections that other traditional security analysis tools could not detect. In the same context of Beehive, some of Beehive project team developed an Early-Stage Enterprise infection detection framework by Mining Large-Scale Log Data [83]. The proposed framework was based on belief propagation and inspired by graph theory. Authors claimed that their proposed framework can be applied either with seeds of hacked hosts or set of malicious domains given by the enterprise or without seeds. It was claimed that the proposed framework techniques perform well on detecting enterprise infections with high accuracy and low false alarm rates using two months DNS logs datasets released by Los Alamos National Lab (LANL), including APT attacks simulated by LANL domain experts. The proposed framework was further investigated on 38TB of real-world web proxy logs and found that hundreds of malicious domains that were undetected by state-of-the-art traditional security analysis tools, have been identified successfully. Beehive and the framework proposed in [83] showed two examples of frameworks that adopted Big Data analytics to detect malicious incidents. However, the detailed structure of those two frameworks were not revealed as the focus was on the ability of Big Data analytics in extracting insights based on the behavior to detect sophisticated incidents.

5) DOFUR Platform

DOFUR [84] is a distributed denial of service (DDoS) Forensics Platform that uses Hadoop‘s MapReduce as an infrastructure for Big Data analytics. It was designed to detect the packets of DDoS attacks that conventional security analysis platforms fail to detect due to the increase of network log file size. Authors concluded that the use of parallel processing based of Hadoop’s MapReduce reduced the time needed to analyze the massive log file entries and enhanced the efficiency in detecting DDoS attacks.

6) LogRhythm’s Security Intelligence Platform

LogRhythm Platform for security intelligence [85], was designed by LogRhythm company, a next-generation SIEM which means Big Data analytics based on the taxonomy of Figure 7. It combines the processes of log managements, network monitoring, digital forensics and threat management in one platform that also provides procedures for response on real time. It further provides, according to designers, a mechanism to prevent breaches before they happen by providing early detection of attacks behavior.

7) Blue Coat Security Platform

The Blue Coat Security Platform [86] was designed by Blue Coat Company to provide protection against broad types of advanced security breaches by integrating sophisticated technologies. It mainly aims to protect against the most sophisticated APTs and detect sophisticated malware extensions. It further benefits from incident response and forensics techniques designed by Blue Coat to reveal any threats that might be hidden. When forensic discovers any threats, its indicators and impacts are recorded and reported to the security analysts in order to update the security environment parameters.

8) WINE Platform

The Worldwide Intelligence Network Environment (WINE), is a platform proposed by Dumitras and co-authors in [87-89] for performing large scale data analysis. According to [87], WINE contains field data, collected by 240,000 sensors worldwide, to design new experimental studies, and facilitates the evaluation of proposed research works on whole security threats lifecycle. Upon preparing the WINE dataset, a WINE platform was proposed for repeatable experimentation on the WINE datasets and correlating with other metadata collected to understand the results. WINE loads data samples and performs the aggregation on data of millions hosts worldwide. WINE dataset was hosted and used by Symantec’s security engineers and analysts and by academic researchers to conduct empirical studies and validate their results against reference datasets. Due to the different issues related to the use of security related data such as privacy and other legal concerns, there is a need for public dataset that reflects the real life scenario and provide a board for validating different proposed security related solutions. Therefore, WINE made this purpose a reality by allowing researchers using massive amounts of security related data collected by sensors worldwide to perform their experiments and validate their proposed algorithms on such dataset samples. Sine WINE was hosted on Symantec it contains

anti-virus telemetry and intrusion protection telemetry which records the occurrences of host and network threats, accordingly. Figure 12 shows the structure of WINE platform. WINE further allows open-ended experiments using parallel processing techniques on WINE datasets.

