Available online www.jsaer.com

***Journal of Scientific and Engineering Research*, 2018, 5(5):586-597**

Review Article

ISSN: 2394-2630
CODEN(USA): JSERBR

Big Data: An Innovative Tool for Improving Cyber Security Issues

Aru Okereke Eze, Amadi Christopher C.

Department of Computer Engineering, Michael Okpara University of Agriculture Umudike, Nigeria

Abstract Cyber security has become a Big Data problem as the size and complexity of security related data has grown too big to be handled by traditional security tools. Cyber security concern affects organizations across all industries, including retail, financial, communications and transportation industries. Advances in data collection and computational statistics coupled with increases in computer processing power, along with the plunging costs of storage are making technologies to effectively analyze large sets of assorted data everywhere. Applying big data technologies to an ever growing number and variety of internal and external data sources, businesses and institutions can discover hidden correlations between data items, and extract actionable insights needed for innovation and economic growth. While on one hand big data technologies yield great promises, on the other hand, they raise critical security, privacy, and ethical issues. This paper introduced and discussed extensively a means of improving Cyber security using Big Data technologies, ontology, and decision support for preventing or reducing losses from cyber-attacks. This research work expressed data as currently one of the most important assets for companies. The Paper narrates the continuous growth in the importance and volume of data and their associated problems which may be handled by traditional analysis techniques. The paper suggested solving the problems through the creation of a new paradigm: Big Data. In conclusion, Big data technology is invariable an efficient tool which can be used to improve cyber security.

Keywords Big Data, Cyber Security, Communications, Technology

1. Introduction

The ability to accumulate large amounts of data provides the opportunity to examine, observe, and notice irregularities to detect network issues. Better actionable security information reduces the critical time from detection to remediation, enabling cyber specialists to predict and prevent the attack without any delays. Data is analyzed using algorithms which give critical insight to organizations in order to provide assistance in improving their services. Big Data is continuing to be used on bigger platforms including financial services, health services, weather, politics, sports, science and research, automobiles, real estate, and now cyber security. Big Data and analytics are some of the most effective defenses against cyber intrusions. Big data refers to datasets that are not only big, but also high in variety and velocity, which makes them difficult to handle using traditional tools and techniques. Due to the rapid growth of such data, solutions need to be studied and provided in order to handle and extract value and knowledge from these datasets. Furthermore, decision makers need to be able to gain valuable insights from such varied and rapidly changing data, ranging from daily transactions to customer interactions and social network data. Such value can be provided using big data analytics, which is the application of advanced analytics techniques on big data. The volume and variety of data produced by and about individuals, things or the interactions between them have exploded over the last few years. Such data can be replicated at low cost and is typically stored in searchable databases which are publicly (or at least easily) accessible over the Internet. According to recent IBM estimates, 2.5 billion Gigabytes of data are created everyday around the globe, and the creation rate is growing continuously. McKinsey estimates that the amount of digital content on the Internet is expected to grow by 44 times to 2020, at an annual growth rate of 40% [1].

This trend describes a phenomenon broadly known as the emergence of big data. The big data phenomenon itself is in part being enabled by the rising popularity of Web 2.0 (esp. online social networks) applications, the low cost of computation and storage, the rapid emergence of new computing paradigms such as cloud computing, breakthrough innovations in the field of data mining and artificial intelligence, combined with the wide availability of sensor equipped and Internet-compatible mobile devices. Big data is nothing but assortment of such huge and complex data so that it becomes very tedious to capture, store, process, analyze and retrieve it with the help of on-hand database management tools or traditional data base management techniques.

A. Cyber Security Today

Cyber security is a set of tools, practices, and guidelines that can be used to protect computer networks, software programs, and data from attack, damage, or unauthorized access [2]. The Internet allows users to gather, store, process, and transfer vast amounts of data, including proprietary and sensitive business, transactional, and personal data [3]. At the same time that businesses and consumers rely more and more on such capabilities, cyber security threats continue to plague the Internet economy. Cyber security threats evolve as rapidly as the Internet expands, and the associated risks are becoming increasingly global. Staying protected against cyber security threats requires all users, even the most sophisticated ones, to be aware of the threats and improve their security practices on an ongoing basis. Creating incentives to motivate all parties in the Internet economy to make appropriate security investments requires technical and public policy measures that are carefully balanced to heighten cyber security without creating barriers to innovation, economic growth, and the free flow of information. Concern over the proliferation of cyber security threats is well-documented and well-founded.

