Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/cose

Computers & Security

TC 11 Briefing Papers

MARISMA-BiDa pattern: Integrated risk analysis for big data

David G. Rosado[a,*], *Julio Moreno*[b], *Luis E. Sánchez*[b], *Antonio Santos-Olmo*[c], *Manuel A. Serrano*[d], *Eduardo Fernández-Medina*[b]

[a] *Institute of Technologies and Information Systems, University of Castilla-La Mancha, Ciudad Real, Spain*
[b] *GSyA Research Group, University of Castilla-La Mancha, Ciudad Real, Spain*
[c] *I+D+i Department, Marisma Shield S.L and Sicaman Nuevas Tecnologías S.L., Tomelloso, Spain*
[d] *Alarcos Research Group, University of Castilla-La Mancha. Ciudad Real, Spain*

ARTICLE INFO

Article history:
Received 8 January 2020
Revised 22 September 2020
Accepted 14 December 2020
Available online 19 December 2020

Keywords:
Big data
Risk assessment
Risk analysis
Information security
Security standards

ABSTRACT

Data is one of the most important assets for all types of companies, which have undoubtedly grown their quantity and the ways of exploiting them. Big Data appears in this context as a set of technologies that manage data to obtain information that supports decision-making. These systems were not conceived to be secure, resulting in significant risks that must be controlled. Security risks in Big Data must be analyzed and managed in an appropriate manner to protect the system and secure the information and the data being handled. This paper proposes a risk analysis approach for Big Data environments, which is based on a security analysis methodology called MARISMA (Methodology for the Analysis of Risks on Information System), supported by a technological environment in the cloud (eMARISMA tool) already used by numerous clients. Both MARISMA and eMARISMA are specifically designed to be easily adapted to particular contexts, such as Big Data. Our proposal, called MARISMA-BiDa, is based on the main related standards, such as ISO/IEC 27,000 and 31,000, or the NIST Big Data reference architecture or ENISA and CSA recommendations for Big Data.

© 2020 Elsevier Ltd. All rights reserved.

1. Introduction

Data has become increasingly important in companies in any area, not only they are fundamental for organizations related to the area of information technologies, but also crucial for industries as varied as health, education, engineering, or governments. Data is essential for all of them to conduct their daily activities and to help top management to achieve their business objectives and consequently, make better decisions based on the information extracted from that data (Akoka et al., 2017; Armstrong, 2014). Additionally, a greater amount of data is being steadily generated. It is estimated that 90 percent of the data generated by humans throughout their history have been created in recent years. For example, throughout 2003, 5 Exabytes of data were generated. Notwithstanding, in 2013 that amount of data was generated every 2 days (Sagiroglu and Sinanc, 2013). This tendency to increase

* Corresponding author.
E-mail addresses: David.GRosado@uclm.es, david.grosado@uclm.es (D.G. Rosado), LuisE.Sanchez@uclm.es (L.E. Sánchez), asolmo@sicaman-nt.com (A. Santos-Olmo), Manuel.Serrano@uclm.es (M.A. Serrano), Eduardo.FdezMedina@uclm.es (E. Fernández-Medina).
https://doi.org/10.1016/j.cose.2020.102155
0167-4048/© 2020 Elsevier Ltd. All rights reserved.

the amount of data generated does not seem to be changing in the near future, quite the contrary, since 16.1 Zettabytes were generated in 2016 and it is expected that by 2025, 163 Zettabytes will be reached (Reinsel et al., 2017).

The broader use of social networks, multimedia data, and the Internet of Things (IoT) produce an increasing amount of data (Hashem et al., 2015). On the other hand, many of these data have an unstructured format, which, together with the rapid production of them—in many cases in real time—complicates their analysis using traditional systems. This set of characteristics are known as the 3 Vs (Volume, Variety, and Velocity) of Big Data (Chen et al., 2014). Big Data consists of extensive datasets that require a scalable architecture for efficient storage, manipulation, and analysis (NIST, 2018a). Big Data arises in response to the need to analyze and better understand this data, in order to obtain valuable information for the organization.

However, with every new technology, new problems arise, and Big Data is no exception. Using Big Data not only increases the scale of traditional privacy and security issues but also adds new challenges that must be addressed (Moreno et al., 2018; Wang et al., 2015). These problems stem from the fact that Big Data was not initially conceived as a secure environment (Priya and Navdeti, 2015), but instead, the security risks to which a system of this type may be subject are very high. Hence, even security issues in Big Data are one of the main obstacles for its adoption by companies, for the fear of losing confidential information, reputation problems, or even infringement of the legislation on data protection (Akinrolabu et al., 2019; Dong et al., 2015; Zhang, 2018). Therefore, it is of prime importance to have a series of guides, methodologies, and mechanisms to adequately implement both the Big Data environment and its security. In addition to that, it is also widely considered that any global information security management environment in the company should be focused on risks (Barrientos and Areiza, 2005; Disterer, 2013; Fredriksen et al., 2002). Thus, security risks in Big Data must be analyzed and managed in an appropriate manner, together with the risks of other types of information systems assets.

Although there are several proposals to address the risk analysis and management, such as MAGERIT (Spanish Higher Council for Government, 2012), OCTAVE (Alberts et al., 2003) or CRAMM (CCTA, 2003), these proposals frequently present difficulties in their practical application or there are no adequate tools to process them, not very adaptable and reusable, informal and highly subjective outcomes without adequate metrics and indicators to measure risk, between other unresolved problems (Abbass et al., 2015; Acevedo and Satizábal, 2016; Pan, 2018; Pan and Tomlinson, 2016; Rajbhandari, 2013; Shameli-Sendi et al., 2016; Wangen and Snekkenes, 2013; Zambon et al., 2011). This may happen because these proposals are designed to be applied in large companies and they are not sensitive to the context—without an adaptive capacity for special environments, which requires special treatment of their risks. The MARISMA methodology was developed addressing these gaps, as well as a technological environment called eMARISMA that supports it (www.emarisma.com). MARISMA is focused on the reuse of knowledge for the process of risk analysis and management by defining the concept of a meta-pattern, which is a data model representing the key risk analysis components, and which can be specialised to any specific area or environment with particular characteristics, such as IoT, cyber-physical systems, cloud computing, critical systems, etc. through the corresponding risk pattern. The risk patterns are context-based templates that can be reused and customised to be applied in the risk analysis and management process for several clients, helping to drastically reduce the effort of the process. The technological environment eMARISMA is a platform in the cloud that implements the MARISMA processes, allowing the automation of risk analysis and management process, supporting reuse, and allowing real time updating of risk indicators.

In fact, the MARISMA framework, specialized with two initially defined risk patterns (one pattern for the generic management of risks based on ISO/IEC 27001 and another pattern of critical infrastructures based on ISO 62443) has, together with the eMARISMA tool, been used by many international companies for almost a decade now. This has meant that both the framework and the tool are constantly evolving and improving. In addition, this experience with real clients has allowed us to identify the need to have other specific patterns to manage their risks, as in the specific case of the Big Data environments.

