IMPLEMENTATION OF CRITICAL INFORMATION INFRASTRUCTURE PROTECTION TECHNIQUES AGAINST CYBER ATTACKS USING BIG DATA ANALYTICS

AUTHORS:

Torty Vincent
&
Udoyen Prince

ABSTRACT

Big Data analytics is the act of analysing data to discover hidden patterns, trends, preferences, and other important information in order to detect infiltration, prevent fraud, and possibly make the right judgements. This study was carried out to investigate the implementation of big data analytics as a technique for information protection against cyber attacks. This study specifically examined the extent of big data analytics implementation and the challenges militating the adoption and full implementation. The survey research design was employed and a total of one hundred and twenty-one staff members of the Joint Admission and Matriculation Board (JAMB) and the Independent National Electoral Commission (INEC) were enrolled in the study. The instrumentation of questionnaire was used to elicit information from the study participants. The data collected were analysed using the binary logistic regression. Findings from the study revealed that big data analytics provides better advantage to information protection against cyber attacks. Also, the findings showed that to a significant extent, big data analytics has not been full adopted and implemented by institution and organizations in Nigeria. This gap was detected in this study to be a result of high cost of hiring expert personnel, time and large quantity of silos. These factors indirectly affects the decision of top management regarding the adoption and implementation of big data analytics. This study recommends the adoption and implementation of big data analytics by organizations and institutions in Nigeria. Adequate training should be provided for staff members whose job description involves interacting with big data.

Citation: ***Torty Vincent & Udoyen Prince (2021): Implementation Of Critical Information Infrastructure Protection Techniques Against Cyber Attacks Using Big Data Analytics.***

CHAPTER ONE
INTRODUCTION

1.0 Introduction

In 2016, BT, the telecoms company that owns and maintains the physical infrastructure that makes up the UK's broadband network, experienced an outage of a portion of its broadband services, causing hundreds of thousands of customers (including businesses) to lose Internet and phone connections for approximately two hours (Williams, 2016). This was the largest and most extensive network breakdown in years, according to the news item. Despite the fact that the corporation denies it and blames the network outage on a malfunctioning router, it has been speculated that the outage was caused by a cyber-attack. Whether BT's explanation for the outage (a malfunctioning router) is correct, the point is that it is definitely feasible to conduct a cyber-attack that may bring an organization's infrastructure down. What if this was a cyber-attack that affected more routers and lasted days rather than hours? For instance, the devastating effects on the country's economy, not to mention the lives lost when emergency services lost communication? On October 19, 2020, IBM researchers discovered vizom, a new type of stealthy malware that targets Brazilian account holders using remote overlay assaults according to Guillermo (2014) as cited in Brewer (2021). It is currently being used in a campaign in Brazil that aims to compromise bank accounts through online financial services. Due to the coronavirus pandemic, Vizom poses as a popular video conferencing software, which is now critical to businesses and social life (Brewer, 2021). The problem with such attacks is that they can eventually lead to a cascading failure of inter-bank funding, triggering a tipping point for a broader systemic liquidity crisis. In both of these scenarios, the organizations' operations are so intertwined with other organizations in their respective countries that their failure will inevitably trigger a domino effect, causing these other or related organizations to fail. As a result, the safeguarding of such infrastructures, also known as critical information infrastructures, is seen as a national security issue.

1.1 Background of the study

Cyber-attacks are constantly making headlines, putting countries, industries, and businesses at danger of security breaches. With society's reliance on technology and the introduction of the internet of things, things could get even worse. Cyber

criminals are growing more smart and knowledgeable, as seen by the fatal software they use to attack businesses. In the year 2020, hackers used stealthy malware to infiltrate Solar breezes (a United States-based firm that provides network monitoring and other technical services to thousands of companies, including government agencies), and injected malicious code to the firm's software system. Companies utilize the Orion system to manage their information technology resources. The code provided a backdoor into the customer information technology system, which hackers used to spy on businesses, organizations, and government agencies. Because critical information is exposed to the hackers, a hack of this magnitude has a global impact.

Information on healthcare, the electricity grid, disease management, and military operations that might be used to destroy a country. How does one protect against such assaults? Is big data analytics the way to go? We've seen a significant rise in data volume over the previous few years. Global IP traffic reached an estimated 1.2 zettabytes in 2016, according to Cisco Systems. Global IP traffic refers to all digital data that travels over an IP network; it is expected to exceed 20 zettabytes by the end of 2021. Data is collected from a variety of sources, including contracts, call centers, social media, and phones. Interactions between faxes, for example. This data could be very useful in detecting fraud. Large corporations are increasingly using big data analytics for cyber-security and defense because it allows them to see bigger and clearer pictures when detecting threats. As a result, a study of the effectiveness of big data analytics – for cyber-attack detection will be conducted in this research. This would be done by looking at the success rate of employing the technology to detect sophisticated and stealthy cyber-attacks like Advance Persistent Threats (through a survey By questionnaire).

Given that stealthy malware is designed to go undetected and that an attack can compromise a computer system in seconds (Brewer, 2015), the term "effectiveness" is defined as: having a detection speed of seconds, minutes, or hours, but no more than a day, as a day may be too late; being able to detect stealth attacks significantly more often than not - at least 75%

1.2 Problem Statement

The internet is a global network of interconnected systems which serves billions of users worldwide. Its popularity and rapid growth have come at an expensive cost, i.e., loss of information and resources due to cyber threats and attacks.

The first cyber crime was reported in 2000 and infected almost 45 million internet users (Message Labs Intelligence, 2010).

Over the few past years cybercrimes have increased rapidly with cyber criminals continuously exploring new ways to circumvent security solutions to get illegal access to computer systems and networks. Some important cyber attacks includes spamming, Search Poisoning, Botnets, Denial of Service (DoS), Phishing, Malware, hacking, etc

The Importance of protecting Critical Information Infrastructure cannot be overemphasized due to the catastrophic nature of such attacks to governments, attacks of such manner can be devastating and lead to a domino effect of disaster. This cyber-attacks often disguises in form of stealthy malware in attacking critical information sectors such as defense, food and agriculture, financial services, oil and gas, public health care, transportation etc. either to steal information or disrupt the normal operations of a government.

The extensive damage caused by these cyber attacks has lead to the design and implementation of cybersecurity systems. Cybersecurity refers to the techniques, processes and methodologies that are concerned with thwarting illegal or dishonest cyber attacks such as hacking, spamming, SQL injection, etc in order to protect one or more computers on any type of network from any type of damage.

This research proposes to address how critical information infrastructure can be protected against cyber-attacks using big data analytics.

1.3 Aims and Objectives of the study

Main aim of this study is to investigate the implementation of critical information infrastructure protection techniques against cyber attacks using big data analytics. Specifically, the study seeks to:

1. Investigate the efficacy of big data analytics as a protection technique.
2. Examine the extent of big data analytics implementation in government agencies in Abuja.
3. Elucidate on the challenges in implementing big data analytics as a protection technique.

1.4 Research Questions

The following questions guide this research:

1. How effective is big data analytics as a protection technique?

2. What is the level of big data analytics implementation in government agencies in Abuja?
3. What are the challenges militating the implementation of big data analytics protection technique?

1.5 Research hypotheses

Hypothesis refers to an experimental statement, tentative in nature, showing the relationship between two or more variables. It is open to test and can be accepted or rejected depending on whether it agrees or disagrees with the statistical test.

The study will test the validity of the following null hypothesis:

$H0^1$: Big data analytics is not effective as an information protection technique.

$H0^2$: There is no significant implementation of big data analytics in government agencies.

$H0^3$: There are no significant challenges impeding the implementation of big data analytics.

1.6 Significance of the study

Big data analytics as a cyber attack information prevention technique is a tool capable of curbing cybercrime due to the fact that it focuses on studying trends or patterns in which this attacks occur which in turn give organizations how to protect critical information and data. The greatest threat to network security procedures is that everyday hackers develop new malicious software and hacking techniques and no single software can practically keep up with the amount of threat. The aftermath of an initial breach in a system's network is often not helped by modern cyber security measures because of the way this cybersecurity measures are designed.

This study will be of immense benefit to both private and public agencies to first come to the understanding of the height of havoc cyber threats could cause to their databases. This study will also help organizations to identify the various information protection techniques to apply in other to combat and secure their data from cyber theft.

This study will further introduce to organizations under different sectors in Nigeria such as the banking sector, educational sector, insurance etc the benefit that comes with the adoption and implementation big data analytics as a prevention technique against cyber attacks such as helping organizations in providing the path in revolutionary transformations in several fields like inventions, marketing statistical status, etc. Helping big organizations in analyzing big data to achieve good raw data

from it. It makes work easy and examines all the information available and provides only the required data needed by the organization.

This study will further add to existing literature on this study topic and as well serve as a reference material to students, scholars and researchers who may which to carryout further study on this topic or related domain.

1.7 Scope of the study

This study focuses on investigating the efficacy of big data as a protection technique. Also, this study will look into the extent big data is being implemented in government agencies. The study will further examine the challenges countered in implementing big data analytics as a protection technique.

Furthermore, the findings of this study will be restricted to the government agencies due to their high need of information protection against cyber theft. Joint Admission and Matriculation Board (JAMB) and Independent National Electoral Commission (INEC), Abuja serve as the enrolled participants for this study.

1.8 Limitation of the study

This study focused mainly on studying Big Data Analytics (BDA) as a major information infrastructure protection techniques against cyber attacks instead of exploring and evaluating other information infrastructure protection techniques that could as well serve same purpose as that of Big Data Analytics (BDA).

Also, this study focused mainly on the theoretical aspect of Big Data Analytics (BDA) as a protection technique against cyber attacks instead of carrying out its practical by designing a database, try any of the cyberattacks on the developed database, and test-run it using BDA to confirm if it will protect the database from the attacks.

Furthermore, the respondents of this study was another limitation to this work because the study was not carried out in many or all sectors of the Nigeria economy in order to generate more valid facts for better conclusion for this study.

CHAPTER TWO

LITERATURE REVIEW

2.1 INTRODUCTION

Finding dynamic or proactive security measures from data analytics is what cyber security analysis is all about. When network traffic is monitored in order to detect compromise before a real danger arises, this is one example of this. When it comes to assaults and threats, no infrastructure or organization can predict the future, but with the correct security analytic tools in place to monitor security events, it is possible to detect a danger before it arises or has a chance to create havoc.

Literature review refers to the critical examination of state of knowledge including substantive findings as well as theoretical and methodological contribution to a particular topic. In line with this definition, the literature reviewed revolved around the exploration of the intrinsic meaning of variables under study.

Our focus in this chapter is to critically examine relevant literature that would assist in explaining the research problem and furthermore recognize the efforts of scholars who had previously contributed immensely to similar research. The chapter intends to deepen the understanding of the study and close the perceived gaps.

Precisely, the chapter will be considered in three sub-headings:

- Review/Explanation of important/relevant terms and technologies
- Review of Similar existing systems/previous related works
- Identification of gap from existing systems reviewed and solution to be proffered by this project

2.2 REVIEW/EXPLANATION OF IMPORTANT/RELEVANT TERMS AND TECHNOLOGIES

2.2.1 Concept of Critical Information Structure

Critical information infrastructure is described by Aladenusi (2015) in his presentation at the Nigeria Computer Society's 12th international conference as those ICT infrastructures that are dependent on core assets that are important for the running of the organization. He went on to say that if such assets are compromised, it

has a disastrous effect on national security, government, the economy, and the country's overall status.

Food and agriculture, dams, financial services, oil and gas, commercial facilities, communication, defense, emergency services, power and energy, government and facilities, information technology, healthcare, transportation systems, and water and sanitation are among the 15 industry sectors defined as critical information infrastructure in Nigeria, according to Aladenusi (2015).

The importance of critical infrastructure in nation-building is demonstrated by the fact that critical information infrastructures are interdependent on a large number of services and infrastructure, and the failure of any of these CII infrastructures causes a catastrophic domino effect that negatively impacts other services.

2.2.2 Concept of Big Data

Big data is data that is too complicated to be managed, searched, or analyzed using typical data storage systems, algorithms, or query techniques (MessageLabs Intelligence, 2010). The three V's define the "complexity" of big data:

1) volume - refers to the information of data held in terabytes, petabytes, or even exabytes (10006 bytes).

2) variety – this refers to the coexistence of unstructured, semi-structured, and structured data, as well as

3) velocity — the rate at which big data is created. The fourth V, veracity, has been introduced by some academics to emphasize the necessity of keeping high-quality data within an organization.

Data from computer networks, telecommunication networks, banking, healthcare, social media networks, bioinformatics, E-Commerce, surveillance, and other sources are some of the most common sources of big data transactions.

According to Cisco, global IP traffic will surpass 1000 Exabytes (1 zettabyte) by 2016. (Cisco, 2015). To put the size of the data being discussed in context, one zettabyte is the same size as the Great Wall of China (Arthur, 2011). Big data is the term for this avalanche of data. Big data, on the other hand, is about more than simply volume. It's also about velocity and variety. Variety refers to a wide range of data forms and forms, including video, audio, photos, text messages, and email, as well as data created by sensors and machines. The speed (including real time) at which these

data are created, processed, and transferred is referred to as velocity. Despite the fact that there are additional qualities, big data is primarily defined by the "three Vs" - volume, variety, and velocity (Gartner, 2012).