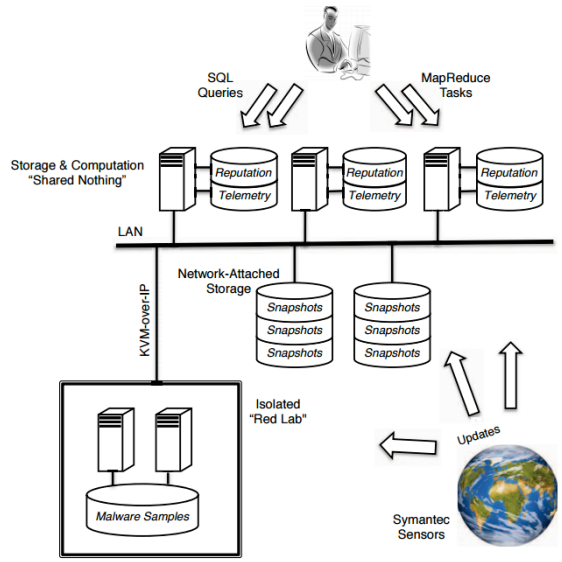

Figure 12: WINE architecture [87]

An example of cyber-security analytics through WINE dataset is the determination of duration of Zero-Day attacks. A study conducted in [90] used WINE dataset platform to measure the duration of 18 zero-day attacks by integrating the anti-virus and reputation telemetry datasets and analyzed data collected from 11 million hosts worldwide. A detail description of this analysis can be found in [90]; however, the results showed that this type of attacks are more common than previously claimed no to be. The outcome of the analysis points out the importance of using Big Data analytics for security research. The detection of zero-day attacks was previously undetected because they are rare events in honeypots or in lab experiments. The use of Big Data identified that 150 hosts were infected out of 11 million of analyzed hosts. The development of platforms such as WINE platform based on Big Data analytics opens up new opportunities for advancing cyber-security research and towards discovering new security breaches that have not seen before.

C. *Challenges of Big Data Analytics adoption in Cyber-security*

In this section, the challenges that face the adoption of Big Data analytics solutions to cyber-security are highlighted.

1) Dealing with Unstructured Data

One of the Vs dimensions that defines the Big Data is the variety. Various types of data characterize Big Data analytics systems, which are structured, semi-structured and unstructured. Traditional security analysis systems such as log mining, intrusion detection and even SIEM systems deal with well-structured data originate from one data source such as system log files, database logs, or even the Netflow records. However, the consideration of semi-structured or unstructured data that combines various sources of information such as e-mails, social media, threat feeds and other security related sources in addition to the well and known structured security related data is still a challenge. Two main factors govern the dealing with unstructured data, which are the rate of data generation and the increasing of data sources. Therefore, dealing with unstructured data issue requires dealing with those two factors together. Big Data analytics for cyber-security systems should be designed in a way that cope with the fast growth of data sources and the vast amount of data generated across the time.

2) Real Time Analysis

As the technology landscape is quickly and dynamically emerged, cyber threats evolve quicker and keep changing their sophisticated strategies in attacking networks or individuals. Therefore, fast and effective detection of potential cyber threats is really a challenge and becomes more necessary in order to avoid severe damage to the enterprises sensitive information, reputations, and further human lives in some extreme and critical systems. Author of [91] stated that, there should be systems that are capable of collecting, processing, and analyzing millions of logs originate from multiple data sources that contains system events, applications logs data, and databases transactions logs as soon as they are generated. This indicates the importance of real time analytics of Big Data in order to provide enterprises with constant and timely monitoring of their activities across the time. It is further stated that, to maximize the benefits of log data rather than log management capabilities, security analytics systems should also incorporate real time analysis and immediate response capabilities together with advanced correlation, behavioral, statistical, and intelligent pattern recognition techniques. According to the report in [70], the problem of real time security and compliance monitoring might even increase with Big Data due to the volume and velocity of data streams generation. The report further stated that the advancements of Big Data analytics infrastructure can close the gap of real time analysis of security information to provide real time security analytics.

3) Streaming Data Analysis

Streaming data analysis involves quickly processing of significant amount of data in near real time. Sometimes, it is called processing data in motion, and it has straight relation with real time analytics discussed in the previous point. Traditional security analysis tools dealt with data streams in batch processing manner, such that historical data is processed after specific amount of time. An example of that is the process of detecting frauds in financial transactions. The data for hours or may be a day is collected and analyzed towards the end of that time period. However, this analysis

should be fast and in real time to be practical and to provide actionable response at the suitable time before further damage may be caused by malicious incidents. Security related data has streaming nature due to the dynamicity of network structure. Furthermore, the advancements of infrastructure is fast, especially with internet of things (IoT) applications which are based on sensors technologies. Therefore, this streaming analysis feature is a challenge for Big Data analytics if the application demands real time analysis and response.