This paper presents a specific risk pattern for Big Data environments to deal with risks related Big Data characteristics. MARISMA represents a comprehensive and consolidated methodological approach that, thanks to the automatic support provided by eMARISMA, and the adaptability properties provided by the use of risk patterns, allows us to tackle the current Big Data risks through a new risk pattern. In addition, these risks can be ready quickly for ongoing and future risks challenges —a request widely demanded by the scientific community (Zio, 2018). MARISMA-BiDa is based on the recommendations proposed by the ISO/IEC 27000-series standard, Big Data recommendations of ENISA (European Union Agency for Network and Information Security) (Rekleitis, 2016) and CSA (Cloud Security Alliance) (Murthy et al., 2014), and the Big Data reference architecture proposed by the NIST organization (U.S. National Institute of Standards and Technology) (NIST, 2015). They consider the inherent characteristics of this type of systems, such as extensive datasets, scalability, massively parallel processing, performance efficiency, non-relational models, computational portability, large scale analytics, etc. This specific pattern can be used as a base for instantiating the particularities of any Big Data system for any context (health, finance, government, education, manufacturing, etc.) given that the types of assets, threats, and risk in these environments are similar from one system to another. It is of prime importance to highlight that these contexts share the same pattern since they all use a Big Data system. The proposed pattern is instantiated by a typical Big Data scenario in a healthcare context (defined as a case study by the NIST) showing how accessible its instantiation is for any given Big Data context.

It is worth emphasizing that this pattern is focused on specific aspects of risk analysis for Big Data, it is not intended to cover all organizational risks that could indirectly affect the Big Data environment for risk management and control since it is assumed that it should be included in a more general risk management framework.

The rest of this work is organized as follows: First, a related work and background section is presented, followed by a section that introduces the main components of the MARISMA methodology and the meta-pattern, as well as the technological environment that supports it and a cost analysis. Next, the MARISMA-BiDa pattern is presented, defining each of its components. Then, a case study is defined and the results of the application of the proposal to this case study are shown. Finally, a section of conclusions and future work is included.

2. Related work and background

In this section, the current main frameworks, methodologies, and standards related to risk analysis and management are discussed, identifying some of their main weaknesses, and stating the way in which MARISMA tries to counteract them. Also, the different proposals that are related to Big Data systems and risk analysis are analyzed. Finally, the specific properties of Big Data ecosystems that justify the need for a new approach are explained.

Among the main proposals for risk analysis and management, the following can be highlighted: MAGERIT, OCTAVE, CRAMM, and the ISO/IEC 27005 (ISO/IEC 27005, 2018), COBIT (ISACA, 2019) or NIST standard (NIST, 2018b). MAGERIT implements the risk management process within a framework so that the governing bodies make decisions considering the risks arising from the use of information technologies. OCTAVE is a strategic planning and consulting technique in security that is based on technological risks. CRAMM covers all phases of risk management, from the actual analysis of risks to the proposal of countermeasures, including the generation of outputs for security documentation.

Regarding the main security management standards, many of them have attempted to incorporate risk analysis and management into their processes: the group of ISO/IEC 27000 standards, and specifically, the ISO/IEC 27005 standard establishes the guidelines for risk management in information security. The ISO/IEC 21827/SSE–CMM standard (ISO/IEC 21827, 2008) establishes a capacity and maturity model in systems security engineering, which describes the essential characteristics of the processes that must exist in an organization to ensure good security in systems, including in the previous phases a risk-oriented process. The ISO/IEC 15443 standard (ISO/IEC TR 15443-1, 2012) classifies the existing methods depending on the level of security and the phase of the assurance. The assurance evaluation is divided into process, product, and environment, while the phases of the risk analysis are design/implementation, integration/verification, replication, transition, and operation. The ISO 31000:2018(ISO 31000, 2018) provides guidelines on managing risks faced by organizations. It provides a common approach to managing any type of risk, it can be used throughout the life of the organization, and it can be applied to any activity, including decision-making at all levels. ITIL (Axelos, 2019) offers an element for a proper risk management: the updated and detailed knowledge of all the assets of the organization and the relationships, weights, and dependencies between them. This knowledge is managed by ITIL throughout the management process of the service support configuration, and by using the basic tool on which a coherent approach to the efficient management of IT is built, the CMDB (Configuration Management Database). COBIT is a methodology for the adequate control of technology projects, information flows, and the risks involved in the lack of adequate controls. It includes a process aimed at assessing risks, which focuses mainly on the criteria of confidentiality, integrity, and availability, and secondarily, on criteria of effectiveness, efficiency, compliance, and reliability. Finally, the NIST (NIST, 2018b) proposed a generic risk management framework that is applicable to any information system. As a result, we can assume that many standards and proposals exist for supporting the process of analyzing the risks of general information systems.

However, concerning the analysis and management of risk security in Big Data systems, no specific proposals have been found. Despite the fact that some papers discuss this issue, none of them can establish a specific methodology. In (Benjelloun and Lahcen, 2018), an analysis of different security controls in Big Data is conducted. Paryasto et al. (Paryasto et al., 2014) conclude that in order to evaluate the security risk in Big Data, the process proposed by the NIST could be adapted. Damiani (Damiani, 2015) defends the need to perform a specific risk analysis for Big Data systems due to the fact that its inherent characteristics can cause the emergence of new threats, such as massive data leaks. Conversely, most of the works that relate risk analysis with Big Data take advantage of this technology to analyze and manage the risks of other information systems (Ale, 2016; Choi and Lambert, 2017). In other words, they use Big Data as a tool, but not as a focus system on which to analyze and manage their risks.

Moreover, many authors criticize that current risk analysis methodologies (OCTAVE, NIST Family, CRAMM, ISO/IEC 27001:2013, CORAS (Lund et al., 2003), MAGERIT, MEHARI (MEHARI, 2010)) have multiple weaknesses that should be corrected by means of new scientific proposals. These main weaknesses, as identified in our research, are as follows: i) It is necessary to have adaptive catalogues with risk sectorial specialization (Garcia and Moreta, 2018) and technology (Rossebø et al., 2017) that is capable of being reused in different situations; ii) One notices the lack of the ability to organize risks via hierarchical and associative structures that are applicable to sectors within companies, groups of companies and other groupings (Zissis and Lekkas, 2012); iii) The methodologies and tools of risk analysis should allow for the reuse of knowledge and learning (Petrescu et al., 2019; Tubío Figueira et al., 2020); iv) The risk analysis should be capable of evolving dynamically once they have been established and so rapidly allow for the changes in the risks produced (Aviad et al., 2018; Haiwen and Xiaofang, 2019); v) The tools of risk analysis should allow for collaboration between companies so as to improve at a global level (Aviad et al., 2018); vi) The risk analysis should be based on a set of Key Risk Indicators (KRIs) which are precise (He and Lu, 2018); vii) The level of subjectivity in the evaluation of the elements of risk analysis should be minimized (Stergiopoulos et al., 2018); and, finally, viii) The process of analyzing the risks should be simple and affordable so as to put it within the reach of companies of all sizes (Hashim et al., 2018).

Our MARISMA framework includes some strategies aimed at counteracting these weaknesses. In particular, the key

concept of MARISMA is the meta-pattern which serves as a support for the risk patterns, as well as the potential to establish hierarchies, dependencies and inheritance upon these patterns, alongside the development of the support tool eMARISMA, provides a response to a fair number of these weaknesses. The tool not only allows for the management of the archives of risk patterns applicable to different companies, but it also permits the creation of hierarchic dependencies between patterns, inheritance, instantiation of patterns and the reuse of contents. It also includes an expert system which assists in the making of decisions about the recalibrating of the risk analyses, and a set of KRIs which help with the management of the procedure. As a result, we can tackle the process of risk analysis in a simpler way and for less cost.