B. Big Data

Big Data has several defining characteristics including volume, variety (of data types and domains-of-origin, and the data flow characteristics of velocity (rate) and variability (change in rate) in which the data is generated and collected; Traditional data systems collect data and curate it into information stored in a data warehouse, with a schema tuned for the specific analytics for which the data warehouse was built. Velocity refers to a characteristic that has been previously referred to as streaming data. The log data from cell phones for example flows rapidly into systems, and alerting and analytics are done on the fly prior to the curation and routing of data or aggregated information into persistent storage. In a Big Data Architecture this implies the addition of application servers to handle the load. Variability refers to changes in the velocity of data flow, which for cost effectiveness leads to the automated spawning of additional processors in cloud systems to handle the load as it increases, and release the resources as the load diminishes. Volume is the dataset characteristic most identified with Big Data. The engineering revolution began due to the massive datasets from web and system logs. The implication has been the storage of the data in its raw format, onto distributed resources, with the curation and imposition of a schema only when the data is read. The variety characteristic often is used to refer to multiple formats of data, recognizing that the much larger amounts of unstructured text, image and video data have vital information to be harvested. This results in more sophisticated curation and pre-analytics to extract useful information, but not a change in architecture. Variety more broadly refers to the use of data from multiple domains. While volume and velocity are revolutionary in the information technology (IT) engineering, the variety characteristic drives a revolution for the organization in both the engineering and the mission by allowing previously impractical or impossible analytics. Techniques for handling variety will change our analytical capabilities. Big data refers to the mining of usable information from the massive amounts of data being created worldwide every day across all industries [4]. While businesses and government agencies take advantage of this influx of information to improve operations, increase sales and lower costs, cyber criminals are mining the same data for unethical reasons.

C. Big Data as a Double-Edged Sword for Cyber Security

Big data has developed a new role in preventing adversaries from taking advantage of the massive amounts of military intelligence, trade secrets, and personal and financial data available through systems at all risk levels. Organizations are being encouraged to transition to intelligence-driven security for a broader view of risk and

vulnerabilities. This requires analyzing external threat intelligence feeds, cloud-based calendars and documents, social network activity logs, website-generated information feeds and other non-traditional sources of security information. Big data's advantages lie in the ability to analyze massive numbers of potential security events and make connections between them to create a prioritized list of threats. With big data, seemingly disparate pieces of data connect to form a clear picture, enabling cyber security professionals to stay ahead of possible threats and help prevent attacks from happening. However, just as organizations and cyber security teams are using big data to increase their efficiencies, so are hackers. Using sophisticated technologies, they are able to distill the data they want from millions of Trojan-infected computers. Cyber criminals have developed plug-in to query databases to transfer certain information, like credit card numbers, bank URLs or social security numbers into separate databases that they have full access to. In addition to creating ways to mine big data for illicit gain, cyber criminals are also using it to monitor their processes and improve their own efficiency. They use big data to learn more about infected machines, breached databases and compromised information systems. They use it to spot trends, failures and successes and to make their next attack more effective [5].

2. Big Data Analytics

Much of the development of Big Data Engineering is a result of the need to analyze massive web log data. Massive Web logs were first filtered by page for aggregate page counts, to determine the popularity of 2 pages. Then the pages were analyzed for sessions (spawning the now massive "cookie" industry to make this simpler). "Sessions" are the sequence of activities that describe a customer's interaction with the site at a "single-setting", with the analyst describing what time-window is considered a session. The next step in analytics capability came from the realization that these sessions could be abstracted into patterns rather than being treated as just the literal collection of pages. With this step, traversal patterns helped site designers sees the efficiencies in their link structure. Furthermore these usage patterns could in some cases be attached to a customer account record. With this step, the site could now be tuned to the benefit of the most valuable customers, with separate paths being designed for the casual visitor to browse, leaving the easy efficient handling for loyal customers. This pattern-oriented analysis applies to the cyber domain, in analyzing logs from a server. Big Data analytics then moved into what is termed "Social Network Analysis" (SNA) to analyze a link node structure [6]. Beginning with Google's PageRank algorithm, a huge field of analytics has developed to understand the relationships between nodes represented by their link structure. This has applicability to cyber through determining the appropriateness of activity between servers. The complexity in cyber is that additional information on the "nodes" must be integrated in order to analyze for the appropriateness of any activity.

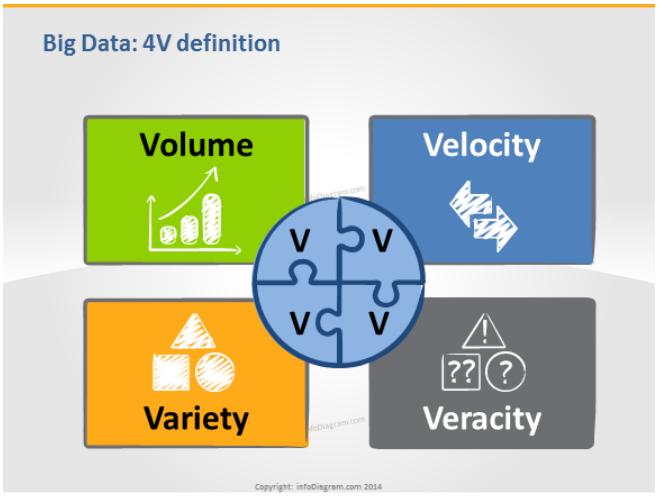

Figure 1: Big data Analytics

- **Volume:** Big data is any set of data that is so large that the organization that owns it faces challenges related to storing or processing it. In reality, trends like ecommerce, mobility, social media and the Internet of Things (IoT) are generating so much information, that nearly every organization probably meets this criterion.