Big data may be characterized using the 5 Vs: volume, velocity, variety, veracity, and value, according to Ishwarapa and Amerada (2015), who used the healthcare business as an example. Every year, hospitals and clinics all over the globe create vast amounts of data in the form of patient records, test results, illness analyses, and other types of information. The Velocity of big data refers to the rate at which this data is created. The term "variety of data" refers to the numerous forms of data (structured, Semi structured and unstructured). The validity of this data refers to its correctness and consistency, and the value of big data refers to the analysis of all of this data to help the medical industry (faster disease detection, better treatment and reduced cost).

Hashem et al. provide a more detailed description of the nature of big data (2015). Big data is classified into five areas, according to them: data sources, content format, data storage, data staging, and data processing.

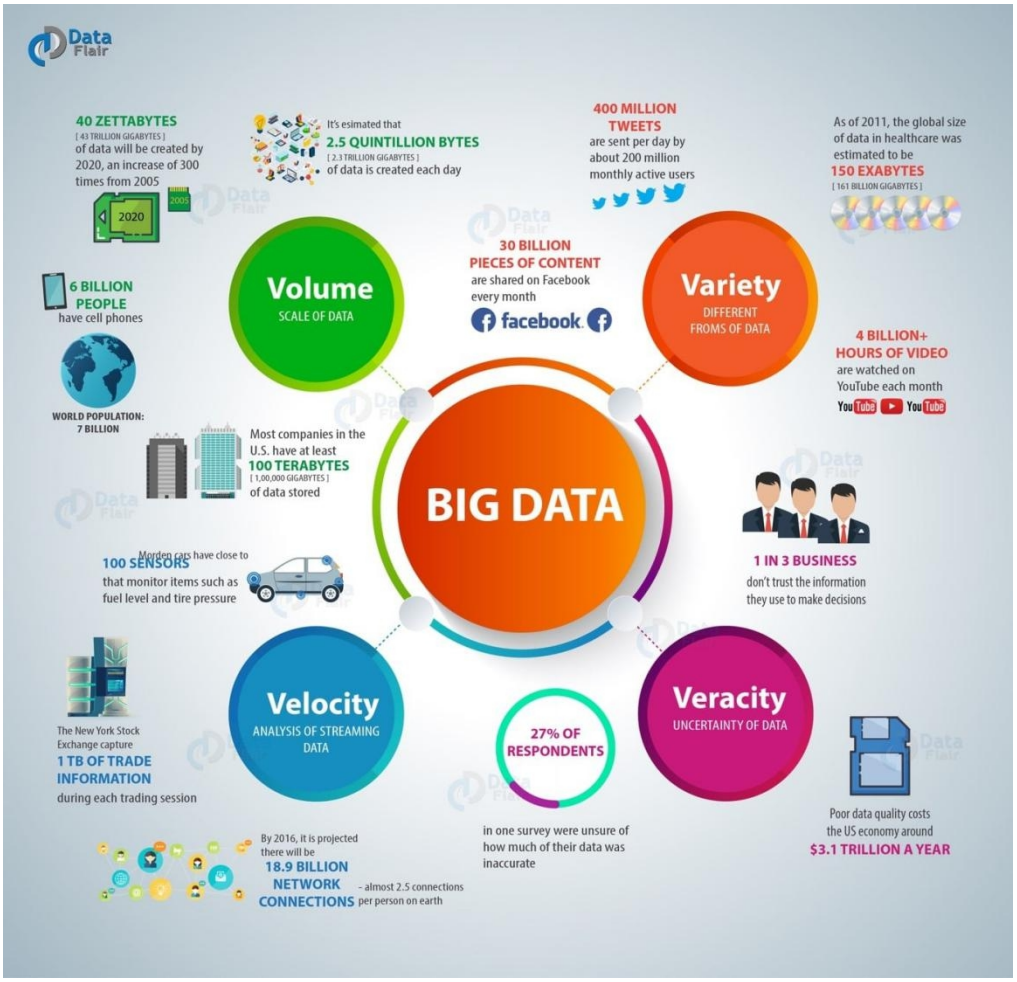


Fig 1: Diagram Showing Big Data (Everett, 2015)

2.2.3 Concept of Big Data Analytics

Big Data Analytics (BDA) is focused with extracting value from big data, i.e. nontrivial, previously unknown, implicit, and possibly beneficial insights. The "From Data to Decision" initiative [http://data-to-decisions.com] is driven by these insights, which have a direct influence on determining or altering existing corporate strategy. The notion is that big data contains patterns of usage, occurrences, or behaviors. BDA uses data mining techniques including Predictive Analytics, Cluster Analysis, Association Rule Mining, and Prescriptive Analytics to try to fit mathematical models on these patterns (Sathi, 2013). These approaches' insights are generally exhibited on interactive dashboards, which help businesses retain a competitive advantage, raise earnings, and improve their CRM.

It's vital to remember that the term "big" in big data is relative; even gigabytes of data might be considered "big" if it's not handled or queried properly. In this case, Apache's Hadoop framework, which is an open-source, entirely fault-tolerant, and highly scalable distributed computing paradigm, is a big help. The MapReduce algorithm (of Google) allows Hadoop to spread BDA jobs among commodity hardware nodes [Sathi, 2013]. At a high level, data is "mapped" in a domain-specific format before being processed at various nodes; the outputs from each node are then "reduced" to obtain the final output. Big firms like Yahoo!, Facebook, Twitter, eBay, and Amazon employ Hadoop, while IBM, Microsoft, Oracle, Talend, Cloudera, Greenplum, Hortonworks, and Datameer are now offering Hadoop-based BDA solutions. Teradata, HP (Vertica), Infobright, Aster Data, and ParAccel are among the companies that provide big data hardware designs (Curry, Kirda, Schwartz, Stewart, and Yoran, 2013).

In our increasingly digitized culture, big data analytics is gradually becoming a vital tool. It is utilized in a variety of fields, including artificial intelligence (AI), health-related research, and information security, as well as by big organizations to improve decision-making.

McAfee and Brnjolfsson (2012) utilize an example of real-time location data from users' smartphones to identify how many consumers were in the Macy's parking lot on Black Friday, the start of the Christmas shopping season in the United States, to demonstrate the relevance of Big data analytics. Analysts were able to estimate the retailer's sales based on this data even before the actual sales were recorded.

Big data has also been utilized to construct artificial intelligence (AI) systems that are better at doing jobs that previously could only be done by people when paired with machine-learning algorithms. IBM's Watson, for example, beat the finest wits in the game of Jeopardy in 2012. (Ferrucci, 2012). Driverless automobiles are another example of machine learning; though these automobiles have not yet overtaken humans, testing (on certain roads) suggest that they have learned the complicated skill of driving (Gibbs, 2014).

The argument is that big data analytics may be a great tool for businesses to make wiser and better decisions since it can provide a more accurate picture of any event even before it occurs. As a result, big data analytics is an ideal and effective technique for detecting cyber-attacks. When discussing the benefits of big data, Tankard (2012) indicated how this application - of employing big data analytics to identify cyber-attacks - may be done. He argues that businesses may use the massive volumes of data they've been gathering to look for cyber security threats like malware and phishing efforts.

There are a variety of technologies that are utilized to do big data analytics. Most specialists in this sector agree that the big data phenomena is still in its early stages, and this appears to be backed by the slew of new technologies – such as storage applications, machine-learning algorithms for analytics, and user interfaces – that are hitting the market today.

Big data analytics, according to Gillette (2016), is the act of studying enormous data sets encompassing a variety of data types in order to find patterns, market trends, and consumer behavior. Companies are embracing big data analytics solutions because the information gleaned from market trends and customer behavior is extremely beneficial in defending against cyber-attacks and driving overall corporate success. If this data is used well, it has the potential to make a significant difference.

In the football business, for example, teams utilize big data to collect information on player performance, such as peak speed, passes completed, shoots on target, and possession, to name a few. This data is now examined, which is where data analytics comes in. Data analytics gives useful information to the club by making sense of the data and providing valuable insights that can be utilized in game strategy, training programs, and player scouting.

Overall, the usage of big data technology aids in prediction and provides a clearer picture of events before they occur, allowing businesses to make better decisions in terms of threat detection and cybercrime prevention.

As a result, big data analytics is an ideal and effective technique for detecting cyber-attacks. When Kumar (2017) outlined the benefits of big data, she stated that firms may mine the massive volumes of data they have been gathering for possible cyber security incidents such as malware and phishing attempts.

Fig 2: Diagram Showing Big Data Analytics (Brewer, 2015)

2.2.4 Overview of Cyber-Crime and Cybersecurity

The meanings of cyberspace, cyber security, and cybercrime have evolved in tandem with technological advancements. It has been suggested that because computer crime may encompass all types of criminal behavior, a definition must stress the uniqueness, expertise, or use of computer technology. The internet's limitless expanse is referred to as cyber-space. It refers to the interconnected network

of information technology components that support many of today's communication technologies (Ibikunle, 2013). Cyber security refers to a set of tools, policies, security concepts, security protections, guidelines, risk management techniques, activities, training, best practices, assurance, and technology that may be used to secure the cyber environment, as well as an organization's and users' assets. Connected computer devices, staff, infrastructure, applications, services, telecommunications systems, and the totality of transmitted and/or stored information in the cyber environment are all assets of organizations and users (Ibikunle, 2013). Cyber security [www.whatis.com] aims to assure the accomplishment and maintenance of the organization's and users' security properties in the cyber environment [www.whatis.com]. The corpus of regulations put in place to defend the cyber realm is known as cybersecurity. However, as we become more reliant on the internet, we will surely confront new threats. The term "cybercrime" refers to a group of organized criminals who target both cyberspace and cyber security. Cyber criminals and nation-states, for example, pose a threat to our economic and national security. Nigeria's economic viability and national security are reliant on a broad network of interconnected and critical cyberspace networks, systems, services, and resources. The way we communicate, travel, power our houses, manage our economy, and access government services has all been revolutionized by cyberspace. Cyber-security refers to a set of technologies, techniques, and procedures for defending networks, computers, programs, and data from assaults, damage, and unauthorized access. The term "security" in the context of computer or cyberspace simply means "cybersecurity" [www.bbc.co.uk]. Ensuring cyber-security necessitates concerted efforts on the part of both people and the country's information infrastructure. The threat presented by cyber-security breaches is evolving quicker than we can keep up with it. It is impossible to focus attention just on one part of the breach since this would imply neglect and enable other components of the breach to proliferate. As a result, we've come to the conclusion that we need to tackle cyber security breaches as a whole. So, what exactly are these breeches?

Criminal action involving computers and the Internet is referred to as cyber-crime. This might range from stealing millions of dollars from internet bank accounts to downloading illicit music downloads. Non-monetary offenses like as generating and distributing viruses on other computers or publishing secret company information on the Internet are included in cybercrime. Identity theft is perhaps the most well-known type of cybercrime, in which criminals exploit the Internet to steal personal

information from other people [http://www.itu.int/en]. “A criminal activity involving an information technology infrastructure, including illegal access (unauthorized access), illegal interception (by technical means of non-public transmissions of computer data to, from, or within a computer system), data interference (unauthorized damaging, deletion, deterioration, or alteration,” according to [Laura, 1995].

Cyber security, according to the International Telecommunication Union (ITU), is a collection of tools, policies, security concepts, security safeguards, guidelines, risk management approaches, actions, training, best practices, assurance, and technologies that can be used to protect the cyber environment, organization, and users' assets. Connected computer devices, staff, infrastructure, applications, services, telecommunications systems, and the totality of transmitted and/or stored information in the cyber environment are all assets of organizations and users. Cyber security aims to ensure the accomplishment and maintenance of the organization's and users' security properties in the cyber environment - the internet - against relevant security dangers (ITU, 2011).

Cyber security is a collection of technologies, procedures, and practices aimed at preventing attacks, damage, and illegal access to networks, computers, programs, and data. Availability, Integrity (which may include authenticity and non-repudiation), and Confidentiality are the general objectives of Cyber Security, according to the ITU (Ravi, 2012).

Fig 3: Diagram Showing Cyber Security (Abdullah, 2019)

Fig 4: Diagram Showing Cyber Security Process (Everett, 2015)

2.2.5 Goals of Cyber Security

The following are the objectives of Cyber-security according to (Ibikunle, 2013).

- To help people reduce the vulnerability of their Information and Communication Technology (ICT) systems and networks.
- To help individuals and institutions develop and nurture a culture of cyber security.
- To help understand the current trends in IT/cybercrime, and develop effective solutions.
- Availability.
- Integrity, which may include authenticity and non-repudiation.
- Confidentiality.

