Risks and security solutions existing in the Internet of things (IoT) in relation to Big Data

Yigliana Alvarez†§, Miguel Angel Leguizamón-Páez†, Tania J. Londoño†

†Universidad Distrital Francisco José de Caldas, Technological Faculty, Telematics Engineering, Bogotá, Colombia

§giglianamendoza@gmail.com, maleguizamonp@correo.udistrital.edu.co, taniajoha93@gmail.com

Recibido: 20 de abril de 2020 – Aceptado: 4 de septiembre de 2020

Abstract

The technological advance of the new era has led to the interconnection of devices, applications, people and data, giving way to the generation of the Internet of Things (IoT). The multiple data collected is so voluminous and variable that it must be stored in Big Data architectures. This evolution has provided the opportunity to have better access, quality and analysis of information, but at the same time, there is a challenge to prevent and mitigate the security risks associated with the relationship between IoT and Big Data, endangering the information collected and the user’s sensitive data, among others. The purpose of this document is to carry out a literature review to collect the security risks found between the relationship of Big Data and IoT, as well as evaluate the current solutions implemented and conclude if they cover the needs for prevention and mitigation of the risk.

Keywords: Big Data, IoT, IT, Risks, Security.

Resumen

El avance tecnológico de la nueva era ha dado pie a la interconexión de dispositivos, aplicaciones, personas y datos, dando paso a la generación del Internet de las Cosas (IoT). Los múltiples datos recolectados son tan voluminosos y variables que deben ser almacenados en arquitecturas Big Data. Esta evolución ha brindado la oportunidad de tener mejor acceso, calidad y análisis de la información, pero a su vez existe un desafío para prevenir y mitigar los riesgos de seguridad asociados a la relación entre IoT y Big Data, poniendo en peligro la información recolectada y los datos sensibles del usuario, entre otros. El propósito del presente documento es realizar una revisión de literatura con el fin de recolectar los riesgos de seguridad encontrados entre la relación de Big Data e IoT, así mismo evaluar las soluciones actuales implementadas y concluir si éstas últimas cubren la necesidad de prevención y mitigación del riesgo.

Palabras clave: Big Data, IoT, Riesgos, Seguridad, TI.
1. Introduction

The increase in the use of devices connected to the Internet in the last decade in what has been called the Internet of Things has allowed the collection of large volumes of data (Big Data) that with its vertiginous increase and coming mainly from the social networks like Facebook, which receives 35 million likes per minute; YouTube, with a load of approximately 100 videos per minute and Twitter with 175 million twits per day \(^1\); they have become difficult to handle, becoming a challenge for the information technology industry \(^2\).

The main challenge lies in the fact that the number of devices connected to the network grows every day, with an estimated growth of 31% since 2017 \(^3\), and which is expected to reach between 22,000 and 50,000 million devices connected between the years 2021 - 2025 \(^1, 3, 4\); taking into account that the amount of data collected can increase exponentially in shorter periods since the curve of information collected through IoT is much greater than the growth of users connected to the network \(^5\). Being a recent technology, it faces challenges related to the security of the information collected and its reliability, since if the required certainty in the collection, storage, treatment, and/or analysis of the data is not given, that Data may be intervened by a third party, resulting in the individualization of people through access to personal data, loss, and modification of information that may affect analysis and decision-making or incorrect prediction patterns may be presented.

The purpose of this article is to carry out a literature review to identify and define the main risks and challenges faced by the Internet of Things related to Big Data, to describe the solutions that have been identified up to the moment and some suggestions that authors have made to improve the storage and processing of data in this new technological era. The mentioned proposal aims to establish whether the current solutions to mitigate the existing risks between IoT and Big Data are enough and define what the consequences are if the risks materialize.

2. Methodology

For the preparation of this document, a Systematic Literature Review was carried out that seeks to find truthful information and reliable sources that allow solving the hypothesis proposed to carry out this review: What are the security risks existing among the relationship of the Internet of things and Big Data? What controls are currently carried out to control and mitigate risks? Are the implemented solutions and controls enough?

**Identification of search terms:** The search terms that were selected for the exploration of scientific articles, reviews, theses, books, conferences that allowed finding information to carry out the proposed literature review were: risks, Internet of things (IoT), Big Data, security, the relationship between IoT and Big Data, cloud storage, attacks, solutions, controls. In addition to considering these keywords, only articles published from 2015 onwards were considered (except for the ISO standards in force with previous dates), that is, with a maximum 5 years in advance, taking into account the evolution of technology and information technology nowadays.

