Application of Big Data Analytic in Cybersecurity

Roberto O Andrade*, Nicole Ontaneda*, Andrea Silva*, Luis Tello-Oquendo†, Susana Cadena‡, Dorys Quiroz§, Walter Fuertes§

*Escuela Politécnica Nacional, Quito, Ecuador
†Universidad Nacional de Chimborazo, Riobamba, Ecuador
‡Universidad Central del Ecuador, Quito, Ecuador
§Universidad de las Fuerzas Armadas ESPE, Sangolquí, Ecuador

Abstract—The technological and social changes in the current information age pose new challenges for security analysts. To overcome these challenges, new strategies and security solutions are sought to improve security operations concerning the detection and analysis of threats and attacks of security. 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 types of structured and unstructured formats. In this paper, we analyze the application of Big Data in order 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.

Index Terms—cognitive security, cognitive sciences, big data, cyber operations

I. INTRODUCTION

Security analysts require to process a large amount of data in order to determine patterns or anomalies that trigger alerts of possible attacks. Logs on servers, networking hardware, personal user devices, and security tools are examples of sources that produce these vast amounts of data. In this context, the Security Operations Centers (SOCs) or Computer Incident Response Teams (CSIRTs) should analyze different visualization and event correlation solutions to be able to carry out the detection process more quickly and effectively. Members of the SOCs and CSIRTs are seeking new strategies based on new technological solutions such as Big Data, Artificial Intelligence, and Data Science [18]. In the field of cybersecurity, the application of cognitive sciences to information security processes drives the concept of cognitive security, which is the integration of cognitive science [33] with data science and security operations.

The availability and access to large amounts of data in the information age allow to introduce proactive security strategies. Predictive and prescriptive analyses could have a forward-looking view of the possible impacts of an attack if a current state of security against existing threats is maintained. International organizations such as the National Institute of Standards and Technology (NIST), started the Data Science Research Program (DSRP) to accelerate the research progress for data analytic methods [34]. In the enterprise field, the CyberSecurity Data Scientist role has become one of the most wanted positions for employers and employees.

Data analytic is not a new field; nevertheless, enhanced solutions such as data mining, Big Data, machine learning, high-performance computing, cloud, and the large amount of available information resources allow data science to provide notable contributions in different fields of society such as agriculture, tourism, health, public management, cybersecurity, among others. Data analytic can generate a significant change in both the training required by the security analyst and the way to execute cybersecurity operations. Different data analytic methods can be used today. Under the perspective of the security analyst facing an attack requires to review relevant information in a short time, they must analyze data of structured type such as logs; also, they require to review unstructured data such as the one coming from websites, news, security feeds, and manufacturers' bulletins. The above-mentioned facts generated the motivation of our work on analyzing the proposals of Big Data analytic in the field of cybersecurity.

The remainder of this study is organized as follows. Section II presents background on the challenges of cybersecurity. Section III focuses on the research methodology that we propose for analyzing the different contributions about the use of Big Data in cybersecurity. Section IV discusses the results of different contributions through of a cybersecurity Big Data model. Finally, Section V presents the conclusions and proposals for future work.

II. BACKGROUND

According to the report by [24], 45 percent of organizations are under-prepared for dedicated cyber attacks and 30 percent have still not fully implemented anti-malware software. The adoption of emerging technologies such as bring your own device (BYOD), cloud, Internet of Things (IoT), among others, increases 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 [31], by 2020 is predicted over 40 trillion gigabytes of digital data or 5,200 gigabytes for every person on earth. In [24], the authors mention that IoT devices are attracted for 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 [23]. FBI Cyber Division mentions that prioritization of knowledge and emerging threats is important since cyber actors adapt and alter their tactics and techniques rapidly [10].

Big data analytics focus 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 correlation of internal and external security data [1]. Through Big data we can take data by twitter feeds and correlate with detected events with security news published on websites or specialized blogs [25]. 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 [34]. On the other hand, [7] 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 data-sets 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 [35];
- IEEE Special Interest Group (SIG) on Big Data for Cyber Security and Privacy [20];
- ITU Study Group 17 (SG17) [21];
- Cognitive Cybersecurity Intelligence (CCSI) Group [19];
- Microsoft Security and Privacy Group [32].