2.2.6 E-crimes that are Peculiar to Nigeria

E-crime is without a doubt a public relations nightmare for Nigeria. Cybercrime is a cause of national anxiety and disgrace (Ibikunle, 2013). The Internet provides limitless economic, social, and educational options. However, as we can see

with cyber-crime, the Internet comes with its own set of perils. The examples given here vary from bogus lotteries to the most sophisticated online frauds. Elekwe, a chubby-faced 28-year-old guy who had been jobless for two years despite holding a diploma in computer technology, acquired a fortune via the fraud. The leader of a fraud group in a business area persuaded him to Lagos from Umuahia. As a result of his activities, he now has three stylish automobiles and two homes. Security officials in Ghana apprehended four Nigerians suspected of running a "419" scam on the internet to defraud unwary overseas investors in July 2001. Prospective investors are said to have lost several millions of dollars as a result of their actions. Two young lads were recently detained after purchasing two computers listed on a woman's website under false pretenses. Government officers apprehended them at the moment of delivery. Mike Amadi received a 16-year sentence for creating a website that advertised lucrative but bogus procurement contracts. An undercover agent acting as an Italian businessman nabbed the man impersonating the EFCC Chairman. Amaka Anajemba, who was sentenced to 212 years in jail, perpetrated the largest international con of all. She was also sentenced to refund $25.5 million of the $242 million stolen from a Brazilian bank with her cooperation.

A recent internet scam case involving a 24-year-old Yekini Labaika of Osun State origin in Nigeria and a 42-year-old nurse of American origin, Thumbelina Hinshaw, was reported in the Sunday PUNCH newspaper on July 16, 2006, involving a 24-year-old Yekini Labaika of Osun State origin in Nigeria and a 42-year-old nurse of American origin, Thumbelina Hinshaw, in search of a Muslim lover to marry was The young guy fooled the victim by pretending to be Phillip Williams, an American Muslim working for an oil business in Nigeria, and promising to marry her. He invented questionable methods to defraud the victim of $16,200 and several expensive goods. After being found guilty of eight crimes against him, the fraudster was sentenced to a total of 1912 years in prison. These kind of incidents are becoming more common. Several young males continue to effectively carry out these illicit activities, robbing unsuspecting persons and organizations (Longe & Chiemeke, 2008). According to a recent research, Nigeria loses roughly $80 million each year due to software piracy. The findings of a study done on behalf of the Business Software Alliance of South Africa by Institute of Digital Communication, a market research and forecasting organization located in South Africa. Nigerian money promises were the fastest growing Internet hoax in 2001, according to the American

National Fraud Information Centre, with a growth rate of up to 90%. Nigerian cybercrime effect per capita was likewise considered extremely high by the Center, according to “The Economic Times” news broadcast in, September 11, 2004.

The majority of those participating are between the ages of 18 and 25, and they live in metropolitan areas. The internet has aided in the modernization of deceptive behaviors among youngsters. The teenagers engaged regard online scamming as a widely recognized technique of obtaining financial support. The emergence of the online criminal subculture has been aided by the governmental leadership's corruption. [Adebusuyi,2008] The priority placed on money gain has been a prominent element in the engagement of youngsters in online fraud.

2.2.7 Concept of Cyber-Attack

A cyber-attack, according to Farhat et al. (2011), is an attack launched from a computer against a website, computer system, or individual computer (collectively, a single computer) that compromises the computer's or information stored on it's confidentiality, integrity, or availability. They went on to say that cyber-attacks can take a variety of forms, including:

1. Spamming: Spamming is sending unsolicited bulk messages to multiple recipients [Banday and Qadri, 2006]. By 2015, the spam volume is forecasted to be 95% of all email traffic [Abdullah, 2019]. Munging, access filtering and content filtering are important anti-spam techniques. Munging makes email addresses unusable for spammers, e.g., abc@gmail.com munges as “abc at gmail dot com”. Access filtering detects spam based on IP and email addresses while content filtering recognizes predefined text patters in emails to detect spam.

2. Search Poisoning: Search poisoning is the dishonest use of Search Engine Optimization techniques to falsely improving the ranking of a webpage [Perdisci and Lee, 2011]. Typically, frequent search keywords are used to illegally direct users towards short-lived websites. The first poisoning case was reported in 2007 [Vaas, 2007], followed by many others.

3. Botnets: Botnets are networks of malware-infected compromised computers managed by an adversary, according to Stone-Grass, Cova, Cavallaro, Gilbert and Szydlowski, [2009]. Attackers use bot software equipped with integrated command and control system to control these zombies (bots) and group them into a network called the bot net [Bailey, Cooke, Jahanian,Xu and Karir,2009].

4. Denial of Service (DoS): A DoS attack makes a system or any other network resource inaccessible to its intended users. It is launched by a large number of distributed hosts, e.g., bot net. Many defensive techniques such as intrusion detection systems, puzzle solution, firewalls etc. have been developed to prevent DoS attacks [Stone-Grass, Cova, Cavallaro, Gilbert and Szydlowski, 2009].

5. Phishing: Phishing fraudulently acquires confidential user data by mimicking e-communication [Jakobsson and Myers, 2007], mainly through email and web spoofing [Shi and Saleem, 2012]. In email spoofing, fraudulent emails direct users to fraudulent web pages which lure to enter confidential data. In web spoofing, fraudulent websites imitate legitimate web pages to deceive users into entering data. Many anti-phishing solutions are in corporate use to counteract this threat.

6. Malware: Malware is software programmed to perform and propagate malicious activities, e.g., viruses, worms and Trojans. Viruses require human intervention for propagation, worms are selfpropagating, while Trojans are non-self-replicating. Damage from malware includes corruption of data or operating system, installation of spyware, stealing personal credentials or hard disk space etc [Shi and Saleem, 2012].

7. Website Threats: Website threats refer to attackers exploiting vulnerabilities in legitimate website, infecting them and indirectly attacking visitors of these sites. SQL injections, malicious ads, search result redirection are the few techniques which are used to infect the legitimate sites (Abdullah, 2019).

8. Hacking: Silver-Greenberg, Goldstein, and Perlroth (2016) stated that Hackers make use of the weaknesses and loop holes in operating systems to destroy data and steal important information from victim's computer. It is normally done through the use of a backdoor program installed on your machine. A lot of hackers also try to gain access to resources through the use of password hacking software. Hackers can also monitor what u do on your computer and can also import files on your computer. A hacker could install several programs on to your system without your knowledge. Such programs could also be used to steal personal information such as passwords and credit card information. Important data of a company can also be hacked to get the secret information of the future plans of the company.

The extensive damage caused by these cyber attacks has lead to the design and implementation of cybersecurity systems. Cybersecurity refers to the techniques, processes and methodologies concerned with thwarting illegal or dishonest cyber

attacks in order to protect one or more computers on any type of network from any type of damage [wikipedia, 2021]. The important goals of cybersecurity are:
1) securely obtain and share information for accurate decision-making, 2) find and deal with vulnerabilities within applications,
3) prevent unauthorized access and
4) protect confidential information. Some of the well-known cybersecurity solutions are being provided by Accenture, HP, Invensys, IBM, EADS, CISCO, Unisys etc.

In contrast to the old strategy of detecting bad signatures, the focus of cybersecurity has lately switched to monitoring network and Internet traffic for the identification of illicit behaviors. Traditional cybersecurity, in particular, focuses on capturing malware by analyzing incoming traffic against malware signatures, which only identify limited-scope threats that have been seen before. Furthermore, signature development lags significantly behind the evolution of cyber assault strategies. Hackers may quickly render tactics like intrusion detection systems, firewalls, and anti-virus software worthless. This situation has grown much more critical given the presence of big data within computer networks — petabytes and exabytes of data being transmitted daily between nodes makes it very easy for hackers to access any network, mask their presence successfully, and inflict significant harm.

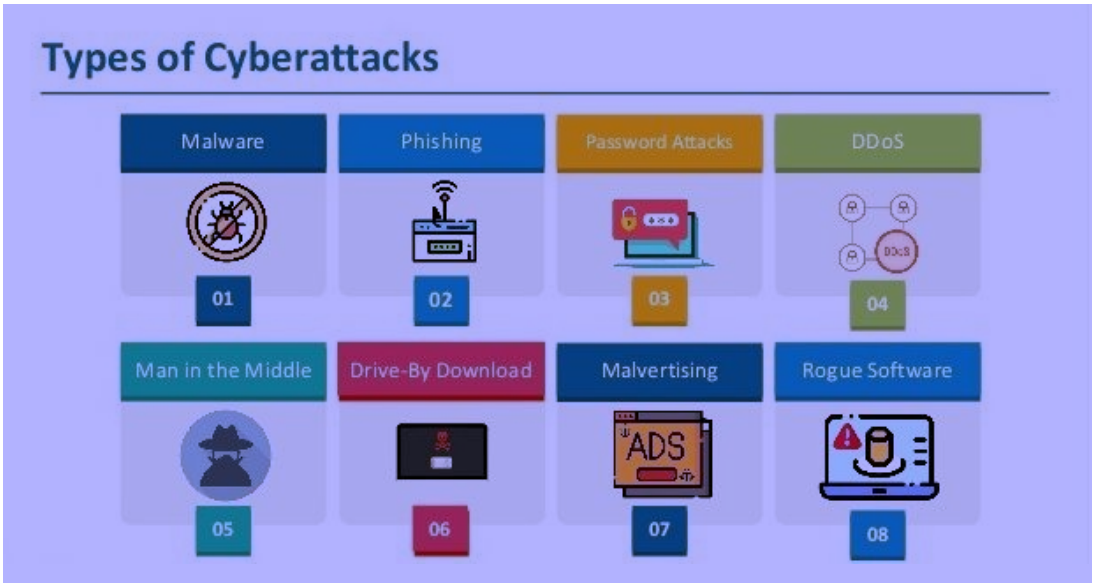


*Fig 5: Diagram Showing Types of Cyber-attacks (*Silver-Greenberg, Goldstein, and Perlroth, 2016)

2.2.8 Big Data Technologies

Big data analytics entails the meticulous examination of the (big) data in order to obtain meaningful and useful information. However, given its nature, it is obvious that any kind of examination of big data would be tricky Therefore, innovative solutions have been required to make the mining of meaningful information from big data less challenging and as easy as possible. There a lot of big data technologies and tools in the market today which shows a lot of progress has been made, According to simplilearn (2018) this data technologies can be divided into the following:

1. Data storage and management:

No SQL database such as Mongo dB, Cassandra, neo4j and Hadoop are popular for data storage and management tools

a. Data cleaning:

Data needs to be cleaned up and well structured (data is reshaped into usable datasets) , example of this tools are Microsoft excel and open refine

b. Data mining:

This involves discovery insights within a database, example of such tools are Teradata and Rapid miner

c. Data visualization tools:

This tools conveys complex data insights in a pictorial way that wat it is easy to understand, common tools are tableu and plotly

d. Data reporting:

Data reporting entails collecting and submitting data which gives rise to adequate analysis Power bi tools are used

e. Data analysis:

This entails asking questions are finding the answers in data, example of this tools are hive, map reduce and hive

Fig 6: Diagram Showing Big Data Technologies (Brewer, 2015)

2.2.9 Threat detection with Big Data Analytics

Antivirus applications, network IDS/IPS, host IDS/IPS, network device events, logging, FIM and whitelisting, and SIEM are the traditional categories for detecting and stopping cyber-attacks. Although these technologies are beneficial in many respects, they are mainly ineffectual against today's covert cyber-attacks. This is because, in addition to working autonomously, these systems create a large volume of data that is difficult and time consuming to analyze without the right tool; as a result, it is possible to overlook crucial cyber-attack occurrences (Shackleford, 2016).

This shows that these disparate systems may be made more efficient and successful with the proper deployment of the relevant tool (such as BDA technology), which can filter through data considerably faster. According to Laitan (2014), an organization that used to employ roughly 35 people to monitor 135,000 data loss prevention (DLP) alerts each day was able to lower that number significantly by adopting a big data analytics solution.

How can a BDA system help an organization detect cyber-attacks or threats? (Laitan, 2014) describes three basic techniques to establishing a big data analytics tool for cyber security, each of which is depending on the data source and analytics setup (canned or ad hoc).

According to the research (Laitan, 2014), the first technique is making current systems – such as SIEM, DLP, and DAP – more intelligent and less noisy so that only the most hazardous cyber-attacks (such as APTs) are recognized and isolated. This typically signifies that the big data system's analytics settings are pre-programmed.

Furthermore, the data to be used will come from the organization's databases, servers, and apps.

The data (for analytics) is gathered from both internal and external sources (such as internet and mobile activities) in the second technique, and the analytics settings are customized (or ad hoc). This implies that businesses may specify their own search parameters, and searches for harmful activity can be done in a Google-like way using some of the big data analytics tools in use.

The analytics in the third way is mostly based on external data on risks and the behaviors of various bad actors. This implies that the big data analytics system is built to scour the Internet (both dark and light) for hostile activity directed at businesses.

In summary, Sullivan (2016) defines a big data technology for cyber security as having the following critical features:

- It must have the ability to scale, as smoothly as possible, to accommodate the increasing size of the security data being collected (from both internal and external sources) without losing performance in its functionalities. This means that the analytics engine must be able to handle the data as it scales horizontally across distributed storage systems. Also, the storage systems it uses must be persistent with low data latency. In other words, the database must be capable of keeping copies of the original data even after it has been modified, and data access must be quick.
- It must have a reporting and visualization function which will allow the information (after the analytics – canned or ad hoc) to be presented in a way that will be useful and meaningful to security analysts.
- The source of the data (for the analytics) must be in context. In other words, analysing weather data for cyber-attack events might not be a good idea. Using just any data might result in higher than necessary false positives, or even worse, it might result in false negatives. This means that the source of the data for the analytics is of absolute importance.