- **Velocity:** If your organization is generating new data at a rapid pace and needs to respond in real time, you have the velocity associated with big data. Most organizations that are involved in ecommerce, social media or IoT satisfy this criterion for big data.
- **Variety:** Traditional systems handled the variety of data through a laborious integration process to standardize terminology, normalize into relational tables, choose indexes, and store into a data warehouse that is tuned for the specific analytics that are needed. Naturally this is an inflexible process that does not easily accommodate new data sources, changes into underlying data feeds, or new analytical requirements. For Web log analysis, this extension to customer session analytics only required the assignment of a customer or visitor ID to the session, allowing integration with a purchasing history. In the cyber analytics case, the integration point is not so simple. The integration of packet data, with server log data, with port-to-port connectivity data, with server type data, with network router settings, etc. provides a much-more complex use case, needing a more sophisticated way to integrate such a variety of data, some of which carries a number of additional attributes that are needed. Recently variety datasets have been addressed through mashups that dynamically integrated a couple of datasets from multiple domains to provide new business capabilities. Early mashups demonstrated this value, for example in the integration of crime data with real estate listings; a valuable analysis that was not possible prior to the availability of open datasets. The limitation to such mashups is that they typically consist of the integration of a limited number of datasets, with the integration variables being manually selected. This type of manual integration is not sufficient for analytics across a variety of large volume datasets with complex inter-relationships [7].

Variety is the Big Data attribute that will enable more sophisticated cyber analytics. The requirement is for an automated mechanism to integrate multiple highly diverse datasets in an automated and scalable way. This is best achieved through a controlled metadata.

An Architecture for a Big Data Platform

This section describes the architecture of a big data platform.

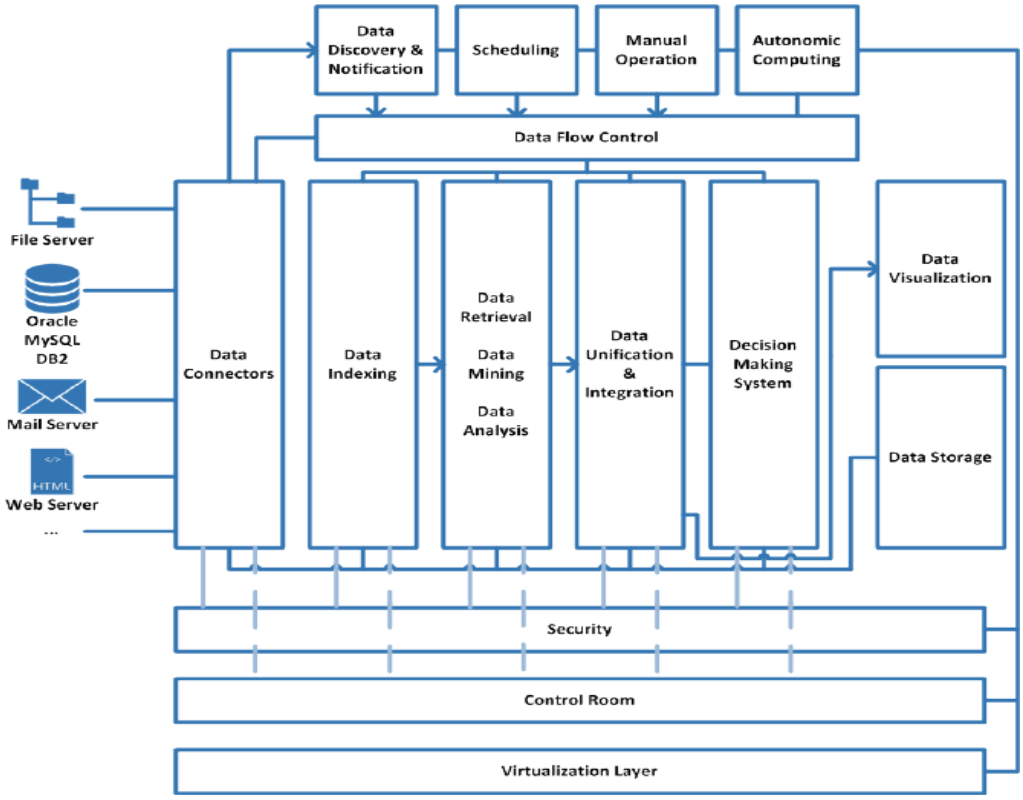

Figure 2: Architecture of Big Data Platform

The architecture is sketched in Figure 1. The architecture shown in Figure 1 has, at its core, a data mining, monitoring, and management platform henceforth referred to as "M3Data". M3Data provides series of components which are interconnected by the Big Data Platform. As represented in Figure 2, M3Data is endowed with connectors for various sources of information (either structured, semi-structured, or unstructured) [8]. It collects the data and puts it in a data flow object. It performs various processing operations on the data, from mining and analytics to extraction, integration, cleansing, and distribution of it. During all these processes, it is monitoring and logging all modifications suffered by the data. The architecture of M3Data consists of six layers of software packages, as shown in Figure 3.