### III. RESEARCH METHODOLOGY

The proposed research methodology is based on the proposal of [8] as follows: (1) Define the research questions; (2) Establish the selection of scientific databases; (3) Establish the inclusion and exclusion criteria; (4) Analyze and synthesize; (5) Report and use the results.

#### A. Research questions and searching criteria

The following research questions were defined for this study:

- What are the domains of cybersecurity where the use of Big data is considered?.
- What proposals have been established to improve cybersecurity operations through use of Big data?.

Performing the literature review, we selected the following academic databases: IEEEExplore, ACM, Scopus, Science Direct, and Web of Science for obtaining specific information about proposals and the use of Big data in cybersecurity. The research has been limited to publication dates from 2014 to 2018 and studies which scope is the use of big data in the field of cybersecurity. According to the established inclusion and exclusion criteria, we defined the following search strings:

- "Big Data" AND "Cybersecurity".
- "Big Data" AND "Security".
- "Big Data" AND "Anomaly Detection".
- "Big Data" AND "Intrusion Detection".

#### B. Synthesizing of Big Data in Cybersecurity

In the academic and scientific field, some alternatives have also been proposed for the use of Big data applied to cybersecurity. In Table I, we present a consolidated list of proposals in which we highlight the technology used, additional techniques (e.g., statistical techniques or machine learning), and the year in which the proposals were made.

Based on our systematic literature review (SLR) we found 40 primary studies between the years 2014 and 2018. Then, we considered the 20 more relevant studies for our analysis. In Fig. 1, we present the number of proposals 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 our study, Hadoop and Apache Spark are Big data solutions, which are mostly used for different scientific proposals; there is no a substantial difference in the number of proposals using Hadoop or Apache Spark.

In Fig. 2, we present 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 case with network monitoring and network analysis. Proposals about DDoS and phishing detection are grouped into ATD. From our study we can observe that most applications of cybersecurity operations mainly focus on anomaly and attack detection while AC and CTI are less developed.
C. Big Data solutions for cybersecurity

To complement the results obtained from the search on scientific databases, we have conducted a research of the commercial solutions of Big data focused on cybersecurity operations.

- **Watson Cognitive Security**, [17] integrated two of its products, Watson a self-learning system that uses natural language processing to analyze unstructured data such as website information and Qradar Advisor a Security Information and Event Management. Qradar correlation the events from different information sources such as firewall, server logs and machines. Using Watson allows correlate local security data in QRadar with unstructured data from sites such as blogs, websites or research articles.

- **[15]** presents a real-time cybersecurity platform composed by three macro components:
  - Telemetry data sources;
  - Telemetry data collectors;
  - Real time processing engine.

Real-Time processing engine is Apache Metron; it is composed of four modules: Data Collection, Message Queue, Stream Process and Enrichment, and Data access.

In Table II, we present the solutions used in each module of Apache Metron.

- **[6]**, presents real-time platform CDH base on Apache Hadoop. Apache Hadoop is a software framework that supports distributed applications across clusters of computers to process large data sets using simple programming models [13]. The CDH configuration consists of three macro steps:
  - Configuring Apache Spot ODM in HDFS;
  - Installing StreamSets;
  - Configure 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.

- **[38]** 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 elastic-search component.
  - Scirius is a web interface for Suricata permits maps signatures from Scirus with Kibana.
  - EveBox is a web-based event viewer to generate reports and alerts.