In this scenario, our MARISMA framework can be considered a consolidated solution allowing for the incorporation of new risk patterns that carry out the risk analysis of specific contexts or technologies (e.g. Big Data), and, at the same time, counteracts some of the main weaknesses of classical risk analysis approaches.

Respecting Big Data ecosystems, as we mentioned previously, they can be regarded as a type of very complex environment, with different technologies and resources working together to extract useful insights for the companies. Big Data supposes a change from traditional systems in three different ways: the amount of data (volume), the rate of data generation and transmission (velocity), and the types of structured and unstructured data (variety) (Chen et al., 2014). These properties are known as the three basic V's of Big Data. Many authors have added new characteristics to this initial group, such as variability, veracity, or value (Khan et al., 2014). These specific dimensions, together with the variety of security domains to be considered (physical security, access control, business continuity, compliance, etc.), the necessary specific controls (big data asset inventory, information labeling and manipulation, access to data, algorithms and applications, control against malicious algorithms, etc.), the specific kinds of assets to be protected (data analytics algorithms and procedures, metadata, volatile data, several types of infrastructure, etc.), and the new threats associated with these environments, suggest the need to deal with the Big Data risks by means of a specific strategy. This strategy will probably be integrated into a global risk analysis approach for the company, but this allows us to make a much more accurate treatment of Big Data risks.

In fact, security reference architectures (SRA) specific to Big Data have been proposed in the scientific community (Moreno et al., 2019), identifying and relating the differentiating components of a Big Data and identifying the necessary bases to incorporate security in the life cycle of the environments of Big Data. These SRAs include concepts such as vulnerabilities, risks or threats to help the different stakeholders gain a better understanding of not only the system and its components, but also of the security aspects that should be addressed during the development of the systems, including the risk analysis process. These SRAs are based on Big Data architectures which are well consolidated in industry and in the scientific community, such as those of the NIST (NIST, 2015) or Microsoft (Microsoft, 2014).

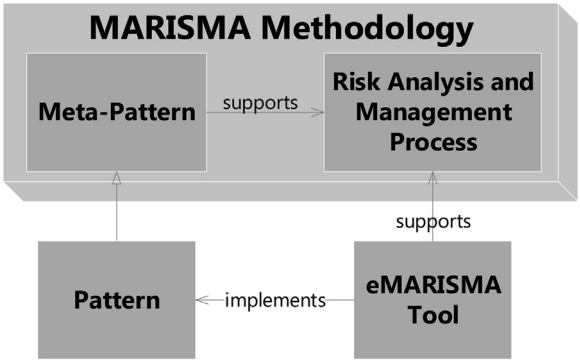


Fig. 1 – General Schema of MARISMA Methodology.

3. MARISMA methodology

A methodology can be defined as the conjunction of a data model that defines its main components and a process that defines a set of activities to conduct its objective. In the case of MARISMA, it is a risk analysis and management methodology that can be adapted to any type of IT environment (Santos-Olmo et al., 2016).

MARISMA's core consists of two components, the meta-pattern and the risk analysis and management process. The process is supported by an automated tool, eMARISMA, which helps in the analysis and decision making, as well as implementing a specific pattern based on the meta-pattern that makes MARISMA applicable to any IT environment. Hence, the two central components of the MARISMA methodology are reusable elements. What makes MARISMA adaptable and applicable to different contexts is the pattern, as different patterns for various contexts can be defined, inheriting the elements common to any risk analysis and management process defined in the meta-pattern. In the case presented, we define a pattern for the Big Data context. Fig. 1 depicts the main components of the MARISMA methodology. In the following subsections, these components will be explained in depth.

3.1. *Meta-pattern*

The risk analysis approach of traditional methodologies such as OCTAVE, NIST Family, CRAMM, ISO/IEC 27001:2013, CORAS, MAGERIT, MEHARI, is oriented to risk analysis based on elements such as Threats, Asset Types, and Vulnerabilities, considering that controls should not be conceived until the risk management phase. Syalim et al. (Syalim et al., 2009) demonstrated that this could be a drawback when performing a risk analysis, and controls should be taken into account not only in the management phase, but also in the risk analysis part.

Other authors highlighted the need for risk methodologies to be adapted to specific sectors (Garcia and Moreta, 2018), with unified frameworks (Al-ahmad and Mohammad, 2013; Chen, 2015; Fenz and Ekelhart, 2011; NATO, 2008; Shamala et al., 2013), prepared for decision support (Bergvall and Svensson, 2015), with updated risk taxonomies (Mukama, 2016; Nurse et al., 2017; Radanliev et al., 2018; Shameli-Sendi et al., 2016), and catalogues of structured

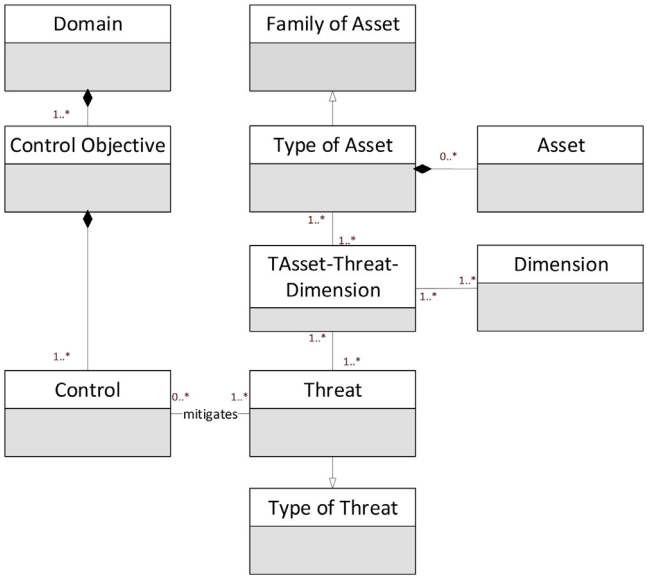


Fig. 2 – Meta-pattern in MARISMA.

elements that could be shared while conducting the risks analysis activities (Agrawal, 2017). The research also highlights the importance of risk analysis being straightforward and with practical orientation (Abbass et al., 2015; Bergvall and Svensson, 2015; Hashim et al., 2018; Oppliger et al., 2017; Pan and Tomlinson, 2016; Wangen, 2017).

As aforementioned, the research related to risk analysis is along the same line of seeking improvements and solutions to make the risk analysis process more efficient, robust, adaptable, and applicable. MARISMA is proposed taking these recommendations into account, in a way that controls are considered from the beginning of the risk analysis process. The adaptability to specific sectors and reuse of artifacts is achieved by defining the concepts of pattern; considering unified frameworks and taxonomies of risks incorporating security standards and recommendations; and decision making is supported by an automatic tool developed (eMARISMA).

As we have seen, research related to risk analysis goes in the line of trying to find improvements and solutions to make the risk analysis process more efficient, robust, adaptable and applicable. MARISMA is proposed with these recommendations in mind, in such a way that controls are taken into account from the beginning of the risk analysis process; adaptability to specific sectors and reuse of artifacts is achieved by defining the concepts of pattern; unified frameworks and taxonomies of risks are incorporated to MARISMA thanks to the use of security standards and recommendations; decision making is supported by an automatic tool developed (eMARISMA).