4) Data Privacy

One of the most important challenges against the successful adoption of Big Data analytics for cyber-security is the privacy of data. Privacy issues violate the reuse principle which involves using the shared data only for the purposes that it was collected for. In the context of traditional data use, privacy was referred to the effects it makes when dealing with sensitive datasets. However, with Big Data analytics, the privacy violation becomes more challenging due to the fact that Big Data analytics facilitate extracting insights and draw conclusions about individuals or enterprises by correlating small pieces of information from different sources. Therefore, Big Data analytics solutions should be designed such that they have less impact on data reuse and privacy preserving regulations. For security related data, the situation become more sensitive. This type of data could not be shared for the sake of Big Data analytics because it may contain private data that disclose the identity of users. By sharing such data and making it available for Big Data analytics parties, this data becomes more exposed to cyber criminals who may use it for further sophisticated cyber security breaches incidents. As a possible mitigation approach, Alvaroe et al [4] suggested that Big Data analytics systems designers and architects should consider safeguarding the data in order to prevent any misuse of Big Data stores during the design stage of the solutions.

5) Data Provenance

Provenance is defined in the literature as a source (origin) of data. It is considered as one of the important challenges for Big Data analytics systems in general and for cyber-security analytics in particular. Although, Big Data facilitates the expanding of data sources used for processing, it is uncertain that these sources meet the trustworthiness requirements of the analytics algorithms used by the proposed analytics solutions to produce accurate and trusted results. As a rule of thumb, authenticity and integrity requirements should be applied to the data of different sources before their actual use in the analytics. The situation becomes further difficult when some malicious data inserted in the pool of Big Data to extract insights that are necessary for making important and critical decisions. This issue converts to a problem of Big Data security, which was discussed in section 3.22. To conclude, while the Big Data analytics solutions are deployed for enterprises systems, new tools to deal with security in such solutions are required besides the use of conventional security analysis tools available before.

6) Visual analytics

It is the use of interactive visual interfaces in order to convey information to humans in an interpreted way in order to get more knowledge and extract insights out of the visualized subject. The importance of visual analytics for cyberthreats intelligence is explored previously in this paper. Keylines [72] was shown as an example of cyberthreats intelligence visualization tools that can be deployed together with cyber-security dashboards in order to show the network connectivity and provide reports about the network state at a point of time. Visual analytics is a challenge for Big Data analytics based security solutions especially when real time analysis of streaming data is considered. The visualization dashboard should be designed such that it can track the dynamicity of events generated on the motion and provide security analysts with clues that were not obvious before, such as the unusual traffic inbound or outbound of host or group of hosts in a network segment. The difficulty of modeling the networks connectivity especially with large-scale enterprises makes the visual analytics mission further difficult and a hard challenge for Big Data analytics cyber-security systems.

7) Adaptiveness

Another important challenge in adopting Big Data analytics solutions for cyber-security is how to cope with the dynamic changes in systems and networks behavior? How to consider the adaptiveness especially when using machine learning and computational intelligence methods to design Big Data analytics algorithms? The issue of adaptiveness become harder when it is considered together with real time and streaming data processing as the learning of the new system state or normal behavior need to be on the motion and should also consider security, privacy, provenance and other requirements at the same time.

V. Future directions

Based on the requirements and challenges of adopting Big Data analytics for cyber-security, that were identified and explored in previous sections, we state some directions for potential future research in the following paragraphs.

A. Big Data Security and Privacy

As shown in the Big Data analytics challenges, security of Big Data is an important challenge that should be overcome. More authentication schemes are needed in order to improve the trustworthiness of data collected from various sources to prove the provenance of data. In addition, existing anomaly data detection methods that have been used successfully for traditional security systems should be properly augmented for Big Data in order to ensure the correctness of data in automated manner and in real time, and detect any insider incidents. In terms of privacy, further privacy preserving

schemes should be designed for Big Data analytics in cyber-security context. Further rules and guidelines should be introduced and monitored by government's agencies to keep the pieces of data related to people identities and their private matters unrevealed.

