ARTICLE HISTORY

Received 01 October 2020
Accepted 02 November 2020

Roberto Andrade
Department of Informatics and Computer Science
Escuela Politécnica Nacional
Quito, Ecuador
roberto.andrade@epn.edu.ec

Luis Tello-Oquendo
College of Engineering
Universidad Nacional de Chimborazo
Riobamba, Ecuador
luis.tello@unach.edu.ec

Susana Cadena-Vela
College of Administrative Sciences
Universidad Central del Ecuador
Quito, Ecuador
scadena@uce.edu.ec

Patricia Jimbo-Santana
College of Administrative Sciences
Universidad Central del Ecuador
Quito, Ecuador
prjimbo@uce.edu.ec

Juan Zaldumbide
Department of Informatics and Computer Science
Escuela Politécnica Nacional
Quito, Ecuador
juan.zaldumbide@epn.edu.ec

Diana Yacchirema
Department of Informatics and Computer Science
Escuela Politécnica Nacional
Quito, Ecuador
diana.yacchirema@epn.edu.ec

Arquitectura de Analítica de Big Data para Aplicaciones de Ciberseguridad

Big Data Analytics Architecture for Cybersecurity Applications

R. Andrade, L. Tello-Oquendo, S. Cadena-Vela, P. Jimbo-Santana, J. Zaldumbide and D. Yacchirema, "Big Data Analytics Architecture for Cybersecurity Applications", Latin-American Journal of Computing (LAJC), vol. 8, no. 1, 2021.

Arquitectura de Analítica de Big Data para Aplicaciones de Ciberseguridad

Big Data Analytics Architecture for Cybersecurity Applications

Roberto Andrade
Department of Informatics and Computer Science
Escuela Politécnica Nacional
Quito, Ecuador
roberto.andrade@epn.edu.ec

Luis Tello-Oquendo
College of Engineering
Universidad Nacional de Chimborazo
Riobamba, Ecuador
luis.tello@unach.edu.ec

Susana Cadena-Vela
College of Administrative Sciences Universidad Central del Ecuador
Quito, Ecuador
scadena@uce.edu.ec

Patricia Jimbo-Santana
College of Administrative Sciences Universidad Central del Ecuador
Quito, Ecuador
prjimbo@uce.edu.ec

Juan Zaldumbide
Department of Informatics and Computer Science
Escuela Politécnica Nacional
Quito, Ecuador
juan.zaldumbide@epn.edu.ec

Diana Yacchirema
Department of Informatics and Computer Science
Escuela Politécnica Nacional
Quito, Ecuador
diana.yacchirema@epn.edu.ec

Resumen— Los cambios tecnológicos y sociales en la era de la información actual plantean nuevos desafíos para los analistas de seguridad. Se buscan nuevas estrategias y soluciones de seguridad para mejorar las operaciones de seguridad relacionadas con la detección y análisis de amenazas y ataques a la seguridad. Los analistas de seguridad abordan los desafíos de seguridad al analizar grandes cantidades de datos de registros de servidores, equipos de comunicación, soluciones de seguridad y blogs relacionados con la seguridad de la información en diferentes formatos estructurados y no estructurados. En este artículo, se examina la aplicación de big data para respaldar algunas actividades de seguridad y modelos conceptuales para generar conocimiento que se pueda utilizar para la toma de decisiones o la automatización de la acción de respuesta de seguridad. En concreto, se presenta una metodología de procesamiento masivo de datos y se introduce una arquitectura de big data ideada para aplicaciones de ciberseguridad. Esta arquitectura identifica patrones de comportamiento anómalos y tendencias para anticipar ataques de ciberseguridad caracterizados como relativamente aleatorios, espontáneos y fuera de lo común.

Palabras clave *— Big data, ciberoperaciones, ciberseguridad*

Abstract— The technological and social changes in the current information age pose new challenges for security analysts. Novel strategies and security solutions are sought to improve security operations concerning the detection and analysis of security threats and attacks. Security analysts address security challenges by analyzing large amounts of data from server logs, communication equipment, security solutions, and blogs related to information security in different structured and unstructured formats. In this paper, we examine the application of big data to support some security activities and conceptual models to generate knowledge that can be used for the decision making or automation of security response action. Concretely, we present a massive data processing methodology and introduce a big data architecture devised for cybersecurity applications. This architecture identifies anomalous behavior patterns and trends to anticipate cybersecurity attacks characterized as relatively random, spontaneous, and out of the ordinary.

Keywords *— Big data, cyber operations, cybersecurity*

I. INTRODUCTION

The increase of digitalization processes, social networks, and interactions generated in digital environments has caused security management requires more complex processes. Besides, another reality has been added related to the diversity of data sources and diverse formats. In this context, security analysts need to implement more controls and thends to know the different attacks that may occur [1].

A challenge for security processes is to establish mechanisms that require processing a large amount of data to determine patterns or anomalies that activate alerts of possible attacks. Security data analysis is not a new field; this process has been growing and developing from data mining solutions, big data, automatic learning, high-performance computing, cloud, and many available information resources to implement data science solutions. The implemented data analysis strategies offer to create a significant change for the treatment of the multiple security problems in both the training of the personnel in charge and how to analyze the companies' problems, thus the analysis of the data to generate contributions in the cybersecurity field.

The amount of data generated within the company operation is significant; therefore, the verification, analysis, and corresponding evaluation by the teams responsible for security become a challenge. Additionally, there may be the need to know different data analysis methods that respond to the security problems encountered. In an attack, the person in charge needs to review the relevant information in a short period and must analyze structured data such as the logs generated from the different infrastructure equipment (server logs, network hardware, personal user devices) and the applications of the implemented information systems; unstructured data such as those coming from websites, news, security feeds, and manufacturers' bulletins.

With this background, a proposal that allows security managers to work with these data types becomes relevant. This study is motivated by these premises and presents an analysis of big data's proposals in cybersecurity matters. A massive data processing methodology is presented jointly with an architecture proposal based on big data comprising five layers: extraction, load, transformation, analytics, and execution.