In Table III, for each solution we present a consolidated of the attributes that we consider most relevant: Real Time
Table III: Relevant attributes of Big Data Cybersecurity Solutions.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Watson</th>
<th>Hortonworks</th>
<th>Cloudera</th>
<th>Selks</th>
</tr>
</thead>
<tbody>
<tr>
<td>RTP</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>NLP</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>IDS</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>ML</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>VA</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>CD</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>ES</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Core</td>
<td>Watson</td>
<td>Spark</td>
<td>Hadoop</td>
<td>ELK</td>
</tr>
</tbody>
</table>

Processing (RTP), Natural Language Processing (NLP), Intrusion Detection System (IDS), Machine Learning (ML), vulnerability analysis (VA), customize dashboard (CD), information from external sources (ES) (e.g., blogs, web pages), and security news.

IV. CYBERSECURITY MODEL BASED ON BIG DATA

Based on the results obtained by the research methodology in this work, we propose four macro-processes for Big data in the domain of the cybersecurity:

- Ingest of heterogeneous large-data;
- Data correlation visualization;
- Identification of anomaly and attacks;
- Automatic threat detection.

A. Big Data in Cybersecurity topics

According to our study, Big data mainly focuses on the detection of anomalies and attacks; however, these activities are passive cyber-defense strategies in which the objective is to generate alerts for the security analyst. Using Big data could establish proactive security strategies such as cyber-deception and threat hunting that allows predicting possible attacks in the future based on the processing of extensive information. 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. In Fig. 3, we present an overview of the topics in which Big data analytics can contribute to the field of cybersecurity.

- Forensic analysis
  Forensics focuses on the preservation, analysis, and interpretation of computer data. According to the Regional Computer Forensics Laboratory (RCFL) by FBI report in the year 2016, 17 088 evidence items were received. This generated 5 667 terabytes for digital forensic examinations. In [49], 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 the integrity and quality of data and avoiding incorrect results due to duplicate data. In [40], 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 to looking at large volumes of data in different heterogeneous sources. Big data solutions provide two essential approaches: the first one integrate information from different sources with structured and unstructured formats and different file types such as images, text, or videos. The second approach has customized visualization tools that include geographic attributes that provide more significant aspects for visibility to the analyst.

- Malware detection
  In [3], the authors propose a scalable clustering approach to identify and group malware that has similar behavior for which they use more than 75 thousand samples and require three hours for the processing. According to [24] in the first half of 2018, IoT devices were attacked with more than 120 000 modifications of malware, so considering the growth of data and the need to reduce processing times, 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 [52], the authors present a method for classifying malware by combining Big data analysis with machine learning, binary instrumentation, and dynamic instruction flow analysis. In [48], the authors present issues and challenges for malware detection:
  - Incremental learning;
  - Active learning;
  - Prediction of malware prevalence;
  - Adversarial Learning.

- Security offense

Figure 3: Cybersecurity applications for Big Data analytic.
- Cyber deception. The main objective of cyber deception is to detect attacks for establishing adaptive cyber defense techniques aiming at confusing 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 [44].

- Threat hunting. It is an iterative activity of active defense searching through the networks and security data to detect advanced threats, instead of waiting for attack alerts [30]. The work proposed by [36] et al. discusses the deployment of threat hunting processes using GRR Rapid Response through two experiments that include tests for remote code execution and the client-side exploits. In [30], 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, we can correlate the most relevant contributions and conclude that treat hunting is focused on the detection of intruders and unknown threats, and 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 the processing of large amounts of information generated by the logs which exceed human capabilities, by using Big data solutions is possible to compensate for this limitation.

- 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 meager false positive rate. In [4], the authors propose two detection mechanisms: Multivariate Dimensionality Reduction Analysis (MDRA) and Principal Component Analysis (PCA). In [41], the authors propose unsupervised anomaly detection on Apache Spark using Principal Component Analysis (PCA) for dimension reduction. Also, they mention that Big data implementations face the following challenges:
  - Selecting relevant features;
  - Scalability;
  - Validation of learned knowledge.

B. Cybersecurity Architecture Based on Big Data

In the following, we detail all the components of the proposed architecture depicted in Fig. 4.