**Search engines:** Table 1 shows the databases used in the collection of information for the construction of this article.

**Information filtering:** This was done by reading all the articles found (160), of which only 35% provided information related to the topic proposed here. After new filtering during its construction, only 41 bibliographic sources were selected.
Table 1. Search engines used in the article creation process

<table>
<thead>
<tr>
<th>Name</th>
<th>Discipline</th>
<th>Access type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academic Journals Database</td>
<td>Multidisciplinary</td>
<td>Free</td>
</tr>
<tr>
<td>Dialnet</td>
<td>Multidisciplinary</td>
<td>Free</td>
</tr>
<tr>
<td>Google Scholar</td>
<td>Multidisciplinary</td>
<td>Free</td>
</tr>
<tr>
<td>IEEE Xplore</td>
<td>Computer science, engineering, electronics</td>
<td>Subscription</td>
</tr>
<tr>
<td>ResearchGate</td>
<td>Multidisciplinary</td>
<td>Free</td>
</tr>
<tr>
<td>Science Direct</td>
<td>Multidisciplinary</td>
<td>Subscription</td>
</tr>
<tr>
<td>Scopus</td>
<td>Multidisciplinary</td>
<td>Subscription</td>
</tr>
</tbody>
</table>

Definition of sections of the article: With the reading and classification of the information found, 4 sections were created that allow us to respond to the hypothesis initially raised: 1. Describe the Internet of Things and Big Data, for the establishment of definitions to understand its operation and relationship. 2. List the risks associated with the Internet of Things concerning Big Data, to understand the challenges and problems that this technology faces today. 3. Current solutions, to understand how these solutions mitigate the risks associated in section 2, and finally, 4. Carry out an analysis and conclusion about the solutions currently implemented and the current challenges that the security faces between the relationship of IoT and Big Data.

3. Description of the Internet of Things and Big Data

In recent years the term "Internet of things" has had a great impact on several factors such as the large volume of data collected (Big Data), vulnerabilities and security aspects generated from data collection and connection between devices; the reason why their definitions are disclosed:

3.1. Internet of Things

Internet of things (IoT) is defined as a giant network in which a large number of "objects" (can be any device with Internet access, which can transfer or receive information) interact with each other through machine to machine communication (M2M) without human intervention (4,6,7). This network is responsible for collecting, processing, and analyzing all the information that passes through the network (3, 8) to make decisions, in other words, millions of devices permanently connected to the Internet act and interact intelligently with each other to feed and benefit thousands of applications that are also connected to the network (9-11). The main objective of the IoT is to automate daily work with the connection and information provided by connected devices to make people's lives easier.

To understand the risks associated with the IoT, it is pertinent to briefly review each of the layers that compose it, namely: Cloud, Fog, Edge, and Extreme Edge (12, 13).

Cloud: It is responsible for providing the IoT with a set of shared computing resources, for example, servers, applications, and services, among others. Its main function is to collect the large volumes of data that are generated in the layer called "Edge" for processing. The main advantage is the availability of the service at any time and place.

Fog: It is defined as a highly virtualized platform that provides connection and storage between end devices, that is, it is responsible for bringing the cloud closer to IoT devices by selecting information based on the use of previously established rules for that selection.

Edge: It is a non-centralized type of computing that carries out its processing and storage independently in each of the devices connected to the IoT network. The main difference between the "Fog" layer is that Edge could make autonomous decisions.
Extreme Edge: This layer is responsible for building the network with each of the devices that contain sensors, increasing the self-awareness of each device since the calculations for decision-making depends on the environment.

3.2. Big Data

Big Data (Large volumes of data) is defined as large data sets, extracted from different and new sources at high speed, which are variable and become almost impossible to handle with conventional processing software (14-16). From this definition the 3 "Vs" of Big Data are born: Variety, Volume, and Velocity, concepts that are detailed below:

Variety: Refers to the amount of data types that can be collected. With the advancement of technology and the various sources from which the data are extracted, multiple types of data are not storable in a conventional database, such data are known as unstructured or semi-structured data, they need additional processing before being stored. Some examples can be email data, financial transactions, audio, and videos, among others.

Volume: As its name implies, it has to do with the amount of data that can be collected in a defined period. With the exponential growth of technology and the interconnection of different devices on the network, information on energy consumption, hours of greatest use, and useful life of an appliance can be collected, up to the interests of people in Google searches, posts of Facebook (2), among others. Considering that these examples can be executed by millions of people connecting to the network all the time, the amount of data collected becomes unmanageable.