The rest of this study is organized as follows. Section II presents background information on the challenges of cybersecurity. Section III presents the related work and analyzes the different contributions that use big data in cybersecurity. Section IV discusses the massive data processing methodology that goes from the business problem to the analytical solution's value. Section V presents the results of the different contributions using a big data cybersecurity model and proposes an architecture based on big data to identify anomalous behavior patterns and trends to anticipate cybersecurity attacks characterized as relatively random, spontaneous, and out of the ordinary. Finally, Section V concludes this study.

II. BACKGROUND

According to the report by [2], 45 percent of organizations are underprepared for dedicated cyber attacks, and 30 percent have still not fully implemented antimalware software. The adoption of emerging technologies such as bring your own device (BYOD), cloud, Internet of Things (IoT), among others, increases the amount of data and complexity of networks that exceed the human capabilities of the security analyst to make sense of interrelationships among data, systems, and users. According to [3], by 2020 is predicted over 40 trillion gigabytes of digital data or 5,200 gigabytes for every person on earth. In [2], the authors mention that IoT devices are attracted to cybercriminals to be used in their illegal activity. In the year 2016, home routers of a European telecom provider were successfully attacked by a version of the Mirai worm, that convert all compromised devices into an army of bots for massive DDoS attacks [6]. FBI Cyber Division mentions that prioritization of knowledge and emerging threats is significant since cyberactors adapt and alter their tactics and techniques rapidly [5].

Big data analytics focuses on knowledge discovery in structured and unstructured data using data science, advanced statistical functions, machine learning algorithms, and visualization tools. Big data presents new alternatives for the detection and prevention of cyber-attacks using the correlation of internal and external security data [12]. Through Big data, we can take data by twitter feeds and correlate with detected events with security news published on websites or specialized blogs [4]. NIST Information Access Division (NIST-IAD) promotes the development of data analytic methods for greater and more accurate access and understanding of the information contained in multimodal heterogeneous data [8]. On the other hand, [18] mentions some cybersecurity challenges that Big data can help to resolve:

- •Data volume: Security analysts need to process large volume of data that demands efficient storage processes, high computer processing and fast access.

- •Data inconsistency: Collected data from heterogeneous sources present different structure and format that require pre-processing to prepare the data.

- •Data visualization: Visualize large datasets in real-time with different types of data require an efficient technique of visualization to present all the information in customized dashboards.

Some working groups focused on the use of Big data for cybersecurity are:

- •NIST Big Data Public Working Group [10];

- IEEE Special Interest Group (SIG) on Big Data for Cyber Security and Privacy [27];
- ITU Study Group 17 (SG17) [28];
- Cognitive Cybersecurity Intelligence (CCSI) Group [26];
- Microsoft Security and Privacy Group [34].

III. RELATED WORK

Some solutions that use big data applied to cybersecurity have been proposed in recent years. Table I presents these solutions in which the scope of the solution, the technology used, and additional techniques (e.g., statistical or machine learning) that complement the solution are highlighted.

Table I: Big Data Proposal.

ID	*Scope*	*Technology*	*Complement*	*Author*
S1	Anomaly detection	Hadoop	None	[20]
S2	Network analysis	Hadoop	None	[40]
S3	Alert correlation	Hadoop	None	[35]
S4	Intrusion detection	Hadoop	None	[45]
S5	Network analysis	Apache Spark	None	[37]
S6	Network monitoring	Hadoop	None	[32]
S7	Phising detection	Apache Spark	None	[7]
S8	DDoS detection	Hadoop	Neuronal network	[46]
S9	Intrusion detection	Hadoop	GPGNU	[13]
S10	Security events	Apache Spark	None	[19]
S11	Cyber Threat Intelligence	Hadoop	None	[42]
S12	DDoS detection	Apache Spark	Neuronal network	[22]
S13	Network monitoring	Hadoop	None	[16]
S14	DDoS detection	Hadoop	None	[47]
S15	Intrusion detection	Apache Spark	None	[24]
S16	Anomaly detection	Apache Spark	PCA	[39]
S17	DDoS detection	Apache Spark	None	[43]
S18	Anomaly detection	Apache Spark	Machine learning	[29]
S19	Anomaly detection	Apache Spark	Machine learning	[30]
S20	Anomaly detection	Apache Spark	Social Media	[31]

Fig. 1 summarizes the number of solutions found using Hadoop and Apache spark and those that have considered complementing the use of big data with other solutions such as: statistical processes or machine learning. From the literature review, it is observed that Hadoop and Apache Spark are Big data solutions mostly used for different scientific proposals; there is no substantial difference in the number of proposals using Hadoop or Apache Spark.

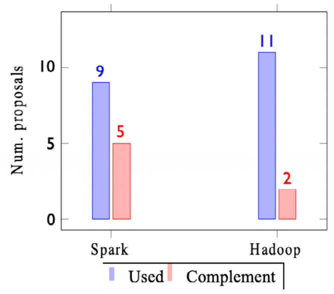

Figure 1: Comparative of Hadoop and Spark proposals.

Some solutions that use big data applied to cybersecurity have been proposed in recent years. Table I presents these solutions in which the scope of the solution, the technology used, and additional techniques (e.g., statistical or machine learning) that complement the solution are highlighted.

Additionally, Fig. 2 presents the cybersecurity operations such as anomaly detection (AD), network analysis (NA), alert correlation (AC), intrusion detection (ID), cyber threat intelligence (CTI), and attack detection (ATD) that are executed using Big data solutions. Security events and CTI have the same scope in the reviewed proposals, similar to network monitoring and network analysis. Proposals about DDoS and phishing detection are grouped into ATD. As observed, most cybersecurity operation applications mainly focus on anomaly and attack detection while AC and CTI are less developed.

A. Big data commercial solutions for cybersecurity

In the following, the commercial big data solutions focused on cybersecurity operations are reviewed.