2.2.10 Security Analytics With Big Data Analytics

Compared to traditional approaches, security analytics provides a “richer” cybersecurity context by separating what is “normal” from what is “abnormal”, i.e., separating the patterns generated by legitimate users from those generated by suspicious or malicious users

1. Big Data Sources for Security Analytics

The concept of "data" for security analytics is expansive, and can be categorized into passive and active sources [Krishnan, 2016]. Passive data sources can include:

- Computer-based data, e.g., geographical IP location, computer security health certificates, keyboard typing and clickstream patterns, WAP data.
- Mobile-based data, e.g., GPS location, network location, WAP data.
- Physical data of user, e.g., time and location of physical access of network.
- Human Resource data, e.g., organizational role and privilege of the user.
- Travel data, e.g., travel patterns, destinations, and itineraries.
- SIEM data, e.g., network logs, threat database, application access data.
- Data from external sources, e.g., rogue IPs, external threats.

Active (relating to real-time) data sources can include:

- Credential data, e.g., user name and password
- One-time passwords, e.g., for online access
- Digital Certificates
- Knowledge-based questions, e.g., "what is your typical activity on Saturdays from 3 pm to 6 pm?"
- Biometric identification data, e.g., fingerprint, facial recognition, voice recognition, handwriting recognition
- Social media data, e.g., Twitter, Facebook, internal office network etc.

Analytics applied on passive and active sources collectively will provide a 360^0 view of network traffic, e.g., by singling out an abnormal behavior in the access pattern of a given user. Appropriate prevention techniques can then be applied, e.g., lock accounts, quarantine, modify network settings, multiple authentications, alert on an on-going fraud etc.

2. Security Analytics Model

A security analytics model has the following primary characteristics according to [Curry, Kirda, Schwartz, Stewart, and Yoran, 2013]:

a. Diverse Data Sources: Data can be acquired from many sources (as mentioned in (A). The frequency and complexity of data sources should have no effect on the final BDA outputs.

b. Enterprise-level Data Warehouse: All network security data should be stored in an enterprise warehouse, with network performance measures stored in fact tables, and query data in dimension tables (time, location, customer etc.). This is important as

data mining (BDA) solutions are now strongly coupled with a warehouse, e.g., Oracle, SQL Server, DB2 etc. Different managers (requiring different analytics) can generate their own security analytics warehousing process.

c. ETL Tools: These tools can help perform ETL at two levels, i.e., at a generic level over a complete data source, and at specific level over selected data in order to apply one or more BDA techniques on it. Some common names are Talend, Informatica, Pentaho etc.

d. BDA Engines: These engines combine Hadoop, sophisticated hardware and analytics softwares to process real-time network streams. As mentioned above, these engines are typically coupled with a warehouse allowing network managers to employ warehouse resources, e.g., ETL, dashboards, data maintenance etc. for the BDA tasks as well.

e. Monitoring Systems: These system monitor network traffic and compare it with behavior and risk models discovered by BDA, or those given by security experts based on their experience.

f. Advanced Security Controls: These are used to impart appropriate security measures in real-time, e.g., additional user authentication, blocking highly suspicious transactions, calling customers while the transaction is in progress etc.

g. Interactive Dashboards: These apply Business Intelligence techniques to display the real-time cybersecurity status using interactive tools, e.g., graphs, charts, tables etc. A lot of research work is being carried out to develop novel visualization techniques which are at par with BDA outputs [Sathi, 2013].

h. Robust Security Infrastructure: This should be able to facilitate efficient communications between different locations or regions along with efficient execution of queries using BDA architectures.

i. Big Data Integration Infrastructure: This will facilitate integration of data from diverse number of data sources, which are quite common as mentioned in (A).

3. Security Analytics for Threat Detection

The largest application of security analytics is in threat monitoring and incident investigations, which is of major concern to both financial and defense institutions. The focus is on discovering and learning both known and unknown cyber attack patterns, which is expected to remarkably influence the efficiency of identifying hidden threats faster, track down attackers and predict future attacks with

increasing accuracy (minimum false positive rate). Some examples of how BDA can help with respect to different security dimensions are as follows:

a. Network Traffic: Detecting and predicting suspicious sources and destinations, along with abnormal traffic patterns.

b. Web Transactions: Detecting and predicting abnormal user access patterns, particularly in the usage of critical resources or activities.

c. Network Servers: Detection and predicting abnormal patterns related to server manipulation, e.g., abnormal or sudden configuration changes, non-compliance with pre-defined policy etc.

d. Network Source: Detecting and predicting abnormal usage patterns of any machine, e.g., related to the type of data the source transmits, processes and receives.

e. User Credentials: Detecting anomalies with respect to a user, or a group of user, not complying with its inherent access behavior, e.g., abnormal access time or transaction amount.

These activities are bringing a revolutionary change in the domains of security management, identity and access management, fraud prevention and governance, risk and compliance, e.g., through centralization of threat data and alert management system, correlating hundred thousands of network events per second, real-time continuous assessment of risk, distinction between legitimate and abnormal activities, along with appropriate prioritization of risks.

4. Steps for Implementing A Security Analytics Solution

According to [Curry, Kirda, Schwartz, Stewart, and Yoran, 2013], the following steps should be followed in adopting a security analytics solution :

a. Develop Security Analytics Business Strategy: The first step is to create a business strategy for implementing the analytics platform. For this, CIOs and CTOs need to initially build the domain knowledge by looking at some of the successful analytics-based security solutions (detailed in next section) to determine their feasibility, impact and value for their own organizations.

b. Participate in Analytics Trainings and Workshops: The real impact of threat detection can only be realized by C-level executives if they understand the technical details of BDA. Hence, attending BDA workshops and training sessions is particularly recommended. Especially because several opensource analytical tools, e.g., Rapid Miner, allow users to experiment with a host of BDA techniques through simple drag-and-drop functionality. Hands-on experiments with these tools on

security data sets can provide in-depth knowledge of the expected BDA outputs and aid in developing the analytics strategy.

c. Implement a Centralized Data Management Infrastructure: This should allow seamless integration of network streams from various sources, along with tools to support ETL, stream processing, data warehousing, data storage and query execution mechanisms. The infrastructure should be flexible to extension as well as to modification, e.g., a change of ETL tools.

d. Implement an Analytics Platform: This should support experimentation with a diverse number of BDA tools, techniques and algorithms. An efficient communication medium should be provided with the data repository to acquire the data and to store BDA outputs.

e. Hire Data Scientist as Consultant: As acquiring BDA outputs is typically lengthy task requiring background knowledge of many domains, it is advised to hire a data scientist as a consultant who can guide the execution of the BDA process at different stages.

f. Implement a "Network Monitoring" Layer: This layer monitors the network streams at run time by employing the set of mathematical models that have been mined by BDA. It can also employ any type of security knowledge given by the system designers based on their experience, or knowledge from some governmental database regarding new types of external threats. This layer should always be "live" as network streams are monitored 24/7. It intimidates the "Suspicion Alert" later if any suspicious behavior is detected.

g. Implement a "Suspicion Alert" Layer: This layer implements all the measures that can be taken to ensure cybersecurity, e.g., alert authorities of an ongoing suspicious transaction, lock access, cross-check user identity etc. It must be ensured that this module is always "live", even if the analytics platform is not being used at a given time.

h. How to Streamline Analytics with Current Workflow? This is the most important question facing C-level executives. Due to the uncertainty which accompanies BDA, it is advisable to complete the aforementioned steps within a minimum level of resources initially. These steps should also be implemented independently of the current workflow, with collaboration occurring only where needed. For instance, a team of 4-5 people along with the data scientist should be given 5-6 months to implement the analytics solution, as an organizational project.

Once the managers are satisfied with the outputs, steps can be taken to infuse the analytics platform gradually in the current workflow.

2.2.11 Big Data Mechanism In Analytics Cybercrime

For future prediction and micro-crime patterns, the crime analysis tool must be able to properly and effectively detect criminal trends. The predictive algorithm's foundation was recognized as the utilization of hotspot groupings. As a result, the Hotspot idea aids with crime prediction. They are used to detect places where crime occurrences are increasing so that appropriate police resources may be sent in specific places, depending on the obsolete events. Data extraction techniques, such as the Weka tool, Quick Tool, R Tool, KNIME, ORANGE, Tanagra, q-radar, and others, can be used to analyze crime data. The k-means assembly methodology is used to extract the data for the crime data analysis. Criminals now employ technical devices to carry out their crimes as a result of technological advancement. These digital data are utilized to investigate crimes.

The criminal data analyzed will aid in the prediction of hotspots. When there is disorganized or semi-structured data, the data used in the analysis and for forecasting the use of data mining is structured data, and data mining techniques cost a lot of time at such time. Obtaining criminal data, prepping it for a fast tool, and using it to conduct k-clustering for groupings. After you've obtained the clusters, examine them to see if you can forecast the crime. Renuka and Rajni (2013) Other data extraction approaches that may be used to analyze crime data and anticipate hotspots are also available. The other technology mostly consists of categorization, the Kassembly method (also known as the expectation algorithm), and so on. It's possible that using the k method improves the results compared to just using k-means clustering. Big Data Analytics can benefit from the KMeans algorithm. Because they require more organized data, these are some of the most conventional and time-consuming ways of crime mapping. [2013, Leong and Chan] The geographical distribution of crime data is likewise a difficult process, but it can now be accomplished with the help of programs like R. The data to be distributed to geographic regions will be greatly affected by several spatial geospatial programs that must be installed and executed using the R tool. These criminal data may be compiled with the help of modern technologies. The GA-based assembly (genetic algorithm) can also be used to analyze or conduct the assembly in Big Data Analytics. The R tool

is used to spatially disseminate the data. This tool may create a spatial representation of data that is geographically scattered. This utility comes with a number of packages that must be installed in order to complete the data dissemination process. This program may be used to do data analysis as well as other forms of distributed data [Nivranshu, Sana Mahajan, & Omkar, 2015].

An artificial neural network (ANN) is a collection of nerve cells or nodes (treatment components) that make predictions based on previously gathered or accessible data. When compared to other systems like Fuzzy Logic Series or Bayesian Network, the artificial neural network's prediction accuracy is generally quite high. The biggest disadvantage of the artificial neural network is that learning how to construct it takes time. [Nikhil & Setu, 2014] Finally, we may summarize the operations as follows: the data gathered will be disseminated mostly based on geographic location and group formation. The groups established with Big Data Analytics are examined in the second step. Finally, clusters are analyzed and fed into an artificial neural network, which generates a prediction pattern. The security authorities can utilize this pattern of forecasts to deploy resources to help reduce crime.

2.2.12 Challenges To Security From The Production, Storage, And Use Of Big Data

According to Guillermo [2014], the biggest challenge for big data from a security point of view is the protection of user's privacy. Big data often contains huge amounts of personal identifiable information, so the privacy of users is a huge concern.

Because of the amount of data stored, breaches affecting big data can have more devastating consequences than the ones we normally see in the press. This is because a big data security breach will potentially affect a much larger number of people, with reputational consequences and enormous legal repercussions.

When producing information for big data, organizations have to ensure they have the right balance between utility of the data and privacy. Before the data is stored it should be adequately anonymized, removing any unique identifier for a user. This in itself can be a security challenge as removing unique identifiers might not be enough to guarantee the data will remain anonymous. The anonymized data could be cross-referenced with other available data following de-anonymization techniques [Guillermo, 2014].

When storing the data, organizations will face the problem of encryption. Data can't be sent encrypted by the users if the cloud needs to perform operations over the data. A solution for this is to use "Fully Homomorphic Encryption" (FHE), which allows data stored in the cloud to perform operations over the encrypted data so new encrypted data will be created. When the data's decrypted, the results will be as if the operations were carried out over plain text data. So the cloud will be able to perform operations over encrypted data without knowledge of the underlying plain text data.
A significant challenge while using big data is establishing ownership of information. If the data's stored in the cloud, a trust boundary should be established between the data owners and the data storage owners. Adequate access control mechanisms are key in protecting the data. Access control's traditionally been provided by operating systems or applications restricting access to the information - this typically exposes all the information if the system or application is hacked [Everett, 2015].

A better approach is to protect the information using encryption that only allows decryption if the entity trying to access the information is authorized by an access control policy. An additional problem is that software commonly used to store big data, such as Hadoop, doesn't always come with user authentication by default. This makes the problem of access control worse, as a default installation would leave the information open to unauthenticated users. Big data solutions often rely on traditional firewalls or implementations at the application layer to restrict access to the information [Everett, 2015].

2.2.13 Best Practices For Managing Big Data In An Organization, From a Security Perspective

Big data's a relatively new concept so there isn't a list of best practices that are widely recognized by the security community. However, there are a number of general security recommendations that can be applied to big data according to [Guillermo, 2014]: These are:

- **Vet your cloud providers**

If you're storing your big data in the cloud, you must make sure your provider has adequate protection mechanisms in place. Check the provider carries out periodic security audits and agree penalties in case adequate security standards aren't met.