All these characteristics that provide improvements in risk analysis and management are supported by the definition of the meta-pattern within the MARISMA methodology. This meta-pattern, whose data model is shown in Fig. 2, defines the key elements of classic risk analysis (Threats, Types of Assets, Dimensions, etc.) where security controls have been integrated along with the control objectives and domains associated with them, to take them into account in the risk analysis phase. This structure is generic, and therefore reusable, where it can be taken into account both security standards to define domains and controls, such as ISO27001 as recommendations, and taxonomies of threats and types of security assets more common to any IT environment. To customize a security risk analysis in specific sectors, a pattern is defined. It inherits the common characteristics of the meta-pattern and it is adapted to a particular sector, for example, for Big Data—explained below.

Following our data model that defines all the concepts to perform risk analysis, the following elements can be seen in Fig. 2:

- Domain: Domains are functional areas of security. For example, one domain to refer to organizational aspects, another domain to refer to legal aspects, and another domain for technological aspects of information security.
- Control Objective: For each of these domains, a series of control objectives that reflect what is intended to be achieved through the implementation of each of the domains are defined. The control objectives are a group of controls that share the same objective against threats.
- Control: The concept of 'control' groups all the actions, documents, procedures, and technical measures adopted to ensure that each threat identified and assessed with a certain risk is minimized, and it aims to achieve the various objectives. The controls allow to stop or mitigate the threats. They can usually be extracted from international standards (e.g.: COBIT, ISO 27001, NIST, etc.)
- Types of Asset: The assets are the resources (physical, software, documents, services, people, facilities, etc.) that add value to the organization and therefore, they need to be protected from potential risks. These assets are categorized by types of asset, which are likewise grouped by family of asset according to their nature. Thus, for example, '*infrastructure*' is one family of assets, whose types of assets could be software, hardware, storage, servers, etc.
- Dimension: A dimension is an inherent characteristic of an asset that can be impacted by a threat, producing a percentage of degradation in the value of the asset (e.g.: confidentiality, availability, complexity, flexibility, etc.)
- Types of Threat and Threat: A threat is a potential situation that damages an asset or control implemented in an organization with a certain probability of occurrence. The threats depend on the type of risk analysis to be performed and they can be grouped into types. For example, the threats of '*loss of information*' and '*destruction of records*' belong to the type of threat of '*unintentional damage*'.
- The Control-Threat Relationship: this relationship establishes how the different existing controls can mitigate the threats identified in the system. This relationship will allow recommendations for the controls needed to protect the system from the threats that can attack it.
- The Types of Asset-Threat-Dimension relationship: in a process of risk analysis and management, the identified threats that can attack a system are focused on one or several types of assets with the objective of damaging the asset by reducing its value. This threat damages the asset by having a direct impact on some of its dimensions. Therefore, for example, the '*destruction of records*' threat attacks

on the asset type of '*Infrastructure*' only for the '*availability*' dimension, making the records no longer available, and thus, reducing their value by 100%. Hence, this relationship or matrix defines what threats can attack a particular type of assets and what is the impact on the dimensions.

This meta-pattern is a generic structure where all the elements and relationships considered are involved in any risk analysis and management process. It is defined in a way that controls, objectives, and domains are taken into consideration in order to comply with many of the security standards or norms, such as ISO27001, COBIT, etc. All the concepts and their relationships are defined to consider the dependencies among the elements when performing any risk analysis and their evaluation, to make it more easily automated, as well as being reusable for any specific sector. Thus, in the case of a sector such as critical infrastructures, following the IEC 62443, NIST-800–82, etc. standards, a pattern with the domains, objectives, and controls defined in that standard is specified. Also, the threats, assets, and general dimensions that usually intervene in this type of environment are identified, and finally the relationships among them. This way, the client will have a pattern for a particular sector, and more specifically for critical infrastructures, which can be used to instantiate it in specific critical systems, such as the energy system, water supply, emergency, etc., All of them share the same pattern; the critical infrastructures pattern, but they detail each specific system to analyze.

A pattern for Big Data environments will be defined in the next section, and Section 5 will instantiate this pattern for a case study of a health system.

3.2. *Risk analysis process*

Organizations, regardless of their size, must be aware of the importance of the IT risks and how they should be managed. ISO 31000 (ISO 31000, 2018) defines risk management as an organizational process which should involve the systematic application of policies, procedures and practices to the activities of communicating and consulting, establishing the context and assessing, treating, monitoring and reviewing risk. All these activities must be controlled by the internal and external stakeholders of the risk analysis and management process (ISO 31000, 2018; Kelemen et al., 2016).

MARISMA defines a risk analysis and management process aligned with international security standards. On the one hand, it performs the analysis identifying assets and potential threats, assessing the impact and probability of occurrence of the threats, and identifying the most appropriate safeguards to evaluate the risk. On the other hand, in the risk management, the most appropriate controls are selected, the events are managed, and the evolution of the risks is monitored, as well as the treatment of the risk. In this paper, we focus on the first part; risk analysis, defining a context-independent and reusable structure to facilitate the risk analysis process and its subsequent management. The success of the process is based on the reuse of patterns, which are knowledge structures with common characteristics for a specific context, for example, a risk pattern for Big Data, a risk pattern for critical systems, or a risk pattern for web services. These patterns have a common base; the meta-pattern, where the common elements for any risk analysis and management process are defined (as defined in the previous section).

The analysis process of MARISMA is shown in Fig. 3. The process begins when a risk analysis and management of the system is performed, and it ends in the analysis phase, presented in this paper—when a risk assessment of the system is achieved. In this phase, an existing pattern must be instantiated, or a new pattern must be generated if it has not been previously created for a particular environment. The creation of a new pattern is based on the meta-pattern where the essential elements are obtained in order to conduct the risk analysis. Once the pattern is ready, the three stages of the risk analysis phase will be conducted, which are; identification, analysis, and evaluation. Each one of these tasks or stages is conducted from three points of view, where each of them is in charge of conducting it. For MARISMA, the points of view are the client, the security expert, and the support tool.

The task for each viewpoint is detailed subsequently: the *Client* viewpoint defines all the iterations and the work that the client must perform during the identification and analysis of system risks (identifying assets, threats, controls, including probability and impact, etc.) It is supported by the eMARISMA tool. The *eMARISMA* viewpoint helps the client to calculate the risk level and its evaluation to take decisions. The *eMARISMA* viewpoint is in charge of automatically managing all the information provided by the client throughout knowledge patterns, instantiating the pattern for a specific context, and making assessments to calculate the risk level of the system. The third viewpoint is the *Expert* one, whose unique important mission in the analysis process is to generate a new non-existent pattern to be used for risk analysis, which is based on the main elements and concepts defined in the meta-pattern as aforementioned, but adapted to the characteristics and particularities of a specific context, for example, a Big Data context.

After that, the system is ready to manage the security event that must be communicated by the client. These security events generate knowledge to adapt the levels associated with the elements of risk management, making the risk dynamically recalculate, and also adapting the elements associated with the selected pattern allowing its evolution. This is part of the risk management process, so it is beyond the scope of this paper, as what is presented in this paper is the risk analysis process.

3.3. *eMARISMA tool*

To support the MARISMA methodology, a tool called eMARISMA was developed using a Software as a Service architecture (see Fig. 4) and using Java stack technologies. This tool supports all the processes of the methodology.

The tool is comprised of two differentiated environments which are located separately to increase their security, but which have a continuous information exchange to facilitate feedback, improvement and learning between them.