Watson Cognitive Security [25] integrated two of its products: (i) Watson: a self-learning system that uses natural lan- guage processing to analyze unstructured data such as website information, and (ii) Qradar advisor: a security information and event management. Qradar correlates the events from different information sources such as firewall, server logs and machines. Using Watson allows correlating local security data in QRadar with unstructured data from sites such as blogs, websites or research articles.

In [23], a real-time cybersecurity platform is presented; it is composed by three macro components: telemetry data sources, telemetry data collectors, and a real-time processing engine. The latter is Apache Metron; it is composed of four modules: data collection, message queue, stream process and enrichment, and data access. Table II presents the solutions used in each module of Apache Metron.

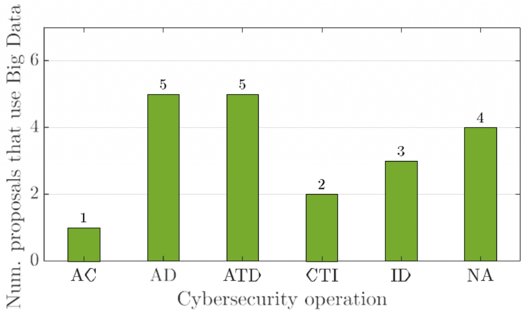

Figure 2: Application of big data in cybersecurity operations.

Table II: Apache Metron Modules.

Module	*Solution*
Data Collection	Pcap
Message Queue	kafka
Stream Process and Enrichment	Spark Storm
Data Access	Hive Solr Hbase

In [17], a real-time platform CDH based on Apache Hadoop is presented. Apache Hadoop is a software framework that supports distributed applications across clusters of computers to process large data sets using simple programming models [21]. The CDH configuration consists of three macro steps: configuring Apache Spot ODM in HDFS, installing Stream-Sets, and configuring StreamSets Data Collector Pipelines. CDH for data management is based on the Apache Spot Open Data Model (ODM), and considers the following data sources: Qualys KnowledgeBase, Qualys Vulnerability Scans, Windows Security Logs, Centrify Identity Platform Logs. CDH architecture defines six core database tables:

- event;
- vulnerability_context;
- user_context;
- endpoint_context;
- threat_intelligence_context;
- network_context.

SELKS [36] is an open distribution of linux based on the suricata ecosystem for the detection of intrusions, uses the ELK stack to correlate and display security events. The components of SELKS are:

- Suricata is a high-performance network IDS, capable of processing more than 10 Gbps.
- Logstash processes the different sources of information.
- Elasticsearch performs indexing from data events.
- Kibana is a visualization platform that allows customized dashboards, read information from elasticsearch component.
- Scirius is a web interface for Suricata that allows maps signatures from Scirus with Kibana.
- EveBox is a web-based event viewer to generate reports and alerts.

Table III: Relevant attributes of big data cybersecurity sol.

Attribute	*Watson*	*Hortonworks*	*Cloudera*	*Selks*
RTP	yes	yes	yes	yes
NLP	yes	yes	no	no
IDS	yes	no	yes	yes
ML	no	no	no	no
VA	yes	no	yes	no
CD	no	no	no	yes
ES	yes	yes	no	no
Open	no	yes	yes	yes
Core	Watson	Spark	Hadoop	ELK

Table III presents a consolidated of the attributes that we consider most relevant in each solution: Real Time Processing (RTP), Natural Language Processing (NLP), Intrusion Detection System (IDS), Machine Learning (ML), vulnerabiliLinksy analysis (VA), customize dashboard (CD), information from external sources (ES) (e.g., blogs, web pages), and security news.

IV. PROCESSING METHODOLOGY USING BIG DATA

Regarding massive data processing, a complete methodology should be considered that goes from the business problem to the analytical solution's value.

In general, a data processing model includes several phases such as the acquisition and registration (data understanding), extraction, cleaning and metadata, integration, aggregation, and representation (treatment), analysis and modeling, visualization and interpretation, communication (presentation of results), application and decision-making (enhancement).

The methodology for processing large volumes of information (big data), which allows the transformation of data into knowledge, has several components, which are:

- ***Business component:*** The business is the one that allows to address the problem and put in the value;
- ***Technology:*** it is one of the most important components, since here is the technology used and the way in which the information will be displayed;
- ***Scientific method:*** The models are built by applying the scientific method to the data. Its phases are data processing and data modeling;
- ***Communication:*** It is considered a key factor to transmit all the data in the clearest and most summarized way, it is important to consider that if the results are not communicated, value is lost.

Fig. 3 indicates the processing methodology used in big data environments, considering the four main stages: deal, technology, scientific method, communication [49], [50].

Within the Knowledge Discovery in Databases (KDD) process, data mining is considered the most important phase, since it brings together the techniques capable of modeling the available information. From the use or understanding of the generated model it is possible to extract knowledge. To be able to use data mining techniques (models) it is important to have a "minable view" of the information, (within the proposed architecture in Section V, it can be viewed as a transformation), which involves several stages within which we find the analysis of the distribution function of each attribute, in order to detect values.

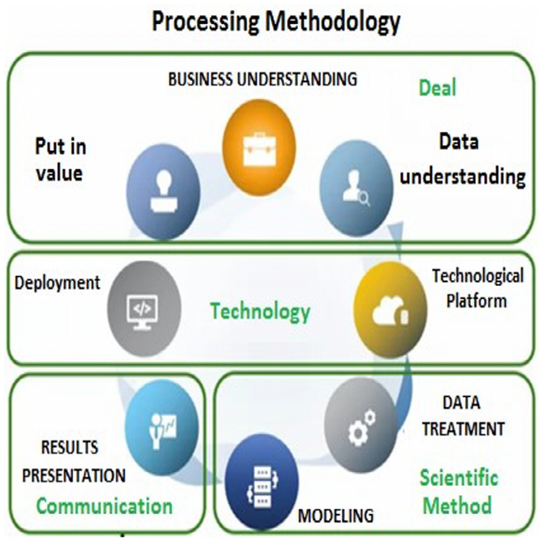

Figure 3: Processing methodology using big data.