Velocity (Speed): Refers to the rate at which data is collected, stored, and used. Usually, the data is received in real-time. However, data storage becomes a challenge for most companies because they require reception, evaluation, and action in real-time.

The study and research about Big Data have brought to the table some additional characteristics such as Complexity, Value, Veracity, and Variability; should be considered due to the importance they have when identifying the risks associated with it (14-17):

Complexity: It is related to the multiplicity of sources that have common data, so it is necessary to connect, correlate, rank, and link the data that comes from different sources.

Value: All data collected has intrinsic value, which is only discovered when a purpose is found for the data. Data analysis is already important; however, it is not enough.

Veracity: Corresponds to the reliability that can be given to the data collected. By having a large volume of data collected in real-time, it must be analyzed if the data is genuine and can be used in analysis and predictions; this is directly related to the value of the data.

Variability: Inconsistency of unstructured and semi-structured data flow in non-periodic peaks. The variability becomes an analysis of the behavior of Volume versus a Variety of the data in each period.

4. Security risks associated with the IoT and its relationship with Big Data

To carry out the literature review regarding current solutions that allow mitigating the risks associated with IoT concerning Big Data, it is pertinent to mention the pillars of information security: Confidentiality, integrity, and availability, which are explained in Figure 1.
In addition to the pillars of computer security, Figure 2 shows factors considered in this article that are responsible for the materialization of a risk, its workflow, and some general definitions that are specified by the ISO 27001 standard, module 8:

**Asset**: Data, devices, or another component that allows the operation of a computer system.

**Threat**: Potential cause of damage to an asset.

**Vulnerability**: Weakness that an asset has and that is exploited by the threat.

**Risk**: Damage caused to an asset.

**Impact**: Consequence of the materialization of the risk.

**Probability**: Possibility of an event happening (for the present case, it will be the materialization of the risk) depending on the conditions given for it to occur.

**Control**: Measure to mitigate and / or prevent risk.

**Residual risk**: Risk remaining after the application of the control.

After understanding the components that include risk, the vulnerabilities, and/or threats associated with the relationship between the Internet of Things and Big Data are mentioned:

**Content privacy**: Many of the applications that are linked to IoT, ask the user for sensitive information that allows individual identification such as email address, date of birth, gender, address, and in some cases it also requests information regarding credit cards when they are paid applications. All the aforementioned information can be manipulated by third parties due to its storage in the cloud.

**Insufficient authentication**: According to OWASP (Open Web Application Security Project), the violation of applications associated with the IoT network through weak authentication is the second most used method by attackers seeking to modify, delete or steal information stored in the Big Data cloud. Most mobile devices use weak passwords and encryption methods that can be easily compromised.

**Lack of transport encryption**: The information sent from IoT devices to Big Data through the local network (LAN) and the Internet sometimes travels flat, since the devices do not have an encryption method and / or security certifications that allow attackers to obtain information through MITM (man in the middle) methods where the attacker acquires the ability to read, insert and modify the information at will.

**Falsification of profiles**: Attackers create multiple fake profiles to saturate the existing resources within the network, giving way to attack. As a consequence, you have access to the functionalities, user roles, and data associated...
with IoT and Big Data devices for their control to carry out fraudulent activities, alteration and/or damage to sensors, cameras, appliances, telephones, among others, that prevent the network from functioning normally.

**Network manipulation** (31): The efficiency of Fog (fog, one of the layers of the IoT) is degraded, delaying the transmission of data, allowing its manipulation and modification. Regarding Big Data, the risk is specified in the manipulation and modification of the data extracted from the sources, giving rise to incorrect results in the analysis prediction.

**Blackhole and Greyhole** (32): They are malicious nodes connected in the Fog of the IoT where: Blackhole oversees discovering the route to send the messages to be part of that trajectory. As soon as the messages arrive at the node, it discards the packages and the Big Database does not receive them, in some cases the packages are manipulated before discarding them. Greyhole is responsible for diverting the package to the Big Data storage when it reaches this node, sending a message to the router confirming its reception. It is a difficult attack to identify because it uses end-to-end connectivity. In effect, you would have received incomplete information, modified information, or even did not receive the information without any type of detection.

**Insufficient input validation and filtering** (31, 33): It is known that Big Data handles a large amount of data that comes from different sources linked to the IoT network, which do not have sufficient validation and filtering for data entry, which leads to a large amount of data that can be extracted from unreliable sources. This is a latent threat in all databases.