The first of these (Fig. 4, right), deals with the patterns management (based on our meta-pattern) and is divided into two sub-systems, one which supports the web and business layer, and the other which contains the different patterns and the knowledge that these acquire from the instantiation of the

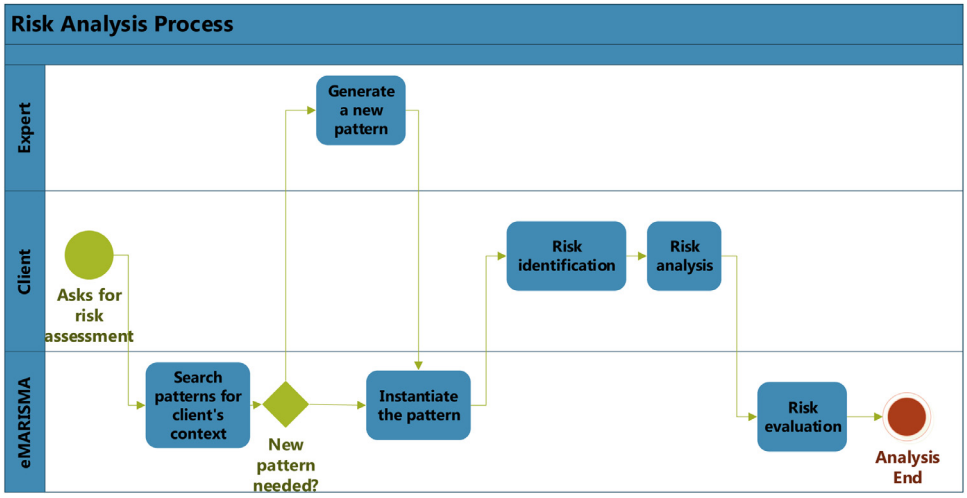


Fig. 3 – General outline of MARISMA Risk Analysis process.

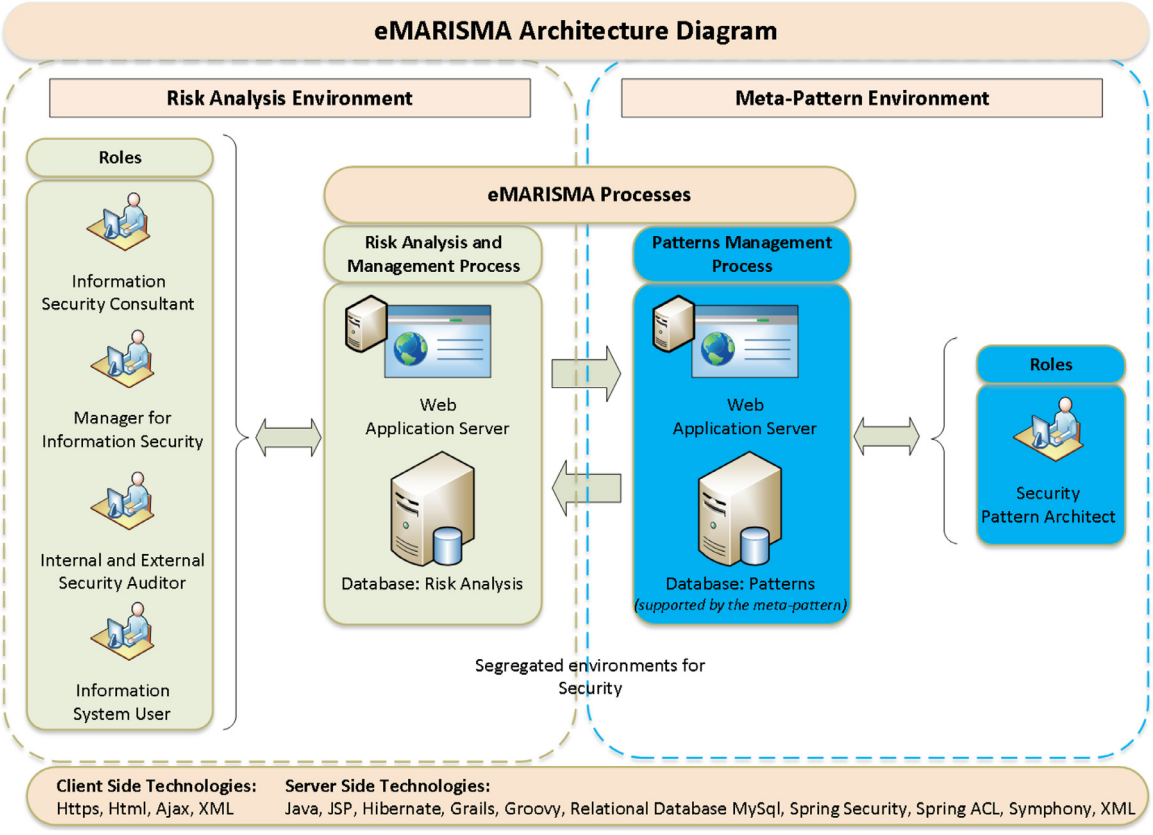


Fig. 4 – eMARISMA Architecture Diagram.

same. Only the architects of the pattern may access this environment, to create new patterns or update the existing ones. The following are some of the functions of this component: i) The meta-pattern defined is reused whenever required to create new patterns, and these new patterns can be used as many times as needed to be adapted to a specific context; ii) To visualize the different existing patterns and to use them as a basis to create other patterns (e.g., sectorial patterns) and iii) To allow the incorporation of elements used in risk analysis and management such as assets, threats, controls, domains, etc.

The second environment (Fig. 4, left), deals with the management of risk analysis, and has two sub-systems which support the web component and the database. In the latter, the information relating to the different risk analyses is stored. Access to this component is given to all stakeholders whom

the client regards as involved and who are necessary for the carrying out of the risk analysis. The following are some of the functions of the tools associated with this environment: i) To obtain a detailed map of the current situation (risk identification) and a plan for recommendations on how to improve it (risk evaluation); ii) The system automatically performs a risk evaluation and it calculates the most appropriate risk treatment plan to allow the company to reach a level of risk within the defined limits in the optimal way; iii) The tool always represents a dashboard with the security levels reached by the company so as to monitor the risk of the company in real time and iv) The results obtained automatically support decision-making and enable one to choose the best way to protect the system and treat the given risk.

This tool allows the integration of several risk patterns. In particular, it integrates patterns for security management (generic) and critical infrastructures, and now also our MARISMA-BiDa pattern for Big Data ecosystems. This tool is being used by multiple national and international companies, since it offers a mature technology that has demonstrated its utility in improving the risk analysis process.

3.4. *Cost analysis*

One of the main problems inherent in the classic techniques of risk assessment is the excessive complexity in its application in companies and the high cost thereof (Korman et al., 2014; Kouns and Minoli, 2010; Macedo and Neto, 2009; Pandey, 2012; Shukla and Kumar, 2012; Wangen and Snekkenes, 2013). This complexity and cost can be absorbed by large companies but not by Small and Medium-Sized Enterprises (SMEs), as these lack the necessary resources (Agrawal, 2017; Tatiana et al., 2017). As a result, many authors highlight as critical factors for systems of analysis and management of risks their procedural efficiency and cost (Rot, 2009; Sánchez et al., 2010).

Addressing this necessity, the MARISMA framework has been conceived so as to carry out the analysis of risks in an agile way and consuming fewer resources. This has been made possible by the use of patterns in a common structure (meta-pattern), and by the reuse of the knowledge supplied by these patterns.