Within the data processing stage, it is important to consider the following processes:

- Cleaning;
- Standardization;
- Transformation;
- Integration;
- Determination of missing values;
- Noise identification;
- Detection of anomalous values.

Variable transformation must be carried out, it must be indicated that it will depend on the type of problem to be solved and the data mining technique to apply if the values are ignored, discarded, or transformed. Fig. 4 indicates the steps to be followed to obtain the vitality [11].

V. CYBERSECURITY ARCHITECTURE BASED ON BIG DATA

In this section, the topics in which Big data analytics can contribute to the field of cybersecurity are presented. Then, an architecture is introduced; it comprises five layers: the extraction layer, the load layer, the transformation layer, the analysis layer, and the execution layer. This architecture pretends to identify anomalous behavior patterns and trends to anticipate cybersecurity attacks characterized as relatively random, spontaneous, and out of the ordinary.

A.Big data in Cybersecurity topics

According to our study, Big data mainly focuses on detecting anomalies and attacks; however, these activities are passive cyber-defense strategies in which the objective is to generate alerts for the security analyst. Big data could establish proactive security strategies such as cyber-deception and threat hunting that allows predicting possible attacks in the future based on extensive information processing. By doing so, attack patterns and profiles of attackers can be determined to establish counterattack strategies. Big data allows analyzing structured and unstructured data like documents, images, and videos used as digital evidence in computer forensic processes. Fig. 5 illustrates an overview of the topics in which Big data analytics can contribute to the field of cybersecurity.

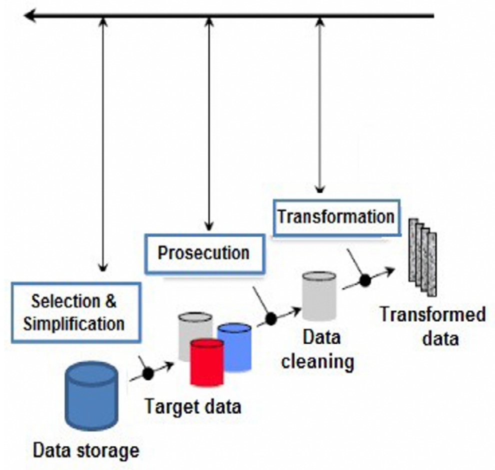

Figure 4: Mineable View.

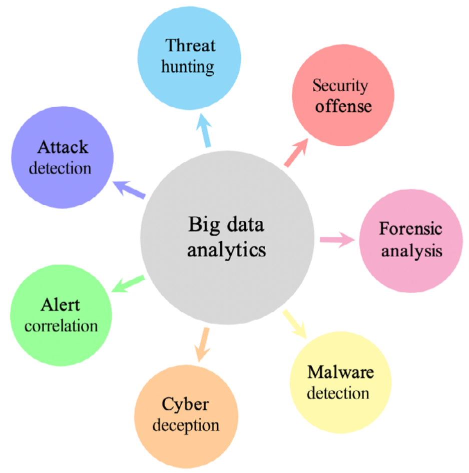

Figure 5: Cybersecurity applications for big data analytics.

Forensic analysis focuses on the preservation, analysis, and interpretation of computer data. According to the Regional Computer Forensics Laboratory (RCFL) by FBI report in 2016, 17 088 evidence items were received. This generated 5 667 terabytes for digital forensic examinations. In [45], the authors define Big data forensics as a particular branch of digital forensics where the identification, collection, organization, and presentation processes deal with a large data-set. Also, they propose a conceptual model for Big data forensics based on Hadoop; the model considers a reduplication layer to remove redundant data. This is a crucial issue in Big data proposals for assuring data integrity and quality and avoiding incorrect results due to duplicate data. In [38], the authors mention that it is possible to reduce the time and improve the effectiveness to find suspicious files by applying visualization techniques. In the current information age, an analyst is faced with looking at large volumes of data in different heterogeneous sources. Big data solutions provide two fundamental approaches: (i) integrating information from different sources with structured and unstructured formats and different file types such as images, text, or videos; (ii) customized visualization tools that include geographic attributes that provide more significant aspects for visibility to the analyst.

Malware detection. In the first half of 2018, IoT devices were attacked with more than 120 000 modifications of malware [2]; so, considering the growth of data and the need to reduce processing times, it is necessary to analyze new technological alternatives. This context motivated the interest of several researchers in analyzing the use of Big data for malware detection. In [14], the authors propose a scalable clustering approach to identify and group malware with similar behavior for which they use more than 75 thousand samples and require three hours for the processing. In [48], the authors present a method for classifying malware by combining Big data analysis with machine learning, binary instrumentation, and dynamic instruction flow analysis. In [44], the authors present issues and challenges for malware detection, such as incremental learning, active learning, malware prediction, prevalence, adversarial learning.

Security offense. It consists of main techniques, namely cyber deception and threat hunting.

Cyber deception aims to detect attacks for establishing adaptive cyber defense techniques to confuse the attacker. Traditional cyber deception techniques use honeypots and honeynets, but some exciting motivations in this research field are to incorporate artificial intelligence, game theory, and Big data to enhance cybersecurity strategies against attackers [41]. Threat hunting is an iterative activity of active defense searching through the networks and security data to detect advanced threats, instead of waiting for attack alerts [33]. In [9], the deployment of threat hunting processes using GRR rapid response is discussed through two experiments that include tests for remote code execution and the clientside exploits. In [33], the authors present the differences between threat hunting and other cybersecurity activities such as cyber defense, penetration testing, forensics, IDS, and cyber intelligence.

From these two works, the most relevant contributions can be correlated and it is concluded that threat hunting is focused on detecting intruders and unknown threats. The identification of vulnerabilities and mechanisms that can be used by an attacker before an attack is made, using basic searching, statistical analysis, visualization techniques, aggregation, machine learning, and Bayesian probability. The process of threat hunting requires processing large amounts of information generated by the logs that exceed human capabilities. By using Big data solutions, it is possible to compensate for this limitation.