B. Behavioral Analytics

Another area of possible research is the behavioral analytics, which involves the consideration of context information in order to increase the possibility of detecting patterns and anomalous behaviors that indicate frauds, thefts, or other cyber-security breaches. Traditional cyber-security solutions were successful with outsider intrusions while it failed to detect insider incidents. However, with monitoring the behavior of normal and legitimate users, Big Data analytics solutions can predict unexpected behaviors and detect insider threats using anomaly detection in interaction behavior such as, how many times file is downloaded? How many times the database is accessed? To conclude, some kinds of insider threats will be difficult be detected unless modeling the normal and anomalous behaviors of users.

C. Visualization

There is a need for advanced visualization tools that provide security analysts with insights that may assist them in saving the time of detecting cyber threats. Some existing security related visualization dashboards are already in use such as Keylines; however, further advancements in this regard are required especially with the increase number of data sources that makes data visualization more complicated, and with the demand of real time processing of streaming data.

D. Internet of Things

IoT applications are examples of Big Data applications such as smart cities, ITS, habitat monitoring and others, which involve the generation of streaming data and real time processing. Therefore, there is a need for adopting Big Data analytics for security issues of such IoT applications. Each type of applications has its own requirements and demands in terms of security and privacy; therefore, it requires suitable augmentation of design to suit such demands.

VI. Conclusions

Cyber threats are getting sophisticated due to rapid advancements of technology landscape, which make the task of mitigating them timely and effectively further difficult. Traditional cyber-security tools such as log management, conventional IDS, IPS, and either SIEM tools are inadequate to deal with the new threats and tactics. Furthermore, the emerge of Big Data that is huge in volume, high in generation rate, and has various data types, makes the detection of cyber threats in such context with traditional and conventional tools insufficient. Therefore, Big Data analytics solutions become necessary for alleviating such sophisticated threats in Big Data. In order to be efficient, Big Data analytics cyber-security solutions should fulfill some basic requirements. They are required to deal with data originated from various sources and have better management strategies with large-scale data. They are also required to handle data of different types and properly visualize it for drawing conclusions and extracting insights in quick and simple manner. To be able to fulfill all these requirements, Big Data analytics cyber-security solutions should have proper and high performance infrastructure that facilitates dealing with Big Data in different contexts. Some Big Data analytics security solutions have been proposed and are in use for some big industrial companies such as IBM and Teradata. However, some challenges cause an incomplete adoption of Big Data analytics for cyber-security. These challenges include the difficulty to deal with unstructured and sophisticated data, and the need for real time and streaming data analysis. This adoption is further challenged by issues related to data privacy and provenance, and by the need for adaption with dynamic changes in data behaviors. To further enhance this adoption, some future directions are suggested such as enhancing the privacy and security schemes that hide the sensitive security related information. In addition, behavioral analytics should be combined with Big Data analytics for mitigating insider threats properly. More visualization tools are needed, so that security analysts can draw some useful and preliminary conclusions about cyber threats and security breaches for further investigation. Finally, the successful adoption should consider the applicability to the various types of IoT applications.

References

[1] Cyber-Security Definitions; National Initiative of Cyber-security Careers and Studies (NICCS), USA. https://niccs.us-cert.gov/glossary; Access date :31/03/2016.

[2] Kasper Security Bulletin 2015: Kaspersky Overal statistics for 2015, in Kaspersky Corporation.

[3] Robert Eastman, Michael Versace, and Alan Webber, Big Data and Predictive Analytics: On the Cyber-security Front Line: White Paper, Feb 2015: International Data Corporation (IDC).

[4] C. Alvaro. A, P.K. Manadhata, and S.P. Rajan, Big Data Analytics for Security. IEEE Security & Privacy, 2013. 11(6): p. 74-76.

[5] J. Oltsik, An-Analytics-based Approach to Cyber-security, May 2015: Enterprise Strategy Group (EGS).

[6] Extending Security Intelligence with Big Data Solutions: Leverage Big Data Technologies to uncover Actionable Insights into Modern, Advanced Data Threats, IBM Software : Thought Leadership White Paper, 2013.

[7] S.H. Ahn, N.U. Kim, and T.M. Chung. Big Data Analysis System Concept for Detecting Unknown Attacks. In: 16th International Conference on Advanced Communication Technology, 2014.