**Table access control** (31): Big Data was created to optimize time and performance in storing information from IoT. However, the security access to the tables where it is stored was not considered, this being a great risk to the integrity of the data. Conventional databases have access controls to the table, columns, and rows that allow them to have a record of all the entries and modifications done. On the other hand, Big Data does not have any access control to the table, allowing attackers to recover information through personalized queries.

**Insecure data storage** (16, 31, 34, 35): considering that Big Data has millions of nodes in which
information is stored; organizing, authenticating, authorizing, and encrypting them becomes difficult to work. Currently, data is moved from the IoT network to cold Big Data storage, reducing storage security. Today, real-time data encryption is not a solution as it can have performance impacts.

DoS / DDoS attacks: This type of attack does not target IoT devices; a third party (attacker or hacker) uses them to compromise other devices, not necessarily devices connected to the IoT network. In the first place, the malware automatically finds a vulnerable device connected to IoT, infecting and associating it with a botnet (an autonomous computer program that is capable of carrying out specific tasks and imitating human behavior), which is then used to perform DDoS flooding the server with malicious network traffic. Network attacks within the IoT can be carried out via HTTP / HTTPS, SMTP, and port scans.

Software: According to a study conducted by HP, 60% of software performs system updates that are not encrypted at the time of download, this gives rise to Attackers to intercept the download and gain access to the application's source code, allowing you to make changes to the source code to steal information.

According to the threats and vulnerabilities related to the Internet of Things and Big Data, the following risks may arise:

- Loss of information, damage to equipment, loss of time in repeated processes
- Individual identification of people associated with data and/or devices.
- Impersonation of individuals and devices in the IoT network and loss and/or alteration of the data collected for Big Data: This risk represents a challenge due to the variety of devices, manufacturers, operating systems, among others; makes it difficult to control access and authentication permissions in IoT, on the other hand, the storage, recovery, and protection of data in Big Data.
- Affecting the availability of the IoT network and data storage in the Big Data cloud.
- Access to sensitive data that has been exposed when transmitting the data to Big Data. The attacker can make multiple uses with the data collected as manipulation of applications linked to the IoT network that alters its operation and give rise to fraudulent actions.
- Predictive analytics not reliable or safe to use: Big Data faces a higher risk due to the amount of data it stores and its difficulty in filtering it. By not guaranteeing reliable identification and filtering of information, it is highly likely that the predictive analyzes are not correct.
- Malfunctioning of systems, destruction of OS, destruction, or modification of applications and information: The attacker to the IoT network can handle both devices and data stored in the Big Data cloud to function as desired to reach a specific objective.
- Alteration in the operation of the code, programs, and sites linked to the IoT network.

5. Current solutions implemented to mitigate security risks associated with IoT and Big Data

Currently, there are solutions and practices, referred to by some authors as controls, to follow to mitigate the mentioned risks regarding the association of IoT and Big Data, such as:
Secure computing code \cite{22, 31}: Due to the large volumes of data in Big Data, access to data, and sending of information from the IoT network, it is necessary to verify and implement access control and dynamic analysis of the code to avoid malicious attacks that affect the well-being of the data. Hadoop encryption solutions are recommended since they incorporate transparent application-level security via API to protect data without changing the database structure.

Data access control \cite{3, 22, 29, 31, 39}: It is necessary an access control that determines the permissibility of network resources and the handling of data to only those devices/users who have certain rights to use the requested resource. To ensure efficient and secure access control, strong credential management policies must be included to ensure the reliability and management of keys considering attribute-based access control or also known as policy-based access control (ABAC).

Authentication \cite{29, 39, 40}: It is the mandatory identification that must be carried out to control access to IoT and Big Data functionalities, this is done through defined profiles and its authentication is done through credentials such as Username, password, and biometric readers, among others.

Security policies \cite{22,23}: To face the security risks linked to IoT and Big Data, it is necessary to establish policies based on cryptography, credential management, passwords, ensuring a strong level of complexity for access in the network application where it covers all routers involved in the traffic to ensure the sending of data packages from IoT to Big Data.

Visibility and Control \cite{17}: Real-time monitoring of services that can interact between IoT devices and Big Data sources of information must be carried out to detect and mitigate threats, allowing control of the operation of the devices and reliable Big Data information.