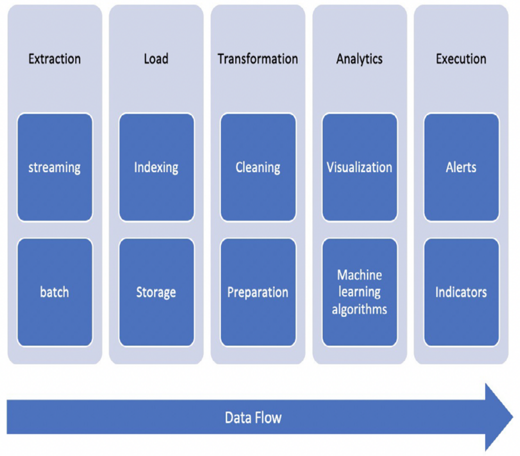

Figure 6: Cybersecurity big data architecture.

Attack detection. Security analysts need to detect attacks in the shortest time possible to reduce the time between detection and attack response. The effective attack detection requires a low false-positive rate. In [15], the authors propose two detection mechanisms: Multivariate Dimensionality Reduction Analysis (MDRA) and Principal Component Analysis (PCA). In [39], the authors propose unsupervised anomaly detection on Apache Spark using PCA for dimension reduction. Also, they mention that Big data implementations face the following challenges: selecting relevant features, scalability, and validation of learned knowledge.

B.Big Data Analytics Architecture for Cybersecurity Appli- cations

The proposed architecture, represented in Fig. 6, contains five functional layers: the extraction layer, the load layer, the transformation layer, the analysis layer, and the execution layer. The different layers are integrated to identify anomalous behavior patterns and trends to anticipate cybersecurity attacks characterized as relatively random, spontaneous, and out of the ordinary.

- Extraction layer. The extraction layer is the foundation of the architecture since it allows us to connect to the source databases and extract the data from those sources, data streams that are received continuously (streaming), and dataset that has a beginning and an end (batch). The main objective of this layer is to extract the data from different sources. Among the data sources, the following can be considered:

- Cyber simulations platforms;
- Sensors;
- Intrusion detection systems;
- Vulnerability analysis;
- Security portals, blogs, or feeds;
- Netflow;
- Servers and networking appliances logs.

It is worth noting that the data format can be in many forms, such as XML, JSON, CSV, and logs. The data can be received on a scheduled or real-time basis. To perform the extraction, many methods can be used for full or incremental loads. The extraction layer is made up of two sub-modules: streaming and batch, which are described below.

Streaming module: The data stream collected is generated in many formats, volume, and almost impossible to regulate or enforce a single data structure or control the data generated volume and frequency. The streaming submodule is in charge of obtaining the data from different data streams, using one data packet at a time, in sequential order. Each data packet includes the source of the data and a time reference to be used for loading.

Batch module: The batch sub-module is designed to obtain data from legacy batches collected from a group of events over a while (usually long). Batch data extraction is an efficient way to extract large volumes of data.

- Load layer. This layer is responsible for loading the data into the data lake for further transformation and analysis. This layer is made up of two sub-modules: storage and loading to carry out the data loading.

Storage module: This module facilitates data storage either on a local or remote platform. To allow large volumes of data to be loaded in a relatively short time, this process has been optimized so that, in the event of a load or storage failure, recovery mechanisms are triggered to restart from the point of failure without loss of data. NoSQL databases will be used to increase the responsiveness and flexibility of formats.

Indexing module: The objective of the data indexing module is to reduce the time it takes to see the results when generating a query for data with an unknown structure, especially in data that forms large tables and complex queries that involve data combinations in many cases. To carry out the indexing, this module uses some variables such as data type (file or in real-time), data size, and way of accessing the data (ad hoc or through structured application interfaces).

- Transformation layer. This layer is in charge of taking the stored data, cleaning it, and preparing it. Many of the indexed and stored data will come with empty or inconsistent fields. These incomplete tuples can affect the next layers of the architecture. The data must also be prepared in the necessary formats that will be inputs for the analysis layer.

Cleaning module: The data cleaning module will be in charge of taking the "raw" data and will be in charge of standardizing the content of the data, taking into account duplicate values, inconsistent heats, additional fields not taken into account, incomplete values, or meaningless fields.

Preparation module: This module is responsible for the preparation of clean data in aspects such as grouping, extrapolation, reduction, and increase of variables, dataset unification. Note that, although there are structured and unstructured data, it must have a logical and standardized structure.

- Analytics layer. This layer is responsible for the analysis of clean and organized data. Different machine learning and data exploration techniques will be applied here. One of the purposes of this layer is to find anomalous patterns and behaviors in them.

Visualization module: The visualization module takes care of a type of exploration based on different types of graphs so that the user can better assimilate the findings in the data.

Machine learning module: This module is designed to apply different machine learning algorithms. The purpose of the module is to find patterns and predict possible behaviors in the data. These anomalous behaviors will allow the next layer to automate early alerts.

- Execution layer. The execution layer is designed to offer different services and applications to generate alerts and perform indicators.

Alerts module: The alerts sub-module is responsible for identifying unusual or anomalous events detected in the data analysis and sending alert messages accordingly to timely inform those responsible for the data of the events detected.

Indicators module: The indicators module will allow the visualization of key performance indicators to obtain the near-real-time status of the data obtained.

Fig. 7 depicts the architecture devised based on ELK stack for covering the extraction, load, and execution layers. The number of collector servers depends on the number of data sources. The massive amount of data in cybersecurity could be a limit for the batch process. The streaming process is an adequate manner to extract data. Data is processed in real-time for collector servers. Cybersecurity data sources, such as logs, NetFlow, or beats, are relevant information sources to detect anomalous behavior patterns.

The visualization layer depends on two main factors: data and indicators of compromise (IoC). The first one is associated with the data type that could establish information relevant to anomalous behaviors. There is a lot of data sources such as firewall, routers, servers, and end devices. Try to process all this information to increase the numbers of collector servers and all big data architecture's total capacity.