- Cyber Collect Layer. It covers the sources of information that will be used to create cybersecurity situation awareness. Among the sources of information, the following can be considered:
  - Cyber simulations platforms;
  - Sensors;

![Figure 4: Cybersecurity Big Data architecture.](image)

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

- Infrastructure layer. The following components are included in the infrastructure layer:
  - Servers of data collection: In these servers, the process of data ingestion of the different sources of information will be carried out. At least three servers are considered for the load balancing process and high availability.
  - Indexing servers: In these servers, the process of indexing the data is performed. For doing so, the attributes are defined based on which both processes, debugging and processing the data are performed for the generation of information in the visualization layer. At least two servers are considered for the load balancing and high availability process.
  - Queue management servers: These servers establish the processes to manage the processing resources of the Big data solution when several requests for information are executed simultaneously.
  - Intrusion Detection server: In this server the rules for the detection of patterns related to security attacks are defined; the server has access with security sensors.
  - Alert management server: In this server the alert management is defined to notify the analyst when
anomalous patterns are detected; an incident management system is included in this server that allows the flow control of escalations in front of the detection of security incidents.

- **Indexing layer.** It contributes to the data search and processes are defined. Based on the results obtained, the rules of the intrusion detection system are optimized.

- **Processing layer.** The bath and stream processing methods can be used in this layer. Batch processing is a technique of processing data that occur in one large group instead of individually. It is usually done to help conserve system resources and allow for any modifications before being processed. On the other hand, stream processing allows some applications to more easily exploit a limited form of parallel processing.

- **Alert Triage Layer.** It defines the alerts that are generated towards security analysts, CSIRT, or other actors in the incident management process. Following good practices, it is advisable to define a categorization of alert levels.

- **Automated Response Layer.** It defines the response actions that can be automated; for this is necessary to establish a security incident management plan.

- **Data Assurance Quality Layer.** It defines processes for data management that ensure the integrity and quality of the data in the different layers; it includes:
  - Collection;
  - Preparation;
  - Analyze;
  - Visualization;
  - Access.

- **Processing Natural Language Layer.** It establishes mechanisms to analyze the textual content of web portals or security information blogs.

V. CONCLUSIONS AND FUTURE WORK

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. The use of Big data analytics is proposed as a new alternative to improve the effectiveness of security operations by providing the ability to process large volumes of data of different formats in a short time. In the field of cybersecurity, Big data is applied majority to monitoring operations and detection of anomalies which focus on reactive security strategies, but other security activities could be enhanced by Big data analytics for proactive strategies such as threat hunting or cyber deception. From our literature review, there are currently few contributions regarding threat hunting and cyber-deception, and through the use of Big data, these operations can be enhanced. In the case of threat Hunting, it allows identifying, in a predictive way, possible attacks by processing large amounts of data to implemented 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.

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). Just implementing a Big data architecture is not enough to solve the problem of dealing with the processing of large amounts of data, we should work on identifying reliable information sources, establishing data quality control processes, generating safety commitment indicators, and defining the update data times.

Big data can work with other solutions to complement its ability to process large amounts of data from heterogeneous sources to detect attack patterns. On the other hand, machine learning allows automating anomaly identification processes through training by the analyst, while natural processing of language allows associate publications made in blogs or security news site blogs with detect patterns. The incorporation of Big data as a cybersecurity tool generates a challenge for the SOCs or CSIRTs since it requires the formation of the security team in Big data analytics, substantial investment in technological infrastructure for Big data implementation, creation of a multidisciplinary team that integrate knowledge of psychology, statistics, security and information systems, and the definition of policies regarding the privacy of the data used by Big data.

There are different proposals in the commercial and academic field regarding the use of Big data in the security field, which are robust and offer different functionalities; however, they have not been widely implemented. We consider that a possible future work is to analyze the reasons that, in general perspective, could be varied as budget, team experience, lack of technical support. A review through a focus group could be an essential contribution to complement the present study.

ACKNOWLEDGMENT

The authors would like to thank the financial support of the Ecuadorian Corporation for the Development of Research and the Academy (RED CEDIA) for the development of this work, under Project Grant GT-II-2017.

REFERENCES