Blockchain \cite{29, 32}: The variety and volume of Big Data that are obtained through IoT devices connected to the network can have security and confidentiality problems. To avoid this, it is advisable to use Blockchain. These security systems can work autonomously in real-time since they have a distributed computing environment to ensure network resources and data transactions, such as trust and security solutions. Thus, it offers to protect the network when a new device connects to it, it can also detect and remove a faulty item that compromises the security of the system.

Secure data storage \cite{16, 31, 34, 35}: In the data storage, the option of sensitive data leakage can be centralized, for this, it is recommended the activation of encrypted data, administrative access audit, and verification of the appropriate API security settings.

Software and server \cite{22, 37, 38}: To prevent IoT devices from being intercepted and manipulated, allowing malicious code to be injected affecting IoT devices and data in Big Data, the purchase of security software is recommended, it is suggested that it can continuously update to avoid new security breaches and those that are found are eliminated as soon as their existence is discovered. To ensure that the updated version for the server has not been altered, it is recommended that it is encrypted as much as possible, not have any reported vulnerabilities and, at the end of the update, perform a secure boot.

Secure Firewall Manager: \cite{35, 36, 41}: The implementation of a Firewall in hardware and/or software in the devices connected to IoT will control the access of users who access private networks connected to the Internet. The firewall will oversee controlling the flow of data packages; if there is any package that does not meet the security criteria, it will be automatically blocked. This solution mitigates the risk
associated with Blackhole and Greyhole and enables reliable data storage in the Big Data cloud.

6. Analysis and conclusions of the literature review

The exponential growth of IoT is imminent, at the same time the increase in the data collected is even greater since a single device connected to the IoT network can generate thousands of data, an important requirement that was not considered initially when it was thought about interconnecting and related devices, data, and people on the network. The risks associated with the relationship between IoT and Big Data are very high, since, as it has been shown in previous sections, the consequences of threats materializing are serious, since not only is the security and availability of data at stake and devices, but in addition to this, people's lives are in danger.

The main objectives of the attackers are the theft, modification, and/or elimination of information, as well as the impersonation of individuals, which leads to fraud, economic losses, personal identification for a specific purpose, theft of devices to obtain control and information from those that are connected to the network; this goes from a cell phone to a smart home connected to IoT. According to the literature review carried out, risks are difficult to identify, since users connected to the network are not prepared with a data security policy that could prevent a risk from materializing.

Usually, people are not used to configuring their devices and follow the instructions provided in its user manual, this being the first open door that an attacker can find to infiltrate the IoT network and thus, access both device control and data. It has been shown that many risks have the same consequence. This means that attackers have multiple options to achieve their mission. For this reason, it is important to prevent risks from materializing with current solutions. Although it is still a challenge to have complete control over the amount of data that is collected from IoT and stored in Big Data, it is possible to mostly prevent having security breaches that compromise the data. Having the greatest possible control over the security of data and devices connected to IoT is the responsibility of all elements connected to the network; Each element connected to the network must have a solution that prevents and mitigates the materialization of the associated risk.

Currently, software, hardware, and personnel risks have been discovered that can be initially mitigated with a data security policy that includes all elements connected to the network. In addition to this, double user authentication with passwords and additional identification such as pins, biometrics, among others; together with the implementation of access control to the data that is stored in the Big Data cloud. On the other hand, there are risks associated especially with the applications used in IoT, how they transmit and store information, as well as how the collected data is consulted and modified. These include threats such as DDoS, insecure software, insufficient access control for querying and modifying tables, data transmission attacks such as Blackhole and Greyhole that trigger the materialization of risks such as loss of information, damage to equipment, identification of personnel, impersonation of individuals, affection of availability, integrity, and confidentiality of information, among others.

Fortunately, some solutions or controls mitigate these risks as much as possible, such as the implementation of a firewall that controls and monitors the transmission of packets that travel from the network to Big Data, additionally, a secure computing code can be launched and security software can be run, to guarantee that the
updates of the applications linked to the network do not leave security breaches, having an access control to the databases included in the Big Data cloud can make the difference in the impersonation of devices and intervention of the data.

Finally, risks are always present in any hardware, software, network deployment, and data storage implementation. Although most risks can be prevented today, the biggest challenge lies in the excessive growth of the information collected for Big Data through IoT. The transmission of data in real-time carries a great risk that does not yet have a relevant solution, it can become a risk that becomes a snowball difficult to control. Future work is expected to carry out research to identify new solutions for this challenge.

7. Funding Statement

The author(s) received no specific funding for this work.

8. References