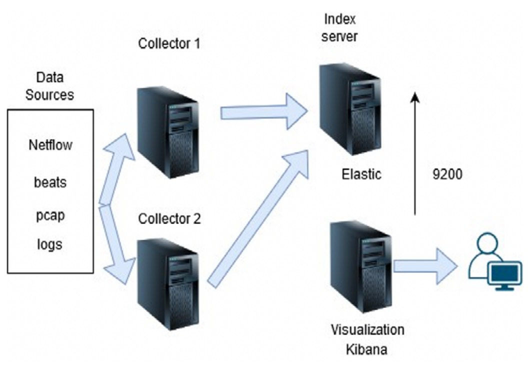

Figure 7: Architecture implementation.

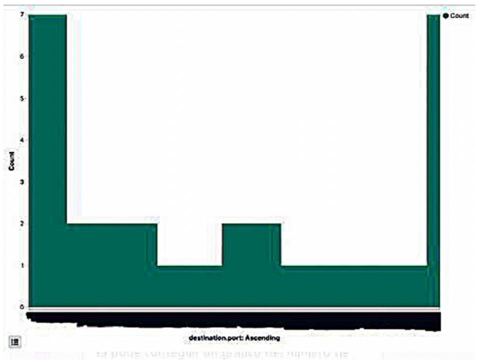

Figure 8: Number of ports during a period.

The second one depends on the first one; if the information generated based on data sources is not relevant is not possible to generate useful alerts about anomalous behaviors. The IoC allows the cybersecurity analyst to know if one event is malicious or not. For instance, Fig. 8 exhibits the number of ports that were used in a specif time; the security analyst could not identify to simple view if this is part of a cybersecurity attack. This figure was generated based on NetFlow traffic.

Another example is DNS traffic, as illustrated in Fig. 9. The ELK architecture can process this kind of data. However, without adequate IoCs, it is not possible to define by the cybersecurity analyst if a high number of connections are part of an event or not.

The analyst in this scenario needs to evaluate past events for identifying if this DNS high-rate count is normal or not. This could be a relevant factor in the streaming process, where the speed of processing is more critical than the storage of the data. To cover this lack, machine learning techniques could be considered as an alternative.

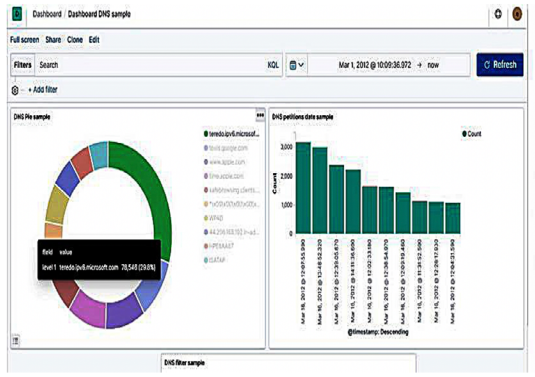
Dashboard / Dashboard DNS sample
Full screen
Share
Clone
Edit
Filters
Search
KQL
Mar 1, 2012 @ 10:09:36.972 → now
Refresh
+ Add filter
DNS Pie sample
teredo.ipv6.microsof...
tools.google.com
www.apple.com
time.apple.com
safebrowsing.clients...
WPAD
44.236.168.192.in-ad...
HPE8AA67
ISATAP
field value
level 1 teredo.ipv6.microsoft.com 78,548 (29.8%)
DNS petitions date sample
Count
3,000
2,000
1,000
0
Count
Mar 16, 2012 @ 12:07:55.990
Mar 16, 2012 @ 13:48:52.320
Mar 16, 2012 @ 12:38:05.670
Mar 16, 2012 @ 14:11:36.600
Mar 16, 2012 @ 12:02:33.180
Mar 16, 2012 @ 12:38:54.070
Mar 16, 2012 @ 12:03:19.460
Mar 16, 2012 @ 11:31:52.500
Mar 16, 2012 @ 12:28:17.030
Mar 16, 2012 @ 12:04:31.590
@timestamp: Descending
DNS filter sample

Figure 9: Number of DNS connections in a period.

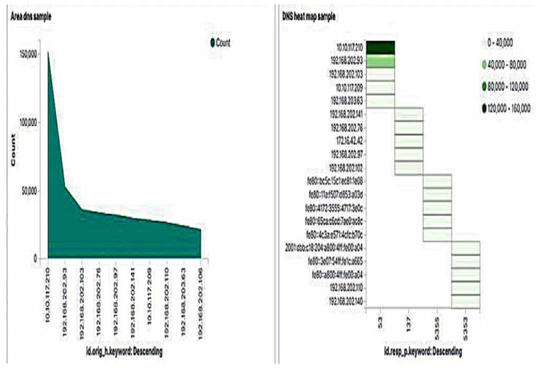

Figure 10: Number of DNS connections associated with an IP address.

Data could be combined in the ELK architecture. For instance, DNS data and NetFlow data could be associated; this allows the cybersecurity analyst to identify DNS requests with the IP address (see Fig. 10). This aspect could be relevant to establish the geo-reference of the attacker.

VI. DISCUSSION

The technological and social changes generate dynamic and complex environments that produce large amounts of data, posing new challenges to security analysts who must process this data to determine patterns or anomalies that allow identifying threats or security attacks. Big data analytics is proposed as a new alternative to improve security operations' effectiveness by processing large volumes of data of different formats in a short time.

In cybersecurity, big data is applied most to monitoring operations and detecting anomalies, focusing on reactive security strategies. However, big data analytics could enhance other security activities for proactive strategies such as threat hunting or cyber deception.

Big data can work with other solutions to complement its ability to process large amounts of data from heterogeneous sources to detect attack patterns. For instance, machine learning allows automating anomaly identification processes through training by the analyst, while natural language processing allows associate publications made in blogs or security news site blogs with detect patterns.

Note that the proposed big data architecture with ELK stack could process different types of data sources. However, the data needs a cleaning process. Another relevant aspect to consider is encrypted traffic because the ELK architecture, in its basic configuration, does not have a way to process this kind of data.