- **Create an adequate access control policy**

Create policies that allow access to authorized users only.

- **Protect the data**

Both the raw data and the outcome from analytics should be adequately protected. Encryption should be used accordingly to ensure no sensitive data is leaked.

- **Protect communications**

Data in transit should be adequately protected to ensure its confidentiality and integrity.

- **Use real-time security monitoring**

Access to the data should be monitored. Threat intelligence should be used to prevent unauthorized access to the data.

2.2.14 Technological Solutions Available To Secure Big Data And Ensure It's Gathered And Used Properly

The main solution to ensuring data remains protected is the adequate use of encryption. For example, Attribute-Based Encryption can help in providing fine-grained access control of encrypted data. Anonymizing the data's also important to making sure privacy concerns are addressed. It should be ensured that all sensitive information is removed from the set of records collected.

Real-time security monitoring is also a key security component for a big data project. It's important organizations monitor access to make sure there's no unauthorized access. It's also important threat intelligence is in place to guarantee more sophisticated attacks are detected and the organizations can react to threats accordingly [Guillermo, 2014].

2.2.15 Strategic And Tactical Policy Approaches Exist To Do The Same

Organizations should run a risk assessment over the data they're collecting. They should consider whether they're collecting any customer information that should be kept private, and establish adequate policies that protect the data and the right to privacy of their clients.

If the data is shared with other organizations, how this is done must be considered. Deliberately released data that turns out to infringe on privacy can have a huge impact on an organization from a reputational and economic point of view. Organizations should also carefully consider regional laws around handling customer data, such as the EU Data Directive [Everett, 2015].

2.2.16 How The Use Of Big Data Different To The Use Of Large Datasets In The Past

For example, many big data solutions look for emergent patterns in real time, whereas data warehouses often focused on infrequent batch runs. How do these different usage models impact security issues and compliance risk? In the past, large data sets were stored in highly structured relational databases. If you wanted to look for sensitive data such as health records of a patient, you knew exactly where to look and how to access the data.

Removing any identifiable information was also easier in relational databases. Big data makes this a more complex process, especially if the data is unstructured. Organizations will have to track down what pieces of information in their big data are sensitive and then carefully isolate this information to ensure compliance.

Another challenge with big data is that you can have a big variety of users each needing access to a particular subset of information. This means the encryption solution you choose to protect the data has to reflect this new reality. Access control to the data will also need to be more granular to ensure people can only access information they are authorized to see [Brewer, 2015].

2.2.17 How Companies Can Ensure And Prove Compliance While Using Big Data

The main challenge introduced by big data is identifying sensitive pieces of information stored within the unstructured data set. Organizations must make sure they isolate sensitive information and are able to prove they have adequate processes in place to achieve it.

Some vendors are starting to offer compliance toolkits designed to work in a big data environment. Anyone using third party cloud providers to store or process data will need to make sure the providers are complying with regulations [Elgendy & Elragal, 2014].

2.2.18 How Traditional Notions Of Information Lifecycle Management Relate To Big Data

Security is a process, not a product. Therefore organizations using big data will need to introduce adequate processes that help them effectively manage and protect the data.

The traditional information lifecycle management can be applied to big data to guarantee the data isn't being stored once it's no longer needed. Also policies related to availability and recovery times will still apply to big data. However, organizations have to consider the volume, velocity, and complexity of big data and amend their information lifecycle management accordingly [Guillermo, 2014].

2.2.19 How Governance Frameworks Can Be Adapted To Handle Big Data Security Issues And Risk

If an adequate governance framework isn't applied to big data, the data collected could be misleading and cause unexpected costs. The main problem from a governance point of view is that big data's a relatively new concept and therefore no one has created procedures and policies [Guillermo, 2014] as cited in Silver-Greenberg, Goldstein, and Perlroth (2016).

The unstructured nature of the information makes it difficult to categorize, model, and map the data when it's captured and stored. The problem's made worst by the fact the data normally comes from external sources, often making it complicated to confirm its accuracy. Organizations need to identify what information is of value for the business. If they capture all the information available, they risk wasting time and resources processing data that will add little or no value to the business [Guillermo, 2014].

2.3 REVIEW OF SIMILAR EXISTING SYSTEMS/PREVIOUS RELATED WORKS

According to a model proposed by Fatma M Abdullah (2019), cybercrime data analysis is carried out using data extraction tools and k-means assembly technique is used to extract the data. This data received is used to study the crime patterns. K clustering is performed on the data which is a way to classify a given data set through a number of certain cluster, it is a kind of unsupervised learning which is used to find groups in the data. When this clusters are obtained they are subjected to analysis and used to solve crime.

Lenin, William & Ambareen Siraj (2015) in the study "Survey of Crime Analysis and Prediction"stated that crime data analysis can be done using data mining techniques with the tools like weka tool, rapid minor tool, R tool, KNIME, ORANGE and Tanagra etc. Mostly the crime data analysis is done using k-means clustering

technique of data mining. Due to development in technology the criminals are using their technological equipment for doing crime. That digital data is being used to analyze the crime.

Renuka & Rajni (2013) in the study "Crime Analysis using K-Means Clustering" was of the view that the analyzed crime will be useful in predicting the hotspots. Again the data used for analyzing and for prediction purpose using data mining is structured data, when there is unstructured or semi-structured data, data mining techniques are somewhat time consuming at that moment. Also, Obtained criminal data was taken, preparing that data for rapid minor tool and perform k-means clustering on that data to obtain the clusters. After obtaining clusters, analyzing that clusters to predict the crime.

In another research work done by Tirthraj Chauhan and Rajanikanth Aluvalu (2016), they stated the growing crime rate in our society has led to large amount of criminal data which makes it very difficult for this data to be analyzed efficiently by traditional data analyses techniques. A model was proposed consisting of three phases(Using R tool, Big data analytics and Artificial neural network) were this large amount of data would be analyzed to create a crime prediction pattern for security agencies which will serve as a guide to allocate resources in fighting crime.

A slightly similar research work done by Hossam Nodel Rahaman (2020) for the Egyptian Cybercrime Centre aimed at using big data analytic model to combat cybercrime in the region.

The detection algorithm involved identifying clusters that have a greater risk of cybercrime then K means analysis is performed to determine the quality of each identified cluster. The research paper further explained cybercrime detection on Hadoop platform (a platform where advanced analytics is done) to predict cybercrime

In a research carried out by sheoran and Yadav (2017) they highlighted the threats associated with social network sites due to the escalating number of individuals that register on this sites and in most cases share their personal information, they explained further t that social networking sites have become a prey for hackers and cybercriminals and proposed a model for using big data technologies in detecting threats in social networking sites.

The proposal was based on novel data mining techniques in creating threat detection solutions, A large data (raw data collected from social media) from multiple sources is the input then data mining is applied for feature extraction, selection and reduction. Classification techniques are now used to classify data into malware or legitimated data. Prediction of threats and anomalies would now be done based on this classification techniques.

Tushar, Naman , Ankit , Aksh , and Kartik (2019) carried a study on the "Role of Big Data Analytics In Banking". The study identified the architectural frame works which are really very beneficial by examining the frauds which can be performed during the period of online transactions and hence the experts can curb it as much quick as possible. To employ such detections while performing online payments BDA takes help from big data as well as machine learning so we can obtain effective performance. At the time of transactions information are collected or gathered, cleaned and important raw information is stored from each party. It also proves helpful in the bankruptcy and also enables the individuals to give the view that the industry is going to be collapsed or not. Banking firms uses this big data which they called as the occupation of insights. Not only this, it can be extremely useful among various fields like collecting raw data and by providing the current status of market. Employing maximum correlations observations which can vary according to the needs as well the difficulties which are lying in front of the path which were followed. So, the study concluded that if the banks are able to optimize the resources to the fullest, could predict the hidden potentials and various knowledge in vast areas so that banks can work on the way of offering and providing the services to all the consumers.

2.3.1 SUMMARY OF RELATED WORK

SN	Title	Author	Year	Conclusion
1	Using big data analytics to predict and reduce cyber crimes	Fatma M Abdullah	2019	Through the proposed system, it is presented as to how big data in the field of cybercrimes recognition can be accommodated Often it says about how to manage things and become easy when part analysis Become robust while analysing complex data sets and a variety of data. It usually becomes Compelling improved techniques that can be included to avoid or prevent cyber-attacks and cybercrime as well. Also, the analysis of the data by the previously mentioned techniques, there can be many threats without specific problems before eliminating by the use of Weka tool, Quick Tool, R Tool, KNIME, ORANGE, Tanagra and q-radar mechanism. Data can also be analysed for Knowledge and Implementation result in various applications. This process adds an effective approach that can eliminate cybercrime.
2	Using Big Data Analytics for developing Crime Predictive Model	Tirthraj Chauhan and Rajanikanth Aluvalu	2016	the growing crime rate in our society has led to large amount of criminal data which makes it very difficult for this data to be analyzed efficiently by traditional data analyses techniques. A model was proposed consisting of three phases(Using R tool, Big data analytics and Artificial neural network) were this large amount of data would be analyzed to create a crime prediction pattern for security agencies which will serve as a guide to allocate resources in fighting crime
3	Survey of Crime Analysis and Prediction	Lenin, William & AmbareenSiraj	2015	crime data analysis can be done using data mining techniques with the tools like weka tool, rapid minor tool, R tool, KNIME, ORANGE and Tanagra etc. Mostly the crime data analysis is done using k-means

				clustering technique of data mining. Due to development in technology the criminals are using their technological equipment for doing crime. That digital data is being used to analyze the crime
4	Crime Analysis using K-Means Clustering	Renuka & Rajni	2013	Again the data used for analyzing and for prediction purpose using data mining is structured data, when there is unstructured or semi-structured data, data mining techniques are somewhat time consuming at that moment

2.4 IDENTIFICATION OF GAP FROM REVIEWED LITERATURE

From the previous works done by other scholars, we can say that various big data analytics technology are already in use in curbing cybercrime but there is little documentation on how effective it is and its implementation and also if the desired result is achieved. Hence this study will fill this gap by investigating the implementation, effectiveness and challenges of this techniques.

CHAPTER THREE

METHODOLOGY/SYSTEM ANALYSIS AND DESIGN

3.1 Introduction

This chapter gives the methodology that the researcher used in the study. The research design, population of the study, sample and sampling techniques, methods of data collection, variables and measurement, method of data analysis, and ethical consideration

3.2 Research Design

Research designs are perceived to be an overall strategy adopted by the researcher whereby different components of the study are integrated in a logical manner to effectively address a research problem. In this study, the researcher employed the survey research design. This is due to the nature of the study whereby the opinion and views of people are sampled.

3.3 Population of the Study

According to Udoyen (2019), a study population is a group of elements or individuals as the case may be, who share similar characteristics. These similar features can include location, gender, age, sex or specific interest. The emphasis on study population is that it constitute of individuals or elements that are homogeneous in description.

This study was carried out to investigate the implementation of critical information infrastructure protection techniques against cyber attacks using big data analytics. The Staff of Joint Admissions and Matriculation Board (JAMB) and Independent National Electoral Commission (INEC), Abuja form the population of the Study.

Statistics derived from the sampled respondents website shows that the estimated population is 502.

3.4 Sample and Sampling Techniques

3.4.1 Sample Size

A study sample is simply a systematic selected part of a population that infers its result on the population. In essence, it is that part of a whole that represents the whole and its members share characteristics in like similitude (Udoyen, 2019). In this

study, the researcher adopted the simple random sampling (srs.) method to determine the sample size.

3.4.2 Sample Technique

The Taro Yamane (1967:886) provides a simplified formula to calculate sample sizes.

Assumption

95% confidence level

P = 0.05

$$n = \frac{N}{1 + N(e)^2}$$

Where:

n= Sample size (502)

e= p-value (0.05)

n= 502/1+502 (0.05)2

n= 502/1+502 (0.0025)

n= 502/1+5.5

n=121

Therefore, for this study, the sample size is 121

3.5 Method of Data Collection Result

3.5.1 Primary

The primary sources of information were raw data obtained through questionnaires and interviews. The questionnaires were structured because of the simple fact that respondents feel more at home with questionnaires than with those that require them to indicate their responses. The questions were unambiguous and easy to answer with enough spaces provided for open-ended questions.

3.5.2 Instrument of Data Collection

Data collection is very crucial in any research process. Questionnaire as a research instrument was mainly used for collection of primary data. A range of data from book, journals and reports was also used.

3.5.3 Administration of Instrument

The research instrument used in this study is the questionnaire titled "Implementation of Critical Information Infrastructure Protection Techniques Against Cyber Attacks using Big Data Analytics", developed by the researcher. A 20 minutes survey containing 15 questions were administered to the enrolled participants. The questionnaire was divided into two sections, the first section enquired about the response's demographic or personal data while the second sections were in line with the study objectives, aimed at providing answers to the research questions which was raised against a five Likert scale of Strongly Agree(SA), Agree(A), Strongly Disagree(SD), Disagree(D), and Uncertain (U). Participants were required to respond by placing a tick at the appropriate column. The questionnaire was administered using online google forms.