In Section 3.2 the procedure of risk analysis of MARISMA is detailed (see Fig. 3), listing the viewpoints involved and the tasks that are carried out in each one of these. The calculation of the cost of the procedure is obtained principally from the costs associated with the two tasks defined within the client viewpoint, since the costs involved in the remaining tasks can be regarded as negligible for a company carrying out a risk analysis. The analysis of the cost of each one of these viewpoints is as follows:

- Expert viewpoint: these are the consultants who along with the investigators develop the new patterns. This is carried out once only and is then provided to the companies so that they can conduct their own risk analyses. As such, the creation of new patterns, although involving a high cost for its creators, does not result in additional costs for the companies.
- Client viewpoint. The client, with the help of eMARISMA, identifies and sets the parameters of the elements required by the respective pattern (assets, controls, threats, dimensions, etc.).
- eMARISMA viewpoint. The license cost of this tool (in the non-free version) is of little relevance in comparison to the total cost of conducting a risk analysis.
- The estimate of the total cost involved in the development of a risk analysis using the MARISMA framework is arrived at using the following formula:

$$Total\ Time = Total\ Time\ RI + Total\ Time\ RA$$

$$Total\ Time\ RI = \sum SubTask\ (Time \times Number\ of\ Elements) \times \%Readjustment$$

$$Total\ Time\ RA = (Time \times Number\ of\ Registers) \times \%Readjustment$$

Where:

- Total Time RI: is the time required to carry out the task "*Risk Identification*" (see Fig. 3).
- Total Time RA: is the time required to carry out the task "*Risk Analysis*" (see Fig. 3).
- Time: is the estimated time taken by a client to carry out an action upon the system.
- Number of Elements: is the number of elements associated with each one of the subtasks in the task "*Risk Identification*".
- Number of Registers: is the number of registers obtained from the valid combinations of the formula [assets x threats x dimensions]
- %Readjustment: is the percentage of the elements in each subtask which are subject to readjustments in relation to the values recommended by the pattern.
- In the following section we demonstrate an example of how the cost formula of risk analysis is applied:
- Risk Identification Task: This risk analysis is established using 40 assets, 150 controls and 50 threats, with a reuse of the knowledge of the pattern of 40% for the three subtasks, with an average time of 4 min per element taken for the subtasks of setting the parameters of controls and threats and 10 min for the subtask of the identification of assets (see Fig. 5).

- Risk Analysis Task: Based upon the preceding elements, and the relations existing in the pattern, one obtains 1600 viable and configurable candidates (assets x threats x dimensions), with there being 5 dimensions for this example (which will depend on the pattern used in each case), a reuse of knowledge of the pattern of 75%, and an average time taken to adjust each element of 1 min (see Fig. 6).

As a result, to carry out a risk analysis with the elements defined in the example would take some 1.240 min (approximately 20 h). This means an investment of resources that

*Total Time RI = (10 * 40) * 60% + (5 * 150) * 60% + (5 * 50) * 60% = 840 minutes.*

Where:

- *Assets subtask. Time = 10*
- *Assets subtask. Number of elements = 40*
- *Assets subtask. %Readjustment = (100 – 40%) = 60%*
- *Control subtask. Time = 5*
- *Control subtask. Number of elements = 150*
- *Control subtask. %Readjustment = (100 – 40%) = 60%*
- *Threats subtask. Time = 5*
- *Threats subtask. Number of elements = 50*
- *Threats subtask. %Readjustment = (100 – 40%) = 60%*

Fig. 5 – Total Time for Risk Identification Task.

*Total Time RA = (1 * 1.600) * 25% = 400 minutes.*

Where:

- *Time = 1*
- *Number of registers =1.600*
- *%Readjustment = (100 – 75%) = 25%*

***Total Time** = 840 + 400 = 1240 minutes*

Fig. 6 – Total Time for Risk Analysis Task and Total Time.

would be acceptable for all types of companies. To conduct the estimates of time and the percentage of reuse of knowledge, the values obtained from the latest 100 analyses of risk carried out using the eMARIMA tool have been used.

4. MARISMA-BiDa pattern

As previously mentioned, the meta-pattern defined in MARISMA contains the necessary elements to conduct a general risk analysis and management process. This meta-pattern, which is used in the analysis phase, is context-independent and therefore, it can be used to perform a risk analysis and management of any system. To take into account the context, the organization, and the system to be protected, a context-dependent pattern must be defined, where the meta-pattern elements most oriented or focused on the system environment being analyzed must be selected. For this reason, to conduct a risk analysis and management in a specific environment such as Big Data, the elements of the meta-pattern focused on Big Data must be selected, obtaining a new specific pattern with the characteristics and particularities of this type of environment, which we have called MARISMA-BiDa. The creation of this new pattern oriented to Big Data is conducted by the expert user. The process of creating a new pattern can be seen in the general outline of MARISMA shown in Fig. 3, defined in the viewpoint expert.

The elements of the meta-pattern inherited by the MARISMA-BiDa pattern is based on the identification and definition of the specific elements for Big Data, which are explained below.

4.1. *Domains, control objectives and controls*

Domains, control objectives, and controls are usually obtained from international security standards or norms, where those that best suit and relate to the environment to be analyzed must be identified and selected. In the case of Big Data, there is no standard for exclusive security management, and consequently, we can rely on ISO27001, COBIT, ITIL, NIST800–53, etc., and adapt them to the particularities of the context. In this case, the ISO/IEC 27000 family has been considered for Big Data because it has a wide and general set of domains, control objectives and controls (14 domains, 35 control objectives and 114 controls) that are easily adaptable to the characteristics of Big Data. This standard is used as a reference for many other specific standards where domains and controls are defined. For example, for the context of industrial control, the ISO 62443 is defined, or the ISO 27799 defined for the health industry.

For the MARISMA-BiDa pattern, the reference standard of ISO 27001 is used as a base, as it is a reference and a general scope. This standard defines a broad set of domains that cover a global management of the security of an organization. It has organizational domains, human resources, and policies, which are independent of the system. For our adaptation of the standard to Big Data there will be a special focus on the domains that manage the most technical aspects and data management, which are those that can be adapted to a Big Data system, and that can be seen in Fig. 7.

Out of all these selected domains, they must be adapted to a Big Data environment, considering the characteristics of these systems. Therefore, for each domain, the control objectives and controls that are applicable to a Big Data environment are analyzed considering four potential actions: 1) eliminate the control because it is not applicable in these environments, 2) add the control without any modification because it is a general control and it can be applied to any environment, including Big Data, 3) add a control of the existing ones in the standard but with some modification to guide them and adapt them to a Big Data context, and 4) add a new control because it is not considered in the standard and we believe it necessary to protect a specific asset from a specific threat in Big Data.

The result of this analysis is shown in Table 1. In this table, in the first two columns, the domains, and the related control objectives that influence a Big Data environment are depicted. In the third column, all the controls belonging to those control objectives are shown, indicating in brackets the action taken on each of them. Thus, a (N) means 'new control', a (A) means 'adapted with respect to a standard control', and a (NM) means 'not modified, existing control in the standard'.

Thus, for the adapted controls, we can find all the controls of the objective '*Liability for Big Data assets*' from the domain '*Asset management*' where they have been adapted to the specific and exclusive assets of Big Data. Two controls of the '*User access management*' objective from the '*access control*' domain have also been adapted, focusing on the provision of access and the management of access privileges of the different roles that usually participate in Big Data environments. In the '*Security of operations*' domain, there are several controls adapted as it is the control of the installation of algorithms in operation and audit controls, which are adapted to the different algorithms of these environments and to the internal functioning of Big Data for auditing activities. The last adapted control belongs to the domain '*Security of communications*', whose control is '*Security of network services in Big Data*', where the security

Table 1 – Domains, Objectives, and Controls of the MARISMA-BiDa Pattern.