It is crucial to define the problem to be solved or countered with the architecture (e.g., DDoS, phishing, or botnets) because, depending on this, specific data sources and parameters will be necessary. It is recommended to follow the methodology outlined in Fig. 3; in particular, it is essential to understand the business component because it allows addressing the problem to be solved, and it is suggested to work in this phase with the business actors.

It is necessary to consider load balancing and fast-read disks in the architecture that facilitates processing large amounts of data from data sources such as communications equipment or server logs.

VII. CONCLUSIONS

The proposed model based on Big data covers the different components that must be considered for the generation of knowledge regarding the cybersecurity status (Cybersecurity Situation Awareness).

Implementing big data architecture is not enough to solve the problem of dealing with large amounts of data. We should identify reliable information sources, establish data quality control processes, generate safety commitment indicators, and define the update data times. The proposed cybersecurity architecture based on big data comprises five layers that identify anomalous behavior patterns and trends to anticipate cybersecurity attacks characterized as relatively random, spontaneous, and out of the ordinary.

From the conducted literature review, it is evident that few contributions exist regarding threat hunting and cyber-deception, and through the use of Big data, these operations can be enhanced. Therefore, adjust the proposed architecture to these operations is relevant, and it is proposed as future work for the project members. In the case of threat hunting, the architecture will allow identifying, in a predictive way, possible attacks by processing large amounts of data to implement security controls before an attack. In the case of cyber-deception, when identifying patterns of threats or attacks, we can change the functionality of security controls to prevent attack vectors.

ACKNOWLEDGMENT

The authors would like to thank the financial support of the Ecuadorian Corporation for the Development of Research and Academy (Red CEDIA) for the funding support of this work, under Research Group GT-II-2018 (Cybersecurity).

REFERENCES

[1] IBM. AI for cybersecurity. [Online]. Available: https://www.ibm. com/ security/artificial-intelligence [Accessed: Nov.25, 2020].

[2] Kaspersky. New IoT-malware grew three-fold in H1 2018. [Online]. Availablet: https://www.kaspersky.com/ [Accessed: Nov.25, 2020].

[3] Microsoft. Enhancing Cybersecurity with Big Data: Challenges and Opportunities. [Online]. Availablet: https://query.prod. cms.rt. microsoft.com [Accessed: Nov.25, 2020].

[4] SK., Kamaruddin and V. Ravi, "Credit Card Fraud Detection using Big Data Analytics: Use of PSOAANN based One-Class Classification," In Proceedings of the International Conference on Informatics and Analytics (ICIA-16). ACM, New York, NY, USA, Article 33 , 8 pages, 2016.

[5] FBI. Audit of the Federal Bureau of Investigation's Cyber Threat Prioritization . [Online]. Available: https://oig.justice.gov/reports/2016/ [Accessed: Nov.25, 2020].

[6] Kaspersky. DDoS attacks in Q4 2016. [Online]. Available: https: //securelist.com/ddos-attacks-in-q4-2016/77412/ [Accessed: Nov.25, 2020].

[7] P. Las Casas, V. Santos Dias, W. Meira Jr, and D. Guedes, "A Big Data Architecture for Security Data and Its Application to Phishing Characterization," pp.36-41, 2016

[8] NIST. Data Science. [Online]. Available: https://www.nist.gov/ programs-projects/data-science [Accessed: Nov.25, 2020].

[9] H. Rasheed, A. Hadi and M. Khader, "Threat Hunting Using GRR Rapid Response," International Conference on New Trends in Computing Sciences (ICTCS), Amman, 2017, pp. 155–160.

[10] NIST. Big Data Public Working Group. [Online]. Availablet: https: //www.nist.gov/el/cyber-physical-systems/big-data-pwg [Accessed: Nov.25, 2020].

[11] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From data mining to knowledge discovery in databases," AI magazine, 17(3), pp. 37–37, 1996.

[12] R. Alguliyev and Y. Imamverdiyev, "Big data: Big Promises for Information Security," IEEE 8th International Conference on Application of Information and Communication Technologies (AICT), Astana, 2014, pp. 1–4.

[13] S.R. Bandre, and J.N Nandimath, "Design consideration of Network Intrusion detection system using Hadoop and GPGPU," 2015 International Conference on Pervasive Computing (ICPC), Pune, pp. 1– 6.

[14] Bayer, Ulrich, P. Comparetti, C. Hlauschek, Ch. Kruegel, and E. Kirda, "Scalable, behavior-based malware clustering," In NDSS, vol. 9, pp. 8–11. 2009.

[15] J. Bin, M, Yan, H. Xiaohong, L, Zhaowen and S. Yi, "A Novel Real-Time DDoS Attack Detection Mechanism Based on MDRA Algorithm in Big Data," Mathematical Problems in Engineering. 2016. pp. 1–10.

[16] Z. Chen, H. Zhang, W.G. Hatcher, J. Nguyen and W. Yu, "A streaming-based network monitoring and threat detection system," IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA), Towson, MD, 2016, pp. 31–37.

[17] Cloudera. Cloudera cybersecurity. [Online]. Available: https:// www. cloudera.com/ [Accessed: Nov.10, 2020].

[18] A. Dauda, S. Mclean, A. Almehmadi and K. El-Khatib, "Big Data Analytics Architecture for Security Intelligence," Proceedings of the 11th International Conference on Security of Information and Networks, 2018.

[19] L. Fetjah, K. Benzidane, H.E. Alloussi, O.E Warrak, S. Jai-Andaloussi and A. Sekkaki, "Toward a Big Data Architecture for Security Events Analytic," IEEE 3rd International Conference on Cyber Security and Cloud Computing (CSCloud), Beijing, 2016, pp. 190–197.

[20] R. Fontugne, J. Mazel and K. Fukuda, “Hashdoop: A MapReduce framework for network anomaly detection,” IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Toronto, ON, 2014, pp. 494–499.

[21] Hadoop. Apache Hadoop. [Online]. Available: https://hadoop. apache.org/ [Accessed: Nov.10, 2020].