3.5.4 Description Questionnaire

The researcher constructed the questionnaire for the study and submitted to the project supervisor who used her intellectual knowledge to critically, analytically and logically examine the instruments relevance of the contents and statements and then made the instrument valid for the study.

The researcher obtained the consent of all study participants before they were enrolled in the study. Permission was sought from the relevant authorities to carry out the study. Date to visit the place of study for questionnaire distribution was put in place in advance. The researcher visited the enrolled participants, and administered the questionnaire to them using online using google form.

3.5.5 Secondary Data

The secondary sources of data were utilized mainly in the review of related literature. This information was obtaining from textbooks magazine, journals, published research work, seminars workshops papers, micrograph etc.

3.6 Variables and Measurement

Predictor (independent) (x) variables, are:

- Efficacy
- Implementation, and
- Challenges

Dependent Variable:

Big Data Analytics was the dependent variable (y).

$Y = f(X)$

$Y = B_0 + (X1 + X2 + X3) + e$

$Y = B0 + (X_1^E + X_2^I + X_3^C) + e$

Y = BDA

X1 = Efficacy

X2 = Implementation

X3= Challenges

B0 = Slope

e= Error margin

3.7 Method of Data Analysis

3.7.1 Non-inferential Techniques

In the analysis of data collected, statistical method simple percentages and tables were used for descriptive purpose and to answer the research questions as well as described responses obtained.

3.7.2 Inferential Techniques

This study employed the binary logistic regression to test the null hypotheses formulated using SPSS v.23. This enables the researcher to draw a relevant conclusion, based on the findings obtained.

3.8 Ethical Consideration

The study was approved by the Project Committee of the Department. Informed consent was obtained from all study participants before they were enrolled in the study. Permission was sought from the relevant authorities to carry out the study. Date to visit the place of study for questionnaire distribution was put in place in advance.

3.9 Summary

This chapter has been able to explain the methodology that the researcher used in the study. The research design, population of the study, sample and sampling techniques, methods of data collection, variables and measurement, and method of data analysis. The population of the study comprised of 60 staffs of JAMB and 61 staffs of INEC making a total of 121 staffs as the population of this study. Binary logistic regression statistical tool was stated as the inferential techniques for testing of hypotheses of this study in this chapter.

CHAPTER FOUR

DATA PRESENTATION, ANALYSIS AND DISCUSSION

Introduction

In this chapter, the researcher presents an analysis of the data collected from the survey. The data used for this chapter was analysed using the statistical package for social science (SPSS v.23). The demographic analysis of respondents were first discussed, followed by the research questions. Finally, the research null hypotheses were tested using the logistics binary regression.

SECTION A: Demographics of Respondents

		Department of respondents				
		Institution	Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Audit	**INEC**	14	11.6	11.6	11.6
	ICT	**INEC**	37	30.6	30.6	42.2
	Test Administration	**JAMB**	3	2.5	2.5	44.7
	Psychometrics	**JAMB**	9	7.4	7.4	52.1
	Information Technology Services	**JAMB**	58	47.9		100
	Total		121	100.0	100.0	

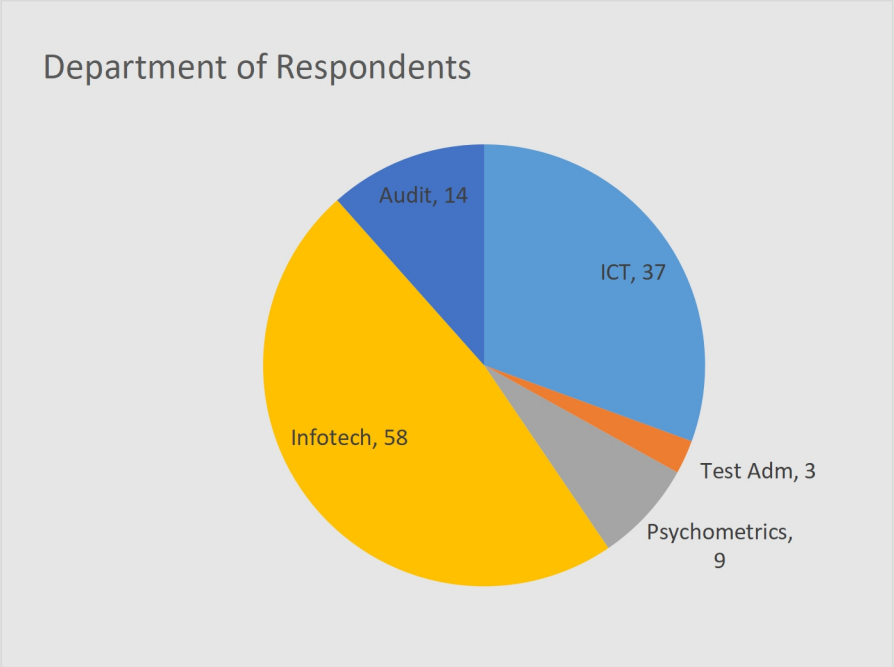


Table 4.2: Analysis of Gender of Respondents

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	female	62	51.2	51.2	51.2
	male	59	48.8	48.8	100.0
	Total	121	100.0	100.0	

Table 4.3: Age of Respondents

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	22-30	33	27.3	27.3	27.3
	30-35	40	33.1	33.1	60.3
	36-45	38	31.4	31.4	91.7
	45 above	10	8.3	8.3	100.0
	Total	121	100.0	100.0	

Table 4.4: Academic Qualification

Academic Qualification					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	B.Sc.	64	52.9	52.9	52.9
	M.Sc.	52	43.0	43.0	95.9
	Ph.D.	5	4.1	4.1	100.0
	Total	121	100.0	100.0	

Table 4.5: Professional Certification

Professional Certification

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Yes	107	88.4	88.4	88.4
	No	14	11.6	11.6	100.0
	Total	121	100.0	100.0	

Big data analytic technology					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Hamma	15	12.3	12.3	12.3
	Hadoop	25	20.6	20.6	32.9
	Storm	5	4.1	4.1	37
	Spark	64	52.8	52.8	89.8

	None	12	9.9	9.9	100
	Total	121	100.0	100.0	

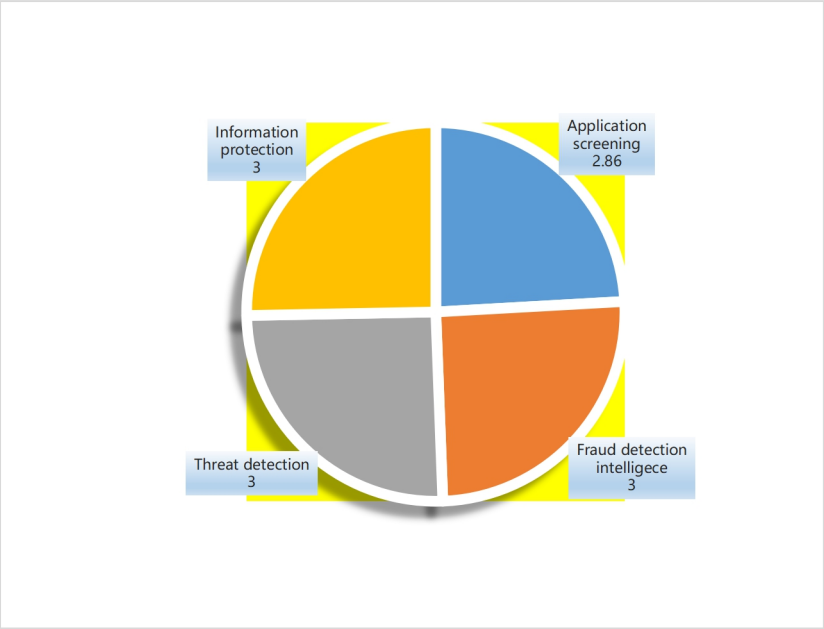
The demographic information of the respondents shows that the respondents selected from the two institutions have had interactions and understanding of big data on a regular basis, majority of the respondents were within the ages of 30-45 years old, while 52.9% have had degrees, 43% have master's degree while 4.1% have PhD. More so, 88.4% of the sampled respondents have professional certifications. The import of this information is that it highlights the intellectual capacity of the respondents to deal with the subject matter. With the above respondent's information, it can be assumed that the respondents are capable of providing reliable information relevant to the study.

SECTION B

Research Questions

How effective is big data analytics as a protection technique?

Statistics					
		Application screening	fraud detection intelligence	Threat detection	Information protection
N	Valid	121	121	121	121
	Missing	0	0	0	0
Mean		2.8678	3.0000	3.0000	3.0000



Application screening					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	uncertain	5	4.1	4.1	4.1
	disagreed	6	5.0	5.0	9.1
	agreed	110	90.9	90.9	100.0
	Total	121	100.0	100.0	

fraud detection intelligence					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	agreed	121	100.0	100.0	100.0

Threat detection

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	agreed	121	100.0	100.0	100.0

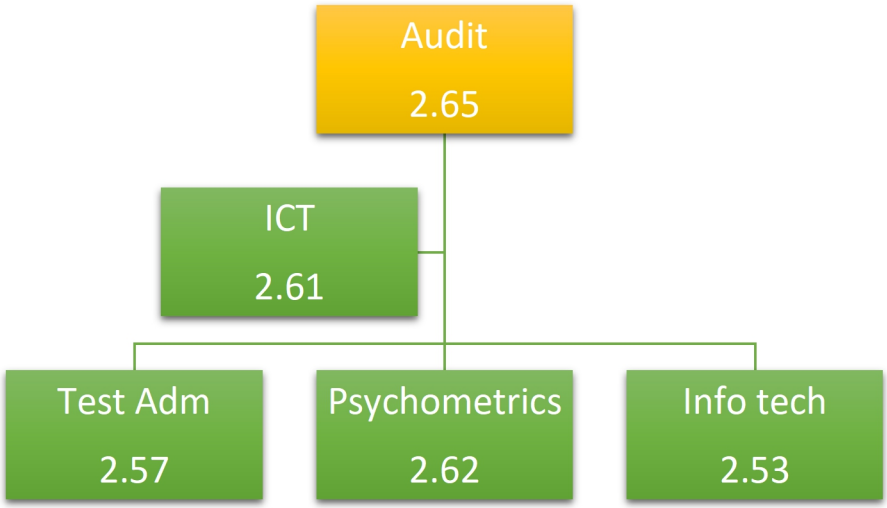
Information protection					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	agreed	121	100.0	100.0	100.0

The responses received and analysed above highlights some of the various advantages associated with the adoption and implementation of big data analytics. With a mean higher than 2.5 the different benefits of big data analytics includes, Application screening (90.9%), fraud detection intelligence (100%), marketing optimization (100%) and information protection (100%). We conclude with these responses that big data analytics is an effective protection technique for information against cyber attacks.

Question 2

What is the level of big data analytics implementation in government agencies?

Statistics						
		Audit	ICT	Test Administration	Psychometrics	Information Technology Services
N	Valid	121	121	121	121	121
	Missing	0	0	0	0	0
Mean		2.6529	2.6198	2.5702	2.6281	2.5372



Audit					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	uncertain	10	8.3	8.3	8.3
	disagreed	89	73.6	73.6	81.9
	agreed	22	18.2	18.2	100.0
	Total	121	100.0	100.0	

ICT					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	uncertain	13	10.7	10.7	10.7
	disagreed	88	72.7	72.7	83.4
	agreed	20	16.5	16.5	100.0
	Total	121	100.0	100.0	

Test Administration					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	uncertain	19	15.7	15.7	15.7
	disagreed	88	72.7	72.7	88.4
	agreed	14	11.6	11.6	100.0
	Total	121	100.0	100.0	
Psychometrics					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	uncertain	10	8.3	8.3	8.3
	disagreed	86	71.1	71.1	79.4
	agreed	25	20.7	20.7	100.0
	Total	121	100.0	100.0	

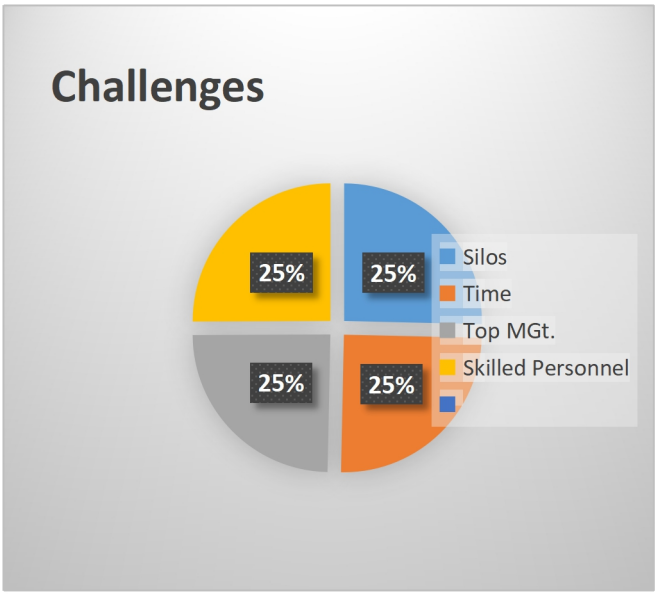
Information Technology Services					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	uncertain	16	13.2	13.2	13.2
	disagreed	81	66.9	66.9	80.1
	agreed	24	19.8	19.8	100.0
	Total	121	100.0	100.0	

On question two, the researcher sought to determine the extent government agencies have applied big data analytics. The responses received from a total of 121 respondents shows that the audit department of JAMB (73.6%) disagreed with the implementation of big data analytic, 18.2%, a total of 72.7% in the ICT department of JAMB disagreed to the implementation of big data. This response (72.7%) was the same with the test administration department. The Psychometrics department (71.1%) and the information technology services (66.9%) of the INEC disagreed to the implementation of big data analytics. From the responses received, researcher concludes on this question, that to a significant extent, government agencies have not implemented the big data analytics.