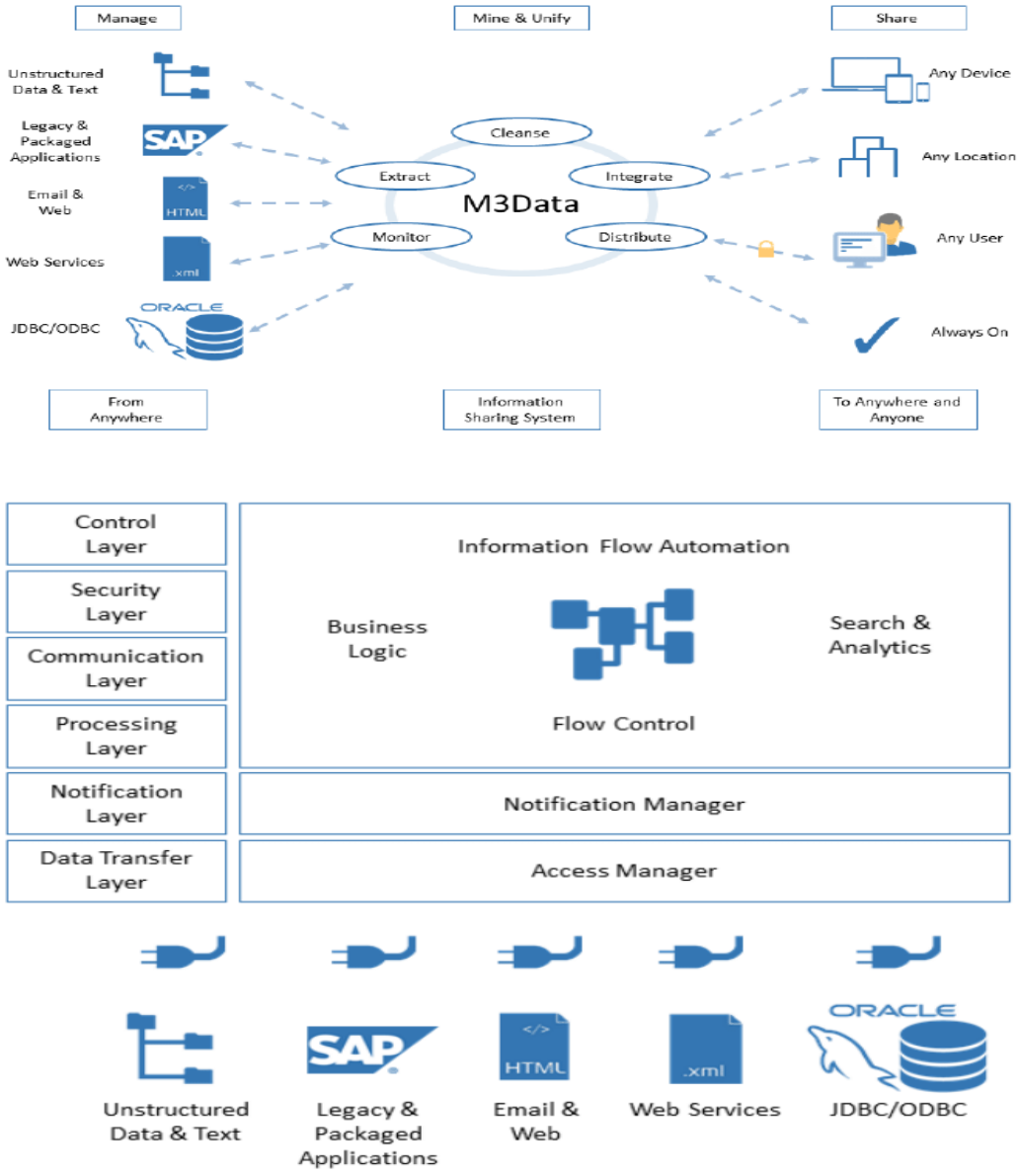

Figure 3: How M3data Works

The layers of the core of the Big Data Platform assure the following functionality, all of which are essential to any big data application, and thus to the Big Data Platform presented in this paper:

(a) platform control: all control functions are implemented in the Control Layer of M3Data;

(b) platform security: specifically, M3Data uses the Role-Based Access Control security principles implemented in a Security Layer, which can also be coupled with an external security mechanism.

(c) platform communication: the Communication Layer implements all protocols needed to communicate with the data sources and inside the building blocks, referred to as "agents", of an M3Data application;

(d) platform processing: a series of mining, unification, and analysis of the data are included in the Processing Layer;
(e) platform event notification: M3Data Notification Layer implements functions for trapping, recording and passing to the Control and Processing Layer all events generated by the database triggers, OS events, emailers, etc.;
(f) data transfer: data transfer from source to destination among all data resources of the platform is implemented in the Data Transfer Layer.
The above layers all concur to the collection, transformation, and transfer of data from source to destination. One of the key modules of M3Data is a graphical programming language which allows users to cut the development time by an order of magnitude [9]. This functionality is embedded in the Data Flow Programming Studio (DFPS) component, which contains all the graphical programming primitives and operators needed to ease the programming burden of designing and implementing a specific big data application. The user therefore only needs to translate the business rules of the application into a data flow describing the path that the data has to go through while various business tasks of the application run and then implement the business logic in a DFPS diagram.

3. Challenges and Issues in Big Data

While the rise of big data yields huge opportunities for individuals, organizations and the society at large, it also raises important privacy and ethical issues. These issues are factors that may lead to situations in which the underlying analytic models and infrastructures are likely to impact privacy negatively from both a legal and an ethical perspective, and hence represent possible obstacles for the big data's potential to be fully realized [10].