[8] K. Gai, M. Qiu, and S.A. Elnagdy. A Novel Secure Big Data Cyber Incident Analytics Framework for Cloud-based Cyber-security Insurance. In: Big Data Security on Cloud, IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS).

[9] J. Hu and A.V. Vasilakos, Energy Big Data Analytics and Security: Challenges and Opportunities. IEEE Transactions on Smart Grid, 2016, 7(5): p. 2423-2436.

[10] T. Mahmood and U. Afzal. Security Analytics: Big Data Analytics for Cyber-security: A Review Of Trends, Techniques and Tools. In: 2013 2nd National Conference on Information Assurance (NCIA), 2013. IEEE.

[11] M. Marchetti,,F. Pierazzi, A. Guido and M. Colajanni, Countering Advanced Persistent Threats through Security Intelligence and Big Data Analytics. In: 2016 8th International Conference on Cyber Conflict (CyCon). 2016. IEEE.

[12] A. Razaq, H. Tianfield, and P. Barrie. A Big Data Analytics based Approach to Anomaly Detection. In: 2016 IEEE/ACM 3rd International Conference on Big Data Computing Applications and Technologies (BDCAT), 2016. IEEE.

[13] B. Blakley, E. McDermott, and D. Geer, Information Security is Information Risk Management. In: Proceedings of the 2001 Workshop on New Security Paradigms2001, ACM: Cloudcroft, New Mexico. p. 97-104.

[14] A.P.H. de Gusmão, L. C. e Silva, M. M. Silva, T. Poleto, and A. P. C. S Costa, Information Security Risk Analysis Model Using Fuzzy Decision Theory. International Journal of Information Management, 2016. 36(1): p. 25-34.

[15] C.C Lo and W.J. Chen, A hybrid Information Security Risk Assessment Procedure Considering Interdependences Between Controls. Expert Systems with Applications, 2012. 39(1): p. 247-257.

[16] N. Feng and M. Li, An Information Systems Security Risk Assessment Model Under Uncertain Environment. Applied Soft Computing, 2011. 11(7): p. 4332-4340.

[17] N. Feng, H.J. Wang, and M. Li, A Security Risk Analysis Model For Information Systems: Causal Relationships of Risk Factors And Vulnerability Propagation Analysis. Information Sciences, 2014. 256: p. 57-73.

[18] Raj Chaudhary and J. Hamilton, The Five Critical Attributes of Effective Cyber-security Risk Management, July 2015, Crawe Horwath.

[19] S. Paul and R. Vignon-Davillier, Unifying Traditional Risk Assessment Approaches With Attack Trees. Journal of Information Security and Applications, 2014. 19(3): p. 165-181.

[20] M. Christodorescu, S. Jha, S. A. Seshia, D. Song, and R. E Bryant. Semantics-Aware Malware Detection. In: 2005 IEEE Symposium on Security and Privacy, 2005.

[21] G. McGraw and G. Morrisett, Attacking Malicious Code: A Report to the Infosec Research Council. IEEE Software, 2000. 17(5): p. 33-41.

[22] A. Vasudevan and R. Yerraballi. Spike: Engineering Malware Analysis Tools Using Unobtrusive Binary-Instrumentation. The 29th Australasian Computer Science Conference. 2006. Australia.

[23] K. Griffin, S. Schneider, X. Hu, and T.C. Chiueh, Automatic Generation of String Signatures for Malware Detection, In: Recent Advances in Intrusion Detection: 12th International Symposium, RAID 2009, Saint-Malo, France, September 23-25, 2009. Proceedings, E. Kirda, S. Jha, and D. Balzarotti, Editors. 2009, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 101-120.

[24] Y. Ye, D. Wang, T. Li, D. Ye, and Q. Jiang, An intelligent PE-malware detection system based on association mining. Journal in Computer Virology, 2008. 4(4): p. 323-334.

[25] M. Jain and P. Bajaj, Techniques in Detection and Analyzing Malware Executables: A Review. International Journal of Computer Science and Mobile Computing, 2014. 3(5): p. 930-935.

[26] R Bace and P. Mell, Intrusion Detection Systems, National Institute of Standards and Technology (NIST), Technical Report 800-31.2001.

[27] L. P, Network Based Anomaly Detection Using Self Organizing Maps, Technical Report, 2002, Nova Scotia: Dalhousie University Halifax.