Domain	Objective	Controls
Asset management	Liability for Big Data assets	Big Data asset inventory (A)
		Ownership of Big Data assets (A)
		Acceptable use of Big Data assets (A)
		Return of Big Data assets (A)
	Classification of information	Classification of information (NM)
		Information labeling (NM)
		Information manipulation (NM)
Access control	Business requirements for access control	Access control policy (NM)
		Access to networks and network services (NM)
	User access management	User registration and deregistration (NM)
		Access provision of typical Big Data roles (A)
		Management of privileges access to typical Big Data roles (A)
		Management of secret user authentication information (NM)
		Review of user access rights (NM)
		Withdrawal or reassignment of access rights (NM)
	User responsibility	Use of authentication secret information (NM)
	Control of access to systems and applications	Restriction of access to information (NM)
		Secure login procedures (NM)
		Password management system (NM)
		Use of privileged utilities (NM)
		Access to the creation/modification of data collection algorithms (N)
		Access to the creation/modification of data preparation algorithms (N)
		Access to the creation/modification of data analysis algorithms (N)
		Access to the creation/modification of data visualization algorithms (N)
		Access to creation/modification of control of access to information (N)
Cryptography	Cryptography controls	Policy for the use of cryptographic controls (NM)
		Key management (NM)
Physical and environmental security	Equipment security	Equipment location and protection (NM)
		Wiring security (NM)
		Maintenance of equipment (NM)
		Safe reuse or disposal of equipment (NM)
		Assurance policy of server availability (N)
Security of operations	Protection against malicious algorithms	Controls against malicious algorithms (N)
	Backup copies	Have backup copies of the information (NM)
	Control of algorithms in operation	Installation of the algorithms in operation (A)
		Supervision of operating algorithms (N)
	Big Data Audit Considerations	Big Data Environment Audit Controls (A)
Security of communications	Network security management	Network controls (NM)
		Security of network services in Big Data (A)
		Segregation of networks (NM)
Acquisition, development, and maintenance of information systems	Security requirements in information systems	Analysis of requirements and specifications of information security (NM)
		Protection of application services transactions (NM)
	Security in development and support processes	Safe development policy (NM)
		System change control procedure (NM)
		Technical review of applications after making changes in the OS (NM)
		Restrictions to changes in software packages (NM)
		Principles of security systems engineering (NM)
		Security development environment (NM)
		Software development outsourcing (NM)
		System security functional tests (NM)
		System acceptance tests (NM)
	Test data	Protection of test data (NM)
Relationship with suppliers	Security in relations with suppliers	Information security policy in relations with suppliers (NM)
		Security requirements in contracts with third parties (NM)
		Information and communication technology supply chain (NM)
		Assurance of legal compliance of suppliers (N)
	Management of the supplier's service provision	Control and review of the provider's service provision (NM)
		Management of changes in the provider's service provision (NM)

(*continued on next page*)

Table 1 (*continued*)

Domain	Objective	Controls
Aspects of info. security for management of business continuity	Redundancy	Availability of information processing resources (NM)
Compliance	Compliance with legal and contractual requirements	Identification of applicable legislation and contractual requirements (NM) Protection and privacy of personal information (NM) Regulation of cryptographic controls (NM)

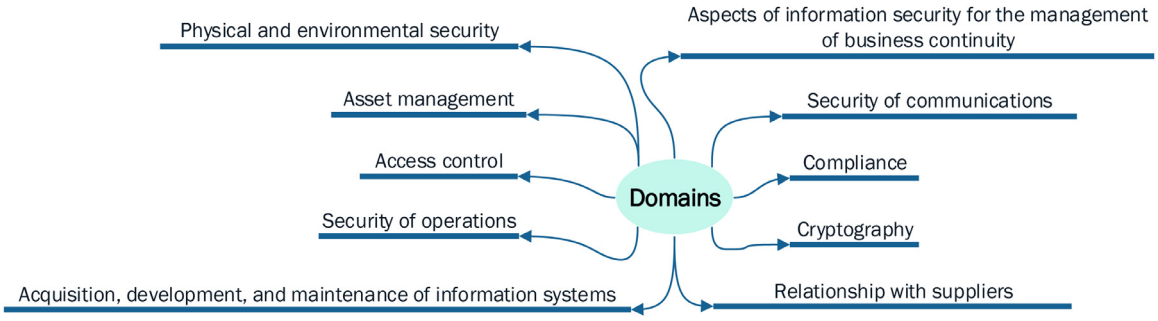


Fig. 7 – Domains for Big Data.

mechanisms and requirements of the network services used in these distributed environments must be specifically identified.

With regards to the new controls to be added to the set of controls categorized by domains, we can find all the access controls for the creation or modification of data for the different processes within Big Data system (collection, preparation, analysis, visualization, control of access to information), which belong to the access control domain. Another new control is '*Assurance policy of server availability*' from the objective '*Equipment security*', since the availability of the servers that are part of the Big Data must always be available for continuous processing. Additionally, for the '*Security of operations*' domain, new controls have been added ('*Controls against malicious algorithms*' and '*Supervision of operating algorithms*'). The former implements detection, prevention, and recovery to protect the malicious code of the specific algorithms used in Big Data, and the latter manages and supervises these new analysis and processing algorithms part of Big Data. The last control added belongs to the domain '*Relationship with suppliers*' and it is the control '*Assurance of legal compliance of suppliers*' since while working with data and information from different sources, they are governed by laws and regulations that must be revised.

4.2. *Types of assets*

The most important institutional organizations, such as NIST (NIST, 2018a) together with ENISA (Rekleitis, 2016) and CSA (Murthy et al., 2014) have studied the taxonomies and reference architectures in Big Data, which allowed to specify the different types of assets. These documents indicate a set of asset types typical from Big Data systems. A set of asset types that are more relevant for the pattern have selected and coincide in all the references studied. These asset types are categorized into five families: Infrastructure, Data, Individuals and roles, Big Data analytics, and Security and privacy techniques. Fig. 8 shows the types of assets grouped by family of assets to be incorporated in our pattern MARISMA-BiDa.

Next, the family and types of assets for Big Data are explained in depth below. The types of assets to be incorporated in our pattern are:

- Infrastructure: this family of type of assets represents the necessary resources for the environment to perform any type of computing. This includes any storage system (relational, NoSQL, NewSQL, etc.), different computing paradigms (batch mode, real time, or streaming data), hardware resources (servers, data centres, physical devices, virtual machines, etc.) and the software itself (operating systems, firmware or applications).
- Data: this is the main asset type family of a Big Data environment. It includes the data itself that can be structured, semi-structured, and unstructured. and it also includes its metadata. The streaming or volatile data is included as well.
- Big Data Analytics: this category includes the different protocols and algorithms (such as machine learning or data mining algorithms) to perform data analysis on Big Data and the visualization of results.
- Security and privacy techniques: this family includes any resource related to security, such as any document with the best security practices, security policies, audits and logs, cryptography methods, or information on how to perform access control.
- Individuals and roles: any type of actor that interacts with the Big Data environment throughout its life cycle, as it can be the data provider, data consumer, and operational roles such as application providers.