A. Challenges to Security and Privacy in Big Data

The massive retention of socioeconomic, demographic, behavioral, financial, and other transactional data for analytic purposes may lead to the erosion of civil liberties due to a loss of privacy and individual autonomy. From a privacy and security perspective, the challenge is to ensure that data subjects (i.e., individuals) have sustainable control over their data, to prevent misuse and abuse by data controllers (i.e., big data holders and other third parties), while preserving data utility, i.e., the value of big data for knowledge/ patterns discovery, innovation and economic growth [11]. The following sections describe some relevant challenges to security and privacy in the context of big data.

Increased Potential for Large-scale Theft or Breach of Sensitive Data

As more data is available, stored in (non-) relational databases accessible on-line, and increasingly shared with third parties, the risk of data breaches also increases. Big data thus raises a number of privacy and security questions related to the access, the storage and the usage of personal/ user related data. A recent series of high-profile data security incidents and scandals, e.g. Edward Snowden's NSA leaks [12] and the breach at the US retail chain Target Corp have demonstrated that data breaches by those who have obtained access to sensitive datasets, legitimately or otherwise, are devastating for both the individuals and the data holders. Unauthorized accesses can possibly involve two types of adversaries: the first type of adversary is interested in gaining access to raw data in order to either compromise the interpretation/analysis process, e.g. by injecting false data into the raw data, or to steal a large volume of sensitive (financial/identity) data. The second type of adversary includes entities primarily interested in accessing different datasets that have already been analysed, as well as the actionable intelligence legitimate analysts have extracted from the data. To breach data privacy, both types of adversaries can exploit software and/or hardware design flaws in the infrastructures behind big data platforms.

B. Challenges in Big Data Integration

The handling of big data is very complex. Some challenges faced during its integration include uncertainty of data Management, big data talent gap, getting data into a big data structure, syncing across data sources, getting useful information out of the big data, volume, skill availability, solution cost etc.

1. The Uncertainty of Data Management: One disruptive facet of big data management is the use of a wide range of innovative data management tools and frameworks whose designs are dedicated to supporting operational and analytical processing. The not only SQL(NoSQL) frameworks are used that differentiate it from traditional relational database management systems and are also largely designed to fulfill performance demands of big data applications such as managing a large amount of data and quick response times. There are a variety of NoSQL approaches such as hierarchical object representation (such as JSON, XML and BSON) and the concept of a key-value storage. The wide range of NoSQL tools, developers and the status of the market are creating uncertainty with the data management.

2. Talent Gap in Big Data: It is difficult to win the respect from media and analysts in tech without being bombarded with content touting the value of the analysis of big data and corresponding reliance on a wide range of disruptive technologies [13]. The new tools evolved in this sector can range from traditional relational database tools with some alternative data layouts designed to maximize access speed while reducing the storage footprints, NoSQL data management frameworks, in-memory analytics, and as well as the broad Hadoop ecosystem. The reality is that there is lack of skills available in the market for big data technologies. The typical expert has also gained experience through tool implementation and its use as a programming model, apart from the big data management aspects.

3. Getting Data into Big Data Structure: It might be obvious that the intent of a big data management involves analyzing and processing a large amount of data. There are many people who have raised expectations considering analyzing huge data sets for a big data platform. They also may not be aware of the complexity behind the transmission, access, and delivery of data and information from a wide range of resources and then loading these data in a big data platform. The intricate aspects of data transmission, access and loading are only part of the challenge. The requirement to navigate transformation and extraction is not limited to conventional relational data sets.

4. Syncing Across Data Sources: Once you import data into big data platforms you may also realize that data copies migrated from a wide range of sources on different rates and schedules can rapidly get out of the synchronization with the originating system. This implies that the data coming from one source is not out of date as compared to the data coming from another source. It also means the commonality of data definitions, concepts, metadata and the like. The traditional data management and data warehouses, the sequence of data transformation, extraction and migrations all arise the situation in which there are risks for data to become unsynchronized.

5. Extracting Information from the Data in Big Data Integration: The most practical use cases for big data involve the availability of data, augmenting existing storage of data as well as allowing access to end-user employing business intelligence tools for the purpose of the discovery of data. This business intelligence must be able to connect different big data platforms and also provide transparency of the data consumers to eliminate the requirement of custom coding. At the same time, if the number of data consumers grow, then one can provide a need to support an increasing collection of many simultaneous user accesses. This increment of demand may also spike at any time in reaction to different aspects of business process cycles. It also becomes a challenge in big data integration to ensure the right-time data availability to the data consumers.

6. Miscellaneous Challenges: Other challenges may occur while integrating big data. Some of the challenges include integration of data, skill availability, solution cost, the volume of data, the rate of transformation of data, veracity and validity of data. The ability to merge data that is not similar in source or structure and to do so at a reasonable cost and in time. It is also a challenge to process a large amount of data at a reasonable speed so that information is available for data consumers when they need it. The validation of data set is also fulfilled while transferring data from one source to another or to consumers as well. This is all about the big data integration and some challenges that one can face during the implementation. These points must be considered and should be taken care of if you are going to manage any big data platform.