[28] Snort; http://www.snort.org.

[29] H.J. Liao, C. H. R. Lin, Y. C. Lin, and K. Y. Tung, Intrusion Detection System: A Comprehensive Review. Journal of Network and Computer Applications, 2013. 36(1): p. 16-24.

[30] C. Day, Chapter 26 - Intrusion Prevention and Detection Systems A2 - Vacca, John R, in Computer and Information Security Handbook (Second Edition). 2013, Morgan Kaufmann: Boston. p. 485-498.

[31] E.W. Fulp, Chapter 6 - Firewalls A2 - Vacca, John R, in Managing Information Security (Second Edition). 2014, Syngress: Boston. p. 143-175.

[32] J.R. Vacca and S.R. Ellis, 3 - Firewall Types, in Firewalls. 2005, Digital Press: Burlington. p. 49-57.

[33] K. Kent, Guide to Computer Security Log Management, 2007.

[34] J.R. Vacca and S.R. Ellis, 18 - Auditing and Logging, in Firewalls. 2005, Digital Press: Burlington. p. 299-309.

[35] J. Myers, M.R. Grimaila, and R.F. Mills, Towards Insider Threat Detection Using Web Server Logs, in Proceedings of the 5th Annual Workshop on Cyber Security and Information Intelligence Research: Cyber Security and Information Intelligence Challenges and Strategies 2009, ACM: Oak Ridge, Tennessee, USA. p. 1-4.

[36] A. Williams, Security Information and Event Management Technologies. Siliconindia, 2006. 10(1): p. 34-35.

[37] R. Gabriel, T. Hoppe, A. Pastwa, and S. Sowa, Analyzing Malware Log Data to Support Security Information and Event Management: Some Research Results. In: '09. First International Conference on Advances in Databases, Knowledge, and Data Applications (DBKDA), 2009.

[38] HPE ArcSight SIEM solution, Hewlett Pacard Enterprise. HPE Tecnical Reports, 2016.

[39] RSA Envision Platform: Simplify compliance and optimize incident management; RSA Technical Reports, https://www.emc.com/collateral/data-sheet/9245-h9037-3in1-ds.pdf.

[40] M.B. Godfrey, netForensics-A Security Information Management Solution, 2002, SANS Institute, InfoSec Reading Room.

[41] OSSIM: The World's Most Widely Used Open Source SIEM; https://www.alienvault.com/products/ossim, AlienVault Tehnical Reports.

[42] LookWise: SIEM ; http://www.s21sec.com/en/technology/lookwise/siem.

[43] LogLogic : Data Security Management; http://www.loglogic.com/about/index.php, LogLogic Tehnical Reports.

[44] G. Suarez-Tangil, E. Palomar, A. Ribagorda, and I. Sanz, Providing SIEM Systems With Self-Adaptation. Information Fusion, 2015. 21: p. 145-158.

[45] C. Di Sarno, A. Garofalo, I. Matteucci, and M. Vallini, , A Novel Security Information And Event Management System For Enhancing Cyber Security In A Hydroelectric Dam. International Journal of Critical Infrastructure Protection.

[46] J. Oltsik, The Big Data Security Analytics Era Is Here, 2013, Enterprise Strategy Group (ESG).

[47] The Top 10 Strategic Technology Trends for 2015; https://www.gartner.com/doc/2964518?ref=ddisp; Access date: 17/04/2015, Gartner Inc Technology Research.

[48] Big Data Definition, Gartner: http://www.gartner.com/it-glossary/big-data/; Access date: 17/04/2016, Gartner Inc. Technology Research.

[49] C.L. Philip Chen and C.Y. Zhang, Data-Intensive Applications, Challenges, Techniques And Technologies: A survey on Big Data. Information Sciences, 2014. 275: p. 314-347.

[50] V. Mayer-Schönberger and K. Cukier, Big Data: A Revolution That Will Transform How We Live, Work, and Think. 2013: Houghton Mifflin Harcourt.

[51] Scaling the Facebook data warehouse to 300 PB; https://code.facebook.com/posts/229861827208629/scaling-the-facebook-data-warehouse-to-300-pb/; Last upated: April, 11 2014; Access date: 17/04/2016.