Question 3

What are the challenges militating the implementation of big data analytics protection technique?

Statistics					
		Large quantity of silos	Time	Top management influence.	High cost of skilled personnel
N	Valid	121	121	121	121
	Missing	0	0	0	0
Mean		2.5207	2.5702	2.4959	2.5289



Large quantity of silos					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	uncertain	18	14.9	14.9	14.9
	disagreed	22	18.2	18.2	33.1
	agreed	81	66.9	66.9	100.0
	Total	121	100.0	100.0	

Time					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	uncertain	14	11.6	11.6	11.6
	disagreed	24	19.8	19.8	31.4
	agreed	83	68.6	68.6	100.0
	Total	121	100.0	100.0	

High cost of skilled personnel					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	uncertain	19	15.7	15.7	15.7
	disagreed	23	19.0	19.0	34.7
	agreed	79	65.3	65.3	100.0
	Total	121	100.0	100.0	

Top management influence					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	uncertain	12	9.9	9.9	9.9
	disagreed	33	27.3	27.3	37.2
	agreed	76	62.8	62.8	100.0
	Total	121	100.0	100.0	

On research question three, the challenges of applying big data analytics were examined. Statistics shows that management decision (2.4) plays a minor role in the adoption and implementation. Of big data analytics This is because management decision is influenced by factors such as quantity of silos, cost of hiring and time. While these factors directly influence management decision, they also double as challenges impeding the effective implementation of big data analytics. Responses from the participants shows that 66.9% of the total responses agreed that quantity of silos is large as data is not pooled for the benefit of the whole organization. Also, 65.3% agreed that big data analytics has been impeded by poor expertise as acquiring professionals might involve costly hiring third-party.

Test of Hypotheses

Hypothesis One

$H_0$1: Big data analytics is not effective as an information protection technique.

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	121	100.0
	Missing Cases	0	.0
	Total	121	100.0
Unselected Cases		0	.0
Total		121	100.0

a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
No	0
Yes	1

Classification Table[a,b]

			Predicted		
			Big data analytics		Percentage Correct
	Observed		No	Yes	
Step 0	Big data analytics	No	0	17	.0
		Yes	0	104	100.0
	Overall Percentage				86.0

a. Constant is included in the model.

b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	1.811	.262	47.931	1	.000	6.118

Variables not in the Equation

		Score	df	Sig.
Step 0	Variables X1	.725	1	.001
	Overall Statistics	.725	1	.001

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	.700	1	.403
	Block	.700	1	.403
	Model	.700	1	.403

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	97.518[a]	.006	.010

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted		
			Big data analytics		Percentage
	Observed		No	Yes	Correct
Step 1	Big data analytics	No	0	17	.0
		Yes	0	104	100.0
	Overall Percentage				86.0

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	X1	.454	.537	.716	1	.001	1.575
	Constant	1.066	.900	1.401	1	.001	2.902

a. Variable(s) entered on step 1: X1.

My result shows that the overall percentage (86%) was correctly predicted as given the model. In this part of the output, this is the null model 86.0 = 104/121. the p-value at step 0 is less than 0.05 at 1 df, this implies that the null hypothesis equals 0. therefore, the hypothesis could be rejected on this basis. However, this is of no so much interest to us. The scores shows that our predictors are statistically significant which puts our data in position to predict the right outcome. The omnibus shows that we specified the full model in the logistics regression command. The p-value (.001) < 0.05 at 1 degree of freedom, therefore, at this point, we reject the null hypothesis that Big data analytics is not effective as an information protection technique.

Hypothesis Two

$H_0$2: There is no significant implementation of big data analytics in government agencies.

Dependent Variable Encoding

Original Value	Internal Value
No	0
Yes	1

Classification Table[a,b]

			Predicted		
			Big data analytics		Percentage
	Observed		No	Yes	Correct
Step 0	Big data analytics	No	0	17	.0
		Yes	0	104	100.0
	Overall Percentage				86.0

a. Constant is included in the model.

b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	1.811	.262	47.931	1	.000	6.118

Variables not in the Equation

		Score	df	Sig.
Step 0	Variables X2	.447	1	.504
	Overall Statistics	.447	1	.504

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	.445	1	.505
	Block	.445	1	.505
	Model	.445	1	.505

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	97.774[a]	.004	.007

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Classification Table[a]

		Predicted		
		Big data analytics		Percentage Correct
	Observed	No	Yes	
Step 1	Big data analytics No	0	17	.0

	Yes	0	104	100.0
Overall Percentage				86.0

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	X2	.350	.524	.444	1	.505	1.418
	Constant	1.282	.821	2.439	1	.118	3.603

a. Variable(s) entered on step 1: X2.

The overall percentage(86%) was correctly predicted as given the model. In this part of the output, this is the null model 86.0 = 104/121. The omnibus shows that we specified the full model in the logistics regression command. The p-value (.505) > 0.05 at 1 degree of freedom, therefore, at this point, we accept the null hypothesis that there is no significant implementation of big data analytics in government agencies.

Hypothesis Three

$H_0$3: There are no significant challenges impeding the implementation of big data analytics.

Dependent Variable Encoding

Original Value	Internal Value
No	0
Yes	1

Classification Table[a,b]

			Predicted		
			Big data analytics		Percentage
	Observed		No	Yes	Correct
Step 0	Big data analytics	No	0	17	.0
		Yes	0	104	100.0
	Overall Percentage				86.0

a. Constant is included in the model.

b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	1.811	.262	47.931	1	.000	6.118

Variables not in the Equation

		Score	df	Sig.
Step 0	Variables X3	5.178	1	.023
	Overall Statistics	5.178	1	.023

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	5.673	1	.017
	Block	5.673	1	.017

	Model	5.673	1	.017

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	92.546[a]	.046	.082

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted		
			Big data analytics		Percentage Correct
	Observed		No	Yes	
Step 1	Big data analytics	No	0	17	.0
		Yes	0	104	100.0
	Overall Percentage				86.0

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	X3	-1.425	.666	4.580	1	.002	.241
	Constant	4.218	1.227	11.826	1	.001	67.907

a. Variable(s) entered on step 1: X3.

The overall percentage(86%) was correctly predicted as given the model. In this part of the output, this is the null model 86.0 = 104/121. The omnibus shows that we specified the full model in the logistics regression command. The p-value (.002)< 0.05 at 1 degree of freedom, therefore, at this point, we reject the null hypothesis that there are no significant challenges impeding the implementation of big data analytics.

CHAPTER FIVE

SUMMARY OF FINDINGS, CONCLUSION AND RECOMMENDATION

5.1 Summary of findings

This study was carried out to examine the implementation of big data analytics as a protection technique for critical information. Specifically, the study examine the efficacy of big data analytics in information protection, it also examined the extent of big data analytics implementation. The motivation for this study is premised on the need for data protection against cyber attacks in Nigeria. The study adopted the survey research design to carry out this study. A total of 121 staff members of the Joint Admission and Matriculation Board (JAMB) and the Independent National Electoral Commission (INEC) were enrolled in the survey. The study employed the binomial logistic regression analysis to test the hypotheses formulated. The findings from the study shows that at 1 degree of freedom p-value < 0.05, big data analytics hold better advantage in protecting large information held by organizations against cyber attacks, also that there was no significant (.505 > 0.05) application of big data analytics in government agencies, and there are significant challenges (.002 < 0.05) impeding the application of big data analytics in institutions and organizations.

5.2 Conclusion and Recommendation

Data comes from a variety of sources, including demographic data, climatic data, scientific and medical data, energy usage data, and so on. All of these data give information about the devices' users' whereabouts, travel, interests, consumption patterns, leisure activities, and projects, among other things. However, there is also data on how infrastructure, machinery, and apparatus are utilised. The volume of digital data is continuously increasing as the number of Internet and mobile phone users continues to rise. We now live in an Informational Society that is transitioning to a Knowledge Based Society. A larger volume of data is required to extract superior understanding. The Information Society is a society in which information plays a significant role in the economic, cultural, and political spheres. Data that exceeds the storage, processing, and computational capability of traditional databases and data analysis methodologies is referred to as big data. Big Data as a resource necessitates the use of tools and methods for analyzing and extracting patterns from enormous amounts of data. Big Data Analytics is a rapidly evolving field. It has been adopted by

the most unlikely industries and has grown into its own industry. However, analyzing these data in the context of Big Data is a process that can be rather intrusive at times. Organizations, institutions, and agencies should fully adopt big data analytics, according to this study. Furthermore, in conjunction with the use of big data, staff should be trained in order to obtain the necessary abilities for executing big data analytics. For optimal implementation, this study also suggests purchasing software and a complete big data package.

REFERENCES

A. Sathi, Big Data Analytics (2013): Disruptive Technologies for Changing the Game. *Mc Press, 1st Edition*

Abdullah, Fatma. (2019). Using big data analytics to predict and reduce cyber crimes. *International Journal of Mechanical Engineering and Technology*. 10. 1540-1546.

Abraham D. Sofaer, David Clark, Whitfield Diffie (, Proceedings of a Workshop on Deterring CyberAttacks: Informing Strategies and Developing Options for U.S. Policy http://www.nap.edu/catalog/12997.html Cyber Security and International Agreements, Internet Corporation for Assigned Names and Numbers pg185 -205

Adebusuyi, A. (2008): The Internet and Emergence of Yahooboys sub-Culture in Nigeria, *International Journal Of CyberCriminology*, 0794-2891, Vol.2(2) 368-381

B. Stone-Grass, M. Cova, L. Cavallaro, B. Gilbert and M. Szydlowski (2009), *"Your Botnet is My Botnet: Analysis of a Botnet Takeover"*, CCS'09, Illinois, USA

B. Whitworth (2001), "Spam and the social technical gap," *IEEE Computer, vol. 37*, no. 10, pp. 38-45, Oct. 2004.

Bartlett, J., Kotrlik, J. and Higgins, C. (2001). Organizational Research: Determining Appropriate Sample Size in Survey Research. *Information Technology, Learning, and Performance Journal*, 19(1).

Brewer, R. (2015). Cyber threats: reducing the time to detection and response. *Network Security*, 2015 (5), pp.5-8.

Cardenas, A., Manadhata, P. and Rajan, S. (2013). Big Data Analytics for Security. *IEEE Security & Privacy,* 11(6), pp.74-76.

Cavelty, M. and Suter, M. (2012). The Art of CIIP Strategy: Tacking Stock of Content and Processes. *Centre for security studies*, pp.27 - 36.

Chauhan, Tirthraj & Aluvalu, Rajanikanth. (2016). Using Big Data Analytics for developing Crime Predictive Model.

Cloud Security Alliance, (2013). Big Data Analytics for Security Intelligence. [online] Cloud Security Alliance. Available at: https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Analytics_for_S ecurity_Intelligence.pdf

Computer Security, Wikipedia, http://en.wikipedia.org/wiki/Computer_security, [Accessed, 25 May, 2021].

Creswell, J. (2014). Research design: Qualitative, quantitative, and mixed methods approaches. 4th ed. California: Sage, pp.3 - 24, 155 - 182.

Cyber Security Analytics. Prevent Intrusion Before Its Too Late: http://bigdata.pervasive.com/Solutions/Cyber-Security.aspx

Data to Decisions, 26th September 2013, http://data-to-decisions.com/

Dugan, K. (2014). Regulator sees cyber attacks on banks causing 'Armageddon'. [online] New York Post. Available at: http://nypost.com/2014/09/22/regulator-sees-cyberattacks-on-banks-causing-armageddon/

Elgendy, N. and Elragal, A. (2014). Big Data Analytics: A Literature Review Paper. Springer, 8557, pp.214 - 227.

Eschelbeck, G. (2014). Smarter, Shadier, Stealthier Malware. Security Threat Report. [online] Sophos. Available at: https://www.sophos.com/enus/medialibrary/PDFs/other/sophos-security-threat-report-2014.pdf

Everett, C. (2015). Big data – the future of cyber-security or its latest threat?. *Computer Fraud & Securit*y, 2015(9), pp.14-17.

Fortscale, www.fortscale.com

Guillermo Lafuente (2014): Big Data Security – Challenges And Solutions

Hadoop.apache.org. (2016). Welcome to Apache™ Hadoop®!. [online] Available at: http://hadoop.apache.org/ [Accessed 25 May 2021].