4. Semantic Technology

Semantic technologies are crucial for the future handling of big datasets across multiple domains. While we have methods for unique concept identification arising through the Semantic Web, these technologies have not made inroads into traditional data management systems. Traditionally the ETL process has been used to enforce

standard terminology across datasets, with foreign keys to external tables for the related information. This is not a scalable solution, since the introduction of a new data source requires the careful construction of foreign keys to each other dataset in the database. This lack of extensibility to add in additional sources highlights the limitations of horizontal scalability in current approaches. In addition there are limitations on the continued expansion in large data warehouses,
highlighting their inability to continue to scale vertically. Semantic technologies have not yet made inroads into Big Data systems. Big datasets that consist of volume tend to be monolithic, having no attempt to integrate across datasets. The data is typically stored in its raw state (as generated), and in the initial big data engineering no joins were allowed. Given this, most Big Data Analytics approaches apply to single datasets. For solutions addressing the integration of variety datasets, the ability to integrate the datasets with uniquely defining semantic technology is thus a fundamental requirement. Two overarching requirements need to be addressed to use ontology for the integration of big data: the construction of the ontology, and the use of the ontology to integrate big datasets.

A. Application to Cyber-Security

The goals of big data applied to cyber-security are generally to improve the effectiveness and timeliness in these four categories of activities:

1. Identity Management
2. Fraud Detection & Prevention
3. Governance, Risk & Compliance
4. Security Management

For the later, security management, one sub-goal is to attempt to achieve near real-
time awareness of security inside and outside of a networked enterprise, and another is just-in-time (JIT) response to attempt to prevent loss from a split second attack. Perimeter security can be improved, as part of a defense-in-depth approach, by JIT blocking and filtering. Today "after-the-loss" scans and forensics are the status quo in the industry. Actionable data can decay in seconds, and losses
can occur in seconds, so it clear that loss prevention requires JIT, if not near real-
time, response. We believe that Big Data analytics combined with near real-time situation assessment and JIT response will provide an additional level of defense-in-depth. Traditional "firewalls" imply filtering of IP addresses and malware patterns that are updated manually as new information is collected. Big Data will provide more complete and more timely information and knowledge that makes it possible to correlate, analyze and use the data in near real-time to do loss prevention actions; automatically in some situations. These actions typically include:

- Incoming message block
- Outgoing message block
- SSL inspection Initiation (encrypted channel)
- Terminate connection
- Quarantine data or hardware unit
- Lock user's account
- Notifications

For example, Big Data log analytics can be used to consolidate information on bad actors and their previous cyber-attacks, for the purpose of predicting future attempts (bad actor behavior) and thus providing better prediction and detection; and before loss is incurred from the attack. This pattern-detection analytics can be viewed as an optimization problem of maximizing cyber-attack situation awareness and precision response in all "kill chain" phases as documented in Wang [14], and minimizing the number of false positives. Examples are: (1) near real-time update of a firewall with new IP addresses or URLs, for immediate update into the firewall for filter/block, and (2) near real-time (NRT) termination of a session (port cutoff) as soon as it is suspected that it is a connection established by a bad actor, or is infiltrated. Surgically precise NRT responses can maximize the prevention of loss of confidentiality and integrity, while minimizing loss of network availability to others that are not directly involved [5]. To illustrate how this would work in practice, consider these three scenarios.

First, logs collect data on all machines with open connections to locations outside of the enterprise, with a JIT analysis including the variety of additional context datasets. One connection is open to China, but the attributes of the particular user connecting in imply they are not supposed to connect anywhere in Asia. Within split seconds the command is given by the system to cutoff the port and a security violation is logged for follow-up. Second, consider a user who initiates an identical second software application, as determined by an identical hash. This is a specific type of pattern, wherever an identical hash is observed in multiple places it is an indication of a compromised account. The NRT response the connection is terminated/cutoff, and the data and memory images are quarantined and saved for forensics. Off-line cyber forensics is scheduled, the account's user-password is disabled, the user is notified to call for a new password, and the situation details are preserved for future reference; and for potential analysis and learning of bad actor behavior. Finally, consider that analysis of the data results in a discovery that a server has ten times more avi files than typical for a server of that type, which is indicative of a user setting-up and running an unauthorized file server. The user's account is disabled and she is notified to call in, the files on the server are quarantined and preserved for forensics, and a forensic investigation is scheduled.

B. Big Data and Science

Big data has the potential to change science as we know it. Progresses in the last decade in the fields of high-performance computer simulation and complex real-time analytics paired with the rapidly increasing volume and heterogeneity of data from various sources (incl. Web browsing/searching records, genomic, health and medical records, earth observation systems, surveillance video, and sensor, wireless and mobile networks) are shaping the vision of a data-intensive science [16]. The vision of a data-intensive science describes a new approach to the pursuit of scientific exploration and discovery, which leverages an ever-growing amount of research data and thus requires new computing, simulation, and data management tools and techniques. The approach promises to integrate life, physical, and social sciences and covers application domains ranging from computational earth and environmental sciences, genomics, to computational social science. The hope being that data-intensive science would enable mankind to better understand and address some of its most pressing challenges: global warming; efficient supply and use of cleaner energy resources; pandemics and global health monitoring among others. To take the example of computational social science [17]. In the recent years, social scientists have begun collecting and analyzing large volumes of data from sources that were barely imaginable a decade ago. Online social network platforms like Facebook and Twitter, and emerging applications such as participatory sensing, two of such sources, allow for the mass-scale collection and sharing of details about peoples' behaviour, as well as the nature and strength of the interactions between individuals and communities in on- and offline environments. It has been demonstrated that the complex and heterogeneous data from these environments can be leveraged alongside statistical techniques to gain insights into online sociological phenomenenon.