[52] P. Russom, Big Data analytics. TDWI Best Practices Report, Fourth Quarter, 2011: p. 1-35.

[53] K. Kambatla, G. Kollias, V. Kumar, and A. Grama, Trends in Big Data analytics. Journal of Parallel and Distributed Computing, 2014. 74(7): p. 2561-2573.

[54] How Big is Big Data in Healthcare; http://blogs.sas.com/content/hls/2011/10/21/how-big-is-big-data-in-healthcare/; Acess date: 18/04/2016.

[55] W. Raghupathi and V. Raghupathi, Big Data Analytics In Healthcare: Promise And Potential. Health Information Science and Systems, 2014. 2(1): p. 1-10.

[56] C. Wang, X. Li, X. Zhou, A. Wang and N. Nedjah, Soft Computing In Big Data Intelligent Transportation Systems. Applied Soft Computing, 2016. 38: p. 1099-1108.

[57] R. Kitchin, The Real-time City? Big Data and Smart Urbanism. GeoJournal, 2014. 79(1): p. 1-14.

[58] X. Tian, R. Han, L. Wang, G. Lu, and J. Zhan, Latency critical Big Data Computing in Finance. The Journal of Finance and Data Science, 2015. 1(1): p. 33-41.

[59] J. Gantz and D. Reinsel, The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East, 2012, IDC.

[60] New York Stock Exchange Oracle Exadata ;http://www.oracle.com/technetwork/database/availability/con8821-nyse-2773005.pdf; Acess date: 19/04/2016.

[61] How the financial services sector uses Big Data analytics to predict client behaviour; http://www.computerweekly.com/feature/How-the-financial-services-sector-uses-big-data-analytics-to-

predict-client-behaviour; Access date: 18/04/2017, in.

[62] WeRSM: we are social media; http://wersm.com/how-much-data-is-generated-every-minute-on-social-media/; Acess date: 19/04/2016., in.

[63] K. Reid-Martinez, Big Data in Education: Harnessing Data for Better Educational Outcomes, 2015, Center of Digital Education.

[64] Center of Digital Education; http://www.centerdigitaled.com/; Acess date: 20/04/2016.

[65] J. Dean and S. Ghemawat, MapReduce: Simplified Data Processing On Large Clusters. Communications of the ACM, 2008. 51(1): p. 107-113.

[66] M.D. Assunção, R. N. Calheiros, S. Bianchi, M. A. Netto, and R. Buyya, Big Data Computing and Clouds: Trends And Future Directions. Journal of Parallel and Distributed Computing, 2015. 79–80: p. 3-15.

[67] J. Choo and H. Park, Customizing Computational Methods for Visual Analytics with Big Data. IEEE Computer Graphics and Applications, 2013. 33(4): p. 22-28.

[68] T. Menzies and T. Zimmermann, Software Analytics: So What? IEEE Software, 2013. 30(4): p. 31-37.

[69] G. Lafuente, The Big Data Security Challenge. Network Security, 2015. 2015(1): p. 12-14.

[70] Top Ten Big Data Security and Privacy Challenges; Technical report, 2012.: Cloud Security Alliance (CSA).

[71] Big Data Analytics for Security Intelligence; Cloud Security Alliance (CSA); Big Data Working Group; Technical Report: Sep 2013.

[72] Visualizaing Cyber Threats with KeyLines; Cambridge Intelligence; http://cambridge-intelligence.com/keylines/. Cambridge Intelligence Technical Reports, 2015.

[73] F. Fischer and D.A. Keim. VACS: Visual Analytics Suite for Cyber Security. IEEE VIS. 2013.

[74] P. Giura and W. Wang, Using large scale distributed computing to unveil advanced persistent threats. Science J, 2012. 1(3): p. 93-105.

[75] QRadar Security Intelligence Client Study. Research Report Independently conducted by Ponemon Institute LLC; Jul 2015.

[76] IBM QRadar Security Intelligence Platform; http://www03.ibm.com/software/products/en/qradar; Access date: 03/05/2016.

[77] Teradata Cyber Security Analytics; http://www.teradata.com/solutions-and-industries/cyber-security-analytics/?ICID=mainnav&LangType=1033&LangSelect=true; Acess date: 04/05/2016.