Han Hu, Yonggang Wen, Tat-Seng Chua, and Xuelong Li, (2014). Toward Scalable Systems for Big Data Analytics: *A Technology Tutorial*. IEEE Access, 2, pp.652-687.

Hashem, I., Yaqoob, I., Anuar, N., Mokhtar, S., Gani, A. and Ullah Khan, S. (2015). The rise of "big data" on cloud computing: *Review and open research issues. Information Systems*, 47, pp.98-115.

http://www.stat.ufl.edu/. (n.d.). Standard Normal Probabilities. [online] Available at: http://www.stat.ufl.edu/~athienit/Tables/Ztable.pdf

Hurst, W., Merabti, M. and Fergus, P. (2014). Big Data Analysis Techniques for Cyberthreat Detection in Critical Infrastructures. *2014 28th International Conference on Advanced Information Networking and Applications Workshops*.

IBM Security Intelligence with Big Data, http://www03.ibm.com/security/solution/intelligence-big-data/

Information Security Forum, (2012). Data Analytics for Information Security: From hindsight to insight. London: *Information Security Forum Ltd,* pp.1 - 3.

International Telecommunications Commission [ITU] (2011) "Making the Online World Safer" document here: http://www.itu.int/net/itunews/issues/2011/05/38.aspx

Internet Security Threat Report. (2016). ISTR. [online] California: Symantec. Available at: https://www.symantec.com/content/dam/symantec/docs/reports/istr-21-2016-en.pdf

J. Shi and S. Saleem, (2012) "Phishing", Technical Report, http://www.cs.arizona.edu/~collberg/Teaching/466-566/2012/Resources/presentations/2012/topic5-final/report.pdf

Krishnan, R. (2016). NSA Data Center Experiencing 300 Million Hacking Attempts Per Day. [online] The Hacker News. Available at: http://thehackernews.com/2016/02/nsautah-data-center.html

L. Lu, R. Perdisci, and W. Lee, (2011) “SURF: Detecting and Measuring Search Poisoning”, CCS’11, October 2011, Illinois, USA.

L. Vaas (2007) “Malware poisoning results for innocent searches”, 27th November, 2007, http://www.eweek.com/c/a/Security/MalwarePoisoning-Results-for-Innocent-Searches

Laura, A. (1995): Cyber Crime and National Security: The Role of the Penal and Procedural Law”,Research Fellow, Nigerian Institute of Advanced Legal Studies., Retrieved from http://nials-nigeria.org/pub/lauraani.pdf

Lee, T. (2014). The Sony hack: how it happened, who is responsible, and what we've learned. [online] Vox. Available at: http://www.vox.com/2014/12/14/7387945/sonyhack-explained

Lenin Mookiah, William Eberle, AmbareenSiraj, (2015): Survey of Crime Analysis and Prediction, *Proceedings of the twenty-Eighth International Florida Artificial Intelligence Research Society Conference.*

Leong, K., and Chan, S. C. (2013): A content analysis of web-based crime mapping in the world’s top 100 highest GDP cities. *Crime Prevention & Community Safety* 15(1):1–22.

Leung, W. (2001). How to design a questionnaire. *Student BMJ,* [online]. Available at: http://www.cochrane.es/files/Recursos/How_to_design_a_questionnaire.pdf

LogRythm Security Analytics: http://logrhythm.com/siem-2.0/theplatform-for-security-analytics/logrhythm-security-analytics.aspx

Longe, O. B, Chiemeke, S. (2008): Cyber Crime and Criminality In Nigeria – What Roles Are Internet Access Points In Playing?, *European Journal Of Social Sciences* – Volume 6, Number 4

M. Bailey, E. Cooke, F. Jahanian, Y. Xu and M. Karir, (2009) “A Survey of Botnet Technology and Defenses”, CATCH’09, *Cybersecurity Applications and Technology*, Washingtin DC, USA

M. Jakobsson and S. Myers, (2007) “Phishing and Countermeasures: Understanding the Increasing Problem of Electronic Identity Theft”. *John Wiley & Sons, Inc.*,

M.T. Banday and J.A. Qadri, (2006) “SPAM – Technological and Legal Aspects”, *Kashmir University Law Review*, Vol. 8, No. 8.

McLaughlin, K., Sezer, S., Smith, P., Ma, Z. and Skopik, F. (2014). PRECYSE: Cyber-attack Detection and Response for Industrial Control Systems. [online] Available at: http://precyse.eu/downloads/

MessageLabs Intelligence: 2010 Annual Security Report, Symantec

Mueller, R. (2012). Combating Threats in the Cyber World: Outsmarting Terrorists, Hackers, and Spies. [online] FBI. Available at: https://www.fbi.gov/news/speeches/combating-threats-in-the-cyber-worldoutsmarting-terrorists-hackers-and-spies

Nivranshu Hans, Sana Mahajan, SN Omkar, (2015): Big Data Clustering Using Genetic Algorithm On Hadoop Mapreduce, *International Journal of Scientific & Technology Research Volume 4, Issue 4*.

Open.edu. (2014). [online] Available at: http://www.open.edu/openlearnworks/mod/resource/view.php?id=52658

Polak, K. (2016). Keeping European datacentres safe from cyber attacks. [online] ComputerWeekly. Available at: http://www.computerweekly.com/feature/KeepingEuropean-datacentres-safe-from-cyber-attacks [Accessed 25 May. 2021].

PRECYSE. (2012). [online] Available at: http://precyse.eu/overview/

Q. Gu and P. Liu, “Denial of Service Attacks”, Technical Report, http://s2.ist.psu.edu/paper/DDoS-Chap-Gu-June-07.pdf

Ravi Sharma (2012): Study of Latest Emerging Trends on Cyber Security and its challenges to Society. *International Journal of Scientific & Engineering Research, Volume 3, Issue 6*, ISSN 2229-5518 IJSER © 2012

Renuka Nagpal, Rajni Sehgal, (2013): Crime Analysis using K-Means Clustering, *International Journal of Computer Applications* (0975 – 8887) Volume 83

Rivera, J. (2014). By 2016, 25 Percent of Large Global Companies Will Have Adopted Big Data Analytics For At Least One Security or Fraud Detection

Use Case. [online] Gartner.com. Available at: http://www.gartner.com/newsroom/id/2663015

RSA Security Analytics, http://www.emc.com/security/securityanalytics/security-analytics.htm

Rushton, K. (2014). Cyber-criminals could spark next financial crisis. [online] Telegraph.co.uk. Available at: http://www.telegraph.co.uk/finance/newsbysector/banksandfinance/11156260/Cybercriminals-could-spark-next-financial-crisis.html

S. Curry, E. Kirda, E. Schwartz, W. H. Stewart, and A. Yoran, (2013) "Big Data Fuels Intelligence-Driven Security", RSA Security Brief.

Secure Analytics, http://www.juniper.net/us/en/productsservices/security/secure-analytics/

Security Analytics Platform: Analyze Actualize, http://www.bluecoat.com/products/security-analytics-platform

Setu Kumar Chaturvedi, Nikhil Dubey, (2014): A Survey Paper on Crime Prediction Technique Using Data Mining, Int. *Journal of Engineering Research and Applications*, Vol. 4, Issue 3(version 1), March 2014

Shackleford, D. (2016). Using Analytics to Predict Future Attacks and Breaches. [online] SANS Institute. Available at: http://www.sas.com/content/dam/SAS/en_us/doc/whitepaper2/sans-using-analyticsto-predict-future-attacks-breaches-108130.pdf

Shead, S. (2016). MIT scientists have built an AI that can detect 85% of cyber attacks — but it still needs human help. [online] Business Insider. Available at: http://uk.businessinsider.com/mit-scientists-build-ai-that-can-detect-85-of-cyberattacks-2016-4 .

Silver-Greenberg, J., Goldstein, M. and Perlroth, N. (2016). JPMorgan Chase Hacking Affects 76 Million Households. [online] DealBook. Available at: http://dealbook.nytimes.com/2014/10/02/jpmorgan-discovers-further-cyber-securityissues/?_php=true&_type=blogs&_r=1

Skopik, F., Friedberg, I. and Fiedler, R. (2014). Dealing with advanced persistent threats in smart grid ICT networks.

Solera Networks, http://www.soleranetworks.com/about/securityanalytics/

Storm.apache.org. (2015). Apache Storm. [online] Available at: http://storm.apache.org/

Sullivan, D. (2016). Introduction to big data security analytics in the enterprise. [online] SearchSecurity. Available at: http://searchsecurity.techtarget.com/feature/Introductionto-big-data-security-analytics-in-the-enterprise [Accessed 25 May, 2021].

Symantec, http://www.symantec.com/index.jsp

Tankard, C. (2012). Big data security. Network Security, (7), pp.5-8.

Teradata and Ponemon Institute (2013), Big Data Analytics in Cyber Defense, February, http://www.ponemon.org/local/upload/file/Big_Data_Analytics_in_Cyber_Defense_V12.pdf

The Economic Times New. September 11, 2004. 1.

Thilla Rajaretnam (2012), The Society of Digital Information and Wireless Communications (SDIWC), *International Journal of Cyber-Security and Digital Forensics* (IJCSDF) 1(3): (ISSN: 2305-0012)

Torty, V. (2021) Research Methodology Made Easy-iprojectblog

Trend Micro, (2015). Report on Cybersecurity and Critical Infrastructure in the Americas. [online] Trend Micro Inc. Available at: http://www.trendmicro.com/cloudcontent/us/pdfs/security-intelligence/reports/critical-infrastructures-westhemisphere.pdf

Tsai, C., Lai, C., Chao, H. and Vasilakos, A. (2015). Big data analytics: a survey. Journal of Big Data, 2(1).

Udoyen, P. (2019) Understanding The Basic Concepts Of a Research Process - iprojectblog

Ullveit-Moe, N., Gjosaeter, T., Assev, S., Koien, G. and Oleshchuk, V. (2013). Privacy Handling for Critical Information Infrastructures. [online] Available at: http://precyse.eu/downloads/

van Kessel, P. and Allan, K. (2014). Get ahead of cybercrime. EY's Global Information Security Survey. [online] EYGM. Available at: http://www.ey.com/Publication/vwLUAssets/EY-global-information-security-survey2014/$FILE/EY-global-information-security-survey-2014.pdf

Virvilis, N. and Gritzalis, D. (2013). The Big Four - What We Did Wrong in Advanced Persistent Threat Detection?. *International Conference on Availability, Reliability and Security*.

Vom Brocke, J., Simons, A., Niehaves, B., Riemer, K., Plattfaut, R. and Cleven, A. (2009). Reconstructing The Giant: On The Importance Of Rigour In Documenting The Literature Search Process. Ecis, 161, Pp.1 - 14.

Williams, R. (2016). BT broadband suffers major outage across UK. [online] The Telegraph. Available at: http://www.telegraph.co.uk/technology/2016/02/02/btbroadband-suffers-major-outage-across-uk/

Zetter, K. (2016). Sony Got Hacked Hard: What We Know and Don’t Know So Far. [online] WIRED. Available at: http://www.wired.com/2014/12/sony-hack-what-we-know/

QUESTIONNAIRE

Please tick the option that best describes your opinion. Thank you for accepting to be part of this study.

Please select department/Unit

Audit { }

ICT { }

Test Administration { }

Psychometrics { }

Information Technology Services { }

Gender

Female { }

Male { }

Age of Respondents

22-30 { }

30-35 { }

36-45 { }

45 above { }

Academic Qualification

B.Sc. { }

M.Sc. { }

Ph.D. { }

Professional Certification

Yes { }

No { }

Which of these big data analytics technology do you know of?

Hamma { }

Hadoop { }

Storm { }

Spark { }

None { }

SECTION B

Big data analytics is useful in large Application screening

Strongly agree { }

Agree { }

Strongly disagree { }

Disagree { }

Uncertain { }

Big data analytics can be a useful tool for fraud detection.

Strongly agree { }

Agree { }

Strongly disagree { }

Disagree { }

Uncertain { }

Big data analytics aids in detecting threats of cyber attacks.

Strongly agree { }

Agree { }

Strongly disagree { }

Disagree { }

Uncertain { }

Big data analytics ensure to an extent the protection of large organizational data.

Strongly agree { }

Agree { }

Strongly disagree { }

Disagree { }

Uncertain { }

Big data analytics has been adopted and fully employed in my organization

Strongly agree { }

Agree { }

Strongly disagree { }

Disagree { }

Uncertain { }

Large quantity of silos can be considered as a challenge to big data analytics implementation.

Strongly agree { }

Agree { }

Strongly disagree { }

Disagree { }

Uncertain { }

Time is another factor impeding the implementation of big data analytics.

Strongly agree { }

Agree { }

Strongly disagree { }

Disagree { }

Uncertain { }

The implementation of big data analytics in organizations can be influenced by the decisions of the top management.

Strongly agree { }

Agree { }

Strongly disagree { }

Disagree { }

Uncertain { }

The cost of hiring skilled personnel is a challenge to the adoption and implementation of big data analytics.

Strongly agree { }

Agree { }

Strongly disagree { }

Disagree { }

Uncertain { }