C. Detecting and Fighting Cyber-Crime with Big Data

Fighting (cyber-) crime does not only require a retrospective analysis of possible evidences but also accurate predictions about criminals' behaviors and their adaptive reactions to countermeasures. As chief (information) security officers are struggling to monitor and protect their corporate networks and enterprise systems against increasingly sophisticated and complex security threats, private companies are slowly but surely moving towards adopting big data security analytics tools, i.e., tools that bring advanced data analytics to enterprise IT security. Unlike current security information and event management (SIEM) solutions, big data security analytics tools such as IBM Security Intelligence and Palantir provide the means required to effectively analyze terabytes of (real-time) network events, packet captures, applications' performances and unstructured data from across/outside the organization. By providing means to discover changing patterns of malicious activities hidden deep in large volumes of organizations data, big data security tools can indeed empower businesses to better understand if and how they have been attacked. In addition to providing continuous and detailed insight into security risks, big data security analytics tools can help organizations in their regulatory compliance efforts. Moreover, the collected statistics and the possibly inferred knowledge about attacks/breaches/criminal activities

could then be shared with other organizations across national jurisdictions, for instance in an effort to achieve collaborative network monitoring / collaborative intrusion detection. Another transformative potential of big data security analytics lies in its power to satisfy the ever-growing interest of law enforcement authorities and intelligence agencies in the flood of information from sources as varied as the Web and mobile networks, financial/tax records, travelers' biometric data, satellite imagery, surveillance video, and so forth. Indeed, a clever aggregation and analysis of such data has the potential to gain useful insights into the identities and/or behavioural patterns of (would be) criminals.

4. Positive Effects of Big Data on Cyber Security

Real-Time Monitoring of Secured Systems

Another benefit that big data provides in the realm of cyber security is the ability to monitor and track systems, usually contained within the cloud, for irregularities and potential breaches. Cloud Security Information and Event Management (CSIEM) allows users to safely transmit and store their private information and files without fear of falling victim to a cyber attack. The way the system works is by implementing a complex series of imprints within a file that can be monitored in real time to protect a user from nefarious outsiders that wish to compromise their data. Cyber security professionals, like those at Blue Coat Systems, use this as the Gold Standard in terms of protection from outside attacks. Think of this particular technology as the fingerprints on our hands. These fingerprints can be monitored and analyzed to ensure our safety and the integrity of our identity during key moments. Similarly, the CSIEM framework gives us a detailed description of the events that have taken place in the form of log files. These log files can be used at a later time to authenticate the validity of certain events and can pinpoint areas of interest should a cyber security breach occur.

Safety and Security for the Masses

Big data, among other blooming technologies, keeps the masses safe in an ever-growing world of complexity and technicality. Looking into the future we can see more of our actions losing the physical aspect of social interaction and moving to a digital realm where professionals are needed to protect and serve. It is likely that we will see more companies like Blue Coat emerge as big data continues to boom and there is a greater need to protect it. Imagine the beauty of a future filled with technologies that not only make our lives easier but warn us in advance for areas that can easily be compromised. Big data is not only a new technology, it will become the framework in which business, social interactions, and organizations operate on in the future of our ever-growing marketplace.

5. A Big Data Solution

Cyber security needs the risk management and actionable intelligence that is common from big data analysis. While it is great to have tools that can analyze data, the key is to automate tasks so that the data is available more quickly and the analysis is sent to the right people on time. This will allow analysts to classify and categorize cyber threats without the long delays that could make the data irrelevant to the attack at hand.

The figure shows how big data analytics can improve on traditional cyber security and operations technology.

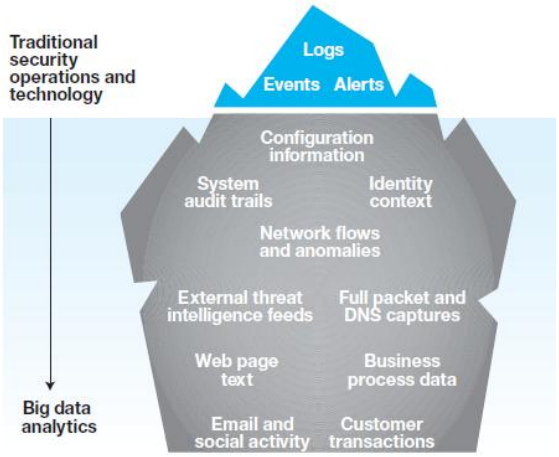

Figure 4: Big data analytics can improve on traditional cyber security and operations technology

Big data will also help analysts to visualize cyber attacks by taking the complexity from various data sources and simplifying the patterns into visualizations.