[78] Cyber-security: A Research Report from the center for digital government; http://assets.teradata.com/resourceCenter/downloads/WhitePapers/CDG13 PCIO SR_Q3.pdf.pdf.pdf.pdf?processed=1; Access date: 05/04/2016.

[79] Cyber Security: Big Data Integration and Analytics for Cyber Security, Teradata Technical E-book, 2015.

[80] J. Fran, S. Wang, W. Bronzi, R. State, T. Engel, BotCloud: Detecting Botnets Using Mapreduce. In: 2011 IEEE International Workshop on Information Forensics and Security. 2011.

[81] J. François, S. Wang, and T. Engel, Bottrack: Tracking Botnets Using Netflow and Pagerank, in Networking 2011, Springer. p. 1-14.

[82] T. F. Yen, A. Oprea, K. Onarlioglu, T. Leetham, W. Robertson, A. Juels, and E. Kirda, Beehive: Large-Scale Log Analysis For Detecting Suspicious Activity, In: Proceedings of enterprise networks, the 29th Annual Computer Security Applications Conference 2013, ACM: New Orleans, Louisiana, USA. p. 199-208.

[83] A. Oprea, Z. Li, T.F. Yen, S. H. Chin and S. Alrwais. Detection of Early-Stage Enterprise Infection by Mining Large-Scale Log Data. In: 2015 45th Annual IEEE/IFIP International Conference on Dependable Systems and Networks. 2015.

[84] R. Khattak, Z. Li, T. F. Yen, S. H. Chin and S. Alrwais. DOFUR: DDoS Forensics Using MapReduce. In: Frontiers of Information Technology (FIT), 2011. 2011.

[85] LogRhythm's Security Intelligence Platform; https://logrhythm.com/pdfs/datasheets/lr-security-intelligence-platform-datasheet.pdf; Access date: 12/05/2016, LogRhythm Company Technical Reports.

[86] Blue Coat Security Platform; https://www.bluecoat.com/products-and-solutions/advanced-threat-protection; Access date: 12/05/2016, Blue Coat Website.

[87] T. Dumitras and D. Shou, Toward A Standard Benchmark For Computer Security Research: The Worldwide Intelligence Network Environment (WINE), in Proceedings of the First Workshop on Building Analysis Datasets and Gathering Experience Returns for Security 2011, ACM: Salzburg, Austria. p. 89-96.

[88] T. Dumitras and I. Neamtiu, Experimental Challenges in Cyber Security: A Story of

Provenance and Lineage for Malware. CSET, 2011. 11: p. 2011.9-9.

[89] T. Dumitras and P. Efstathopoulos. The Provenance of WINE. In: 2012 Ninth European of Dependable Computing Conference (EDCC), 2012.

[90] L. Bilge and T. Dumitras, Before We Knew It: An Empirical Study of Zero-Day Attacks in The Real World, In: Proceedings of the 2012 ACM Conference on Computer and communications security 2012, ACM: Raleigh, North Carolina, USA. p. 833-844.

[91] A. Macrae, Identifying Threats in Real Time. Network Security, 2013. 2013(11): p. 5-8.

Author Biographies

Murad A. Rassam is a Researcher at Department of Computer Science, Faculty of Computing, Universiti Teknologi Malaysia (UTM). He obtained his PhD and MSc of Computer Science from UTM on June 2013 and April 2010, respectively. His main research interests include wireless sensor networks security, sensor data quality assurance, anomaly detection, data mining and the application of machine learning techniques for computer and network security.

Mohd. A. Maarof is a Professor at Faculty of Computing, Universiti Teknologi Malaysia (UTM). He obtained his BSc (Computer Science) and MSc (Computer Science) from USA and his PhD from Aston University, Birmingham, United Kingdom in the area of Information Technology (IT) Security. He is currently leading the Information Assurance & Security Research Group (IASRG) at UTM. Currently, his research interests include intrusion detection system, malware detection, web content filtering and cryptography.

Anazida Zainal is a Senior Lecturer at Department of Computer Science, Faculty of Computing, Universiti Teknologi Malaysia. She obtained her BSc (Computer Science) from USA and her MSc and PhD from UTM. Her research interests include network security, intrusion detection system, wireless sensor networks and data mining.