[22] C. Hsieh and T. Chan, “Detection DDoS attacks based on neural- network using Apache Spark,” International Conference on Applied System Innovation (ICASI), Okinawa, 2016, pp. 1–4.

[23] Hortonworks. Ciberseguridad de los macrodatos. [Online]. Available: https://es.hortonworks.com/ [Accessed: Nov.10, 2020].

[24] G.P.Gupta and M. Kulariya, “A Framework for Fast and Efficient Cyber Security Network Intrusion Detection Using Apache Spark,” Procedia Computer Science.

[25] IBM. Watson and Cybersecurity: The Big Data challenge. [Online].
Available: https://www.ibm.com/blogs/think [Accessed: Nov.10, 2020].

[26] BM. Cognitive Cybersecurity Intelligence (CCSI) Group. [Online].
Available at: https://researcher.watson.ibm.com/researcher [Accessed: Nov.10, 2020].

[27] IEEE. IEEE Special Interest Group (SIG). [Online]. Available: http://computing.northumbria.ac.uk/staff/FGPD3/sigbdcsp/ [Accessed: Nov.10, 2020].

[28] ITU. Study Group 17. [Online]. Available: https://www.itu.int/ en/ITU-T/about/groups/Pages/sg17.aspx [Accessed: Nov.10, 2020].

[29] Z. Jia, C. Shen, X. Yi, Y. Chen, T. Yu and X.Guan, “Bigdata analysis of multi-source logs for anomaly detection on network based system,” 13th IEEE Conference on Automation Science and Engineering (CASE), 2017.

[30] Lighari, S. N., and Hussain, D. M. A. (2017). Testing of algorithms for anomaly detection in Big Data using apache spark. 2017 9th International Conference on Computational Intelligence and Communication Networks (CICN).

[31] H.C. Manjunatha and R.Mohanasundaram, “BRNADS: Big data real-time node anomaly detection in social networks,” 2nd International Conference on Inventive Systems and Control (ICISC), 2018.

[32] S. Marchal, X. Jiang, R. State, R and T. Engel, “A Big Data

AUTHORS

Roberto Andrade

Roberto O. Andrade received the electronics and telecommunications engineering degree from the Escuela Politécnica Nacional (EPN) in 2007, and the master's degree in Network and Telecommunications Management from the Army Polytechnic School (ESPE), in 2013. He is currently PhD. Candidate in security systems with the School of Systems Engineering, EPN. He worked in the security areas of the Ministry of Education of Ecuador (MINEDUC) SENPLADES. He has been a certified technical instructor of CCNA, CCNP, and CCNA Security at EPN, since 2010.

Luis Tello-Oquendo

Luis Tello-Oquendo received the electronic and computer engineering degree (Hons.) from Escuela Superior Politécnica de Chimborazo, Ecuador (2010), the M.Sc. degree in telecommunication technologies, systems, and networks (2013), and the Ph.D. degree (Cum Laude) in telecommunications from Universitat Politécnica de Valencia (UPV), Spain (2018). He was Graduate Research Assistant with the Broadband Internetworking Research Group, UPV (2013 - 2018) and Research Scholar with the Broadband Wireless Networking Laboratory, Georgia Institute of Technology, Atlanta, GA, USA (2016-2017). He is currently an Associate Professor with the Universidad Nacional de Chimborazo. His research interest includes 5G and beyond cellular systems, IoT, machine learning.

Susana Cadena-Vela

Susana Cadena-Vela is Professor at the Central University of Ecuador (UCE), PhD in Computer Science, in the line of Data Quality and Open Data. Member of the research groups: Indicators for the Management of the Ecuadorian University, State of the IT of the Ecuadorian Universities sponsored by the Ecuadorian Consortium for the Development of Research and the Academy (CEDIA) and Group of Analytics and Big Data for the Cybersecurity, in addition to the Red Ecuatoriana de Datos y Metadatos (REDAM) groups and the Open Science Research Group.

Patricia Jimbo-Santana

Patricia Jimbo-Santana is Full Professor at the Central University of Ecuador, is an Engineer in Computer and Computer Systems, Computer Expert of the Criminology Institute of the Central University of Ecuador. She is PhD in Computer Science at the National University of La Plata, Argentina, and PhD in Computer Science and Mathematics of Security at the University of Virgina Rovaire of Spain, among her research lines are data mining, machine learning, big data, risk, information and communication technologies.

Juan Zaldumbide

Juan Pablo Zaldumbide is a professor at the Technologist Training School of the National Polytechnic School, in addition to the Master of Information Systems and Business Intelligence of the Armed Forces University (ESPE) and of the Big Data subject of the SEK International University. He obtained his degree in Computer and Computing Systems Engineer from the National Polytechnic School. Later, he obtained his master's degree in Systems Management at the ESPE. He then obtained his Master of Science (Computer Science) degree from the University of Melbourne - Australia, graduating with honors. He has been part of several research projects focusing on the area of Big Data, Data Analytics, Cloud Computing and Machine Learning. In addition, he has published several articles and has been part of conferences and talks in these areas. He has also served on several scientific committees and a peer reviewer for indexed journals.

Diana Yacchirema

Diana Yacchirema received the M.Sc. degree and Ph.D. degree (Cum Laude) in Telecommunications from Universitat Politècnica de València, Spain, in 2011 y 2019, respectively, and the M.Sc. degree (Hons.) in communications and information technology from Escuela Politécnica Nacional (EPN), Quito, Ecuador, in 2009. She is currently researcher and professor of the Department of Informatics and Computer Sciences of EPN. Her research activities and interests include a wide range of subjects related to Internet of Things, sensor networks, big data, cloud computing, edge computing, and network security. She received the Best Academic Record Award from EPN in 2009.

R. Andrade, L. Tello-Oquendo, S. Cadena-Vela, P. Jimbo-Santana, J. Zaldumbide and D. Yacchirema, "Big Data Analytics Architecture for Cybersecurity Applications", Latin-American Journal of Computing (LAJC), vol. 8, no. 1, 2021.