Being able to utilize the data in its raw format allows disparate data to be useful not only with what is happening now, but also with historical data. Using this historical data, you can create statistical baselines to identify what is "normal." You will then be able to determine when the data deviates from the norm. Sometimes it's easy to miss indicators when they are offered in real time; however, they may have new meaning when they are viewed over time.

This historical data can also create new possibilities for predictive models, statistical models, and machine learning. This gives the ability to predict future events. However, it's what you can do with this data, if anything, that can make the difference between being attacked or not. After all, data is just really information unless an action is taken towards improving cyber security. Being able to automatically respond to threats noticed in data, and also being able to have a high level of trust in the accuracy of the data is key to a big data security solution.

6. Preparation and Prevention

How can cyber security professionals hope to stay ahead of cyber criminals who have all the advantages of big data at their disposal? With the right preparation and prevention strategies. Here are a few:

- In the era of big data, awareness is the first line of defense against cybercrime. As one recent survey revealed, most cyber security professionals know that they need to worry about big data, but they don't always clearly understand what it means.
- Organizations should integrate customized processes and technical solutions geared to their specific risks and requirements to collect process, store, analyze and share data.
- Integrating big data analytics into a solid infrastructure to provide and develop security solutions is essential – as is employing an expert IT staff to deploy them.
- Strengthening cyber security teams with highly skilled data scientists and analytics experts may become increasingly essential.
- Future investments in technology should lean toward flexible, analytics-based solutions that can change as business requirements and security threats evolve.

Big data offers advantages to both the world of business and the underworld of hackers and cyber criminals. With continuous effort, investment in technology and awareness, cyber security professionals can win the battle against this and other complex challenges that new technology will surely bring.

7. **Conclusion**

Some might believe that big data will quickly solve the problems of the cyber security industry. The reality is that data and analytics will allow companies to identify anomalies and advanced attack vectors. Sentinel

One uses machine learning paired with cloud intelligence and automated responses to detect unusual activity and respond when you need it. Big Data for security will undoubtedly be of high interest and value for society to use, both now and in the near future. Applications for event security and subversive crime as examples seem to be very promising. The usability will increase when Big Data is not only used for enhanced situational awareness but also for better situational understanding. Big Data in security applications should adhere to standards and principles for responsible innovation. Handling Big Data with AI brings forward many ethical issues. This needs to be balanced by an expert ethical board. It should be considered for smaller organizations to have a centralized board to pull resources and have good quality capacity available.

References

[1]. Manyika, James, et al. (2011). "Big data: The next frontier for innovation, competition, and productivity."

[2]. Craigen, D., N. Diakun-Thibault, and R. Purse, (2014). Defining cybersecurity. Technology Innovation Management Review, 4(10).

[3]. Gary Locke, (2011). Cybersecurity, Innovation and the Internet Economy.

[4]. Kanungo, T., et al., An efficient k-means clustering algorithm: Analysis and implementation. IEEE transactions on pattern analysis and machine intelligence, 2002. 24(7): p. 881-892.

[5]. Challenger, J.R., et al., Efficiently serving dynamic data at highly accessed websites. IEEE/ACM transactions on Networking, 2004. 12(2): p. 233-246.

[6]. Anderson, R.J., Security engineering: a guide to building dependable distributed systems. (2010): John Wiley & Sons.

[7]. Church, A.H. and Dutta, S. (2013) The Promise of Big Data for OD: Old Wine in New Bottles or the Next Generation of Data-Driven Methods for Change. *OD Practitioner*, 45, 23-31.

[8]. Dean, J. and Ghemawat, S. (2008) MapReduce: Simplified Data Processing on Large Clusters. *Communications of the ACM*, 51, 107-113.

[9]. Renu, R.S., Mocko, G. and Koneru, A. (2013) Use of Big Data and Knowledge Discovery to Create Data Backbones for Decision Support Systems. *Procedia Computer Science*, 20, 446-453.

[10]. Alexandru Adrian TOLE, (2013). Big Data Challenges. *Database Systems Journal* vol. IV, no. 3/2013

[11]. S. Beisken, T. Meinl, B. Wiswedel, L. de Figueiredo, M. Berthold, and C. Steinbeck. Knime-cdk: Workflow-driven cheminformatics. BMC Bioinformatics, 14(1):257, 2013.

[12]. Greenwald, Glenn. No Place to Hide: Edward Snowden, the NSA, and the US Surveillance State. Metropolitan Books, 2014.

[13]. O'Driscoll, A., Daugelaite, J. and Sleator, R.D. (2013) 'Big Data', Hadoop and Cloud Computing in Genomics. *Journal of Biomedical Informatics*, 46, 774-781.

[14]. Wang, L. and R. Jones, Big Data Analytics for Network Intrusion Detection: A Survey. International Journal of Networks and Communications, 2017. 7(1): p. 24-31.

[15]. Nunan, D. and Di Domenico, M. (2013) Market Research and the Ethics of Big Data. *International Journal of Market Research*, 55, 505-520.

[16]. Tansley, Stewart, and Kristin Michele Tolle, eds." The fourth paradigm: data-intensive scientific discovery." (2009).

[17]. Lazer, David, et al." Life in the network: the coming age of computational social science." Science (NewYork, NY) 323.5915 (2009): 721.