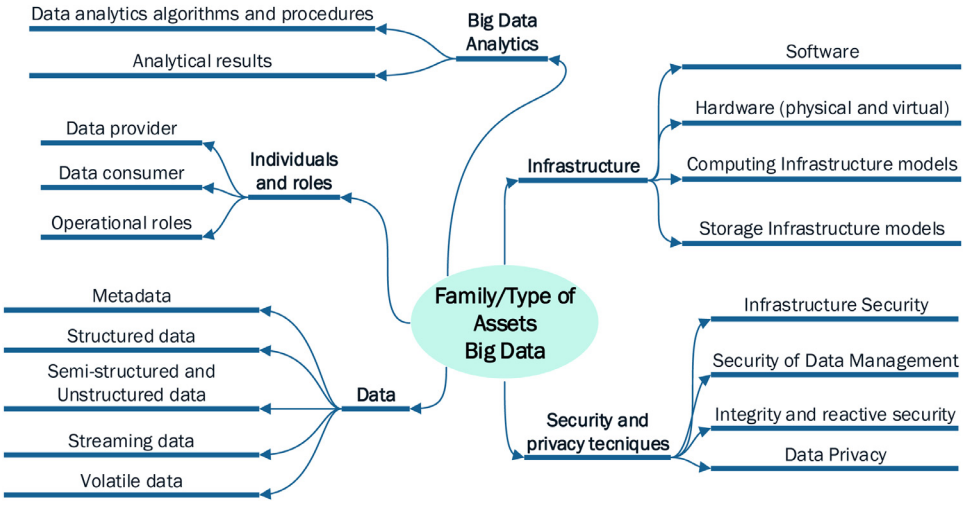


Fig. 8 – Family and types of assets for Big Data.

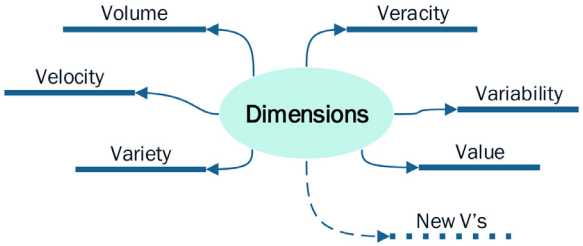


Fig. 9 – Dimensions for Big Data.

4.3. *Dimensions*

For the dimensions of the MARISMA-BiDa pattern, the different Vs typical of Big Data systems have been considered (see Fig. 9). Initially, 3 basic Vs were defined (Chen et al., 2014):

- Volume: refers to the amount of data created from different data sources.
- Velocity: is the rate of data generation or data flow. The speed at which these are created and should be analyzed. It can be in real time.
- Variety: describes the organization of the data, whether the data is structured, semi-structured, or unstructured at the moment, the data is not only in a structured format, such as in a database, but also in many forms such as audio, images, or documents.

These series of characteristics complicate the analysis of the data by means of traditional systems. In addition to these characteristics, several authors have added new Vs that add complexity while clarifying the definition of a Big Data system. Unlike the situation with the 3 basic Vs, the new Vs have no general consensus. Among all the proposals, these three new Vs have been highlighted (Khan et al., 2014; Sun et al., 2019):

- Veracity: the data and data sources from which they are obtained must be reliable so that the resulting information is truthful.
- Variability: the data is not only in different formats, but also constantly changing.
- Value: probably one of the most important Vs. This fact must not be forgotten since the objective of Big Data is to obtain useful information from the available data. A common mistake in organizations is to establish a Big Data system as a trend, without being clear about its main objective.

Therefore, in a first proposal, these six dimensions are considered for the MARISMA-BiDa pattern (see Fig. 9). The MARISMA-BiDa pattern is designed to be extensible and to easily incorporate new V's if necessary, such as Visualization, Validity, Volatility, or Vendee (Patgiri, 2018; Patgiri and Ahmed, 2016). The eMARISMA tool automatically supports the incorporation of new dimensions in the risk analysis process for Big Data.

These dimensions are part of the MARISMA-BiDa pattern, leaving the possibility of incorporating new dimensions in the future if needed. The dimensions are added to the eMARISMA tool to define the pattern and to be able to establish the relationships and complete the matrices that this tool supports to conduct the analysis of risks for Big Data. Fig. 10 shows the incorporation of the dimensions in the MARISMA-BiDa pattern using the eMARISMA tool.

4.4. *Type of threats*

To define the different types of threats that can affect a Big Data environment, a study of taxonomies of threats for Big Data has been conducted. The recommendations given by ENISA are considered the most complete and relevant ones to define threats in Big Data. In (Rekleitis, 2016), an overview, along with a good practice guide on how to manage threats in Big Data, is created. This recommendation defines the most common threats that can occur in a Big Data environment grouped into 5 types: Unintentional damage; Eavesdropping, interception or hijacking; Nefarious activities or abuse; Legal; and Organizational (see Fig. 11).

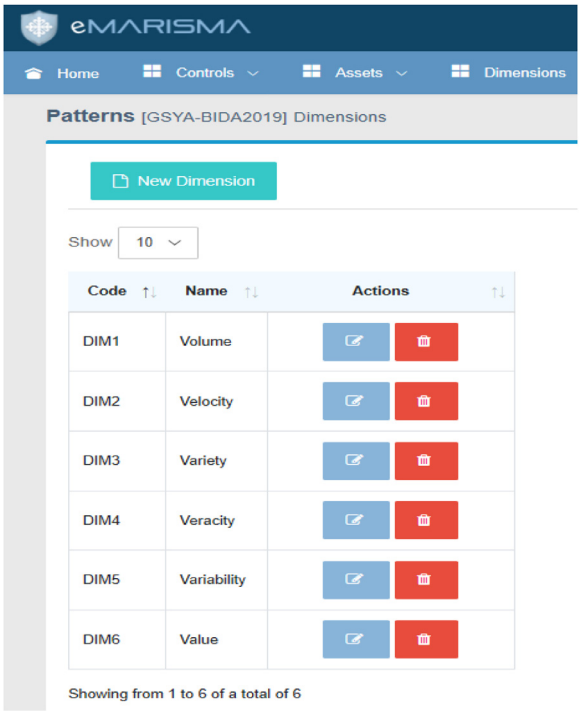


Fig. 10 – Dimensions defined on the eMARISMA tool.

Each of these types are general types of threats that can also affect a Big Data system.

- Unintentional damage: this type of threat includes sharing or leaking information due to human errors, unintentional intervention, or misuse of systems administration.
- Eavesdropping, interception or hijacking: this group includes those threats due to the alteration or manipulation of communications between two components. These attacks do not require the installation of new tools or software on the victim's infrastructure.
- Nefarious activities or abuse: this set of threats includes all those derived from criminal activity. In contrast to the previous type, these threats usually require the attacker to take some prior actions to modify the victim's infrastructure through the use of specific tools and software. An example of this type of threat is a denial of service (DoS) attack.
- Legal: this type of threat includes all the threats that occur due to the legal implications of a Big Data environment, such as non-compliance with laws or regulations, non-compliance with contract requirements, or unauthorised access to resources of intellectual property.
- Organizational: this type of threat includes all those that come from the organizational sphere, such as the lack of trained personnel that manages the Big Data environment.

4.5. *Type of asset - Threats – dimensions relationship*

As it can be seen from the meta-pattern in the Fig. 2, there is a relationship between threats, asset types, and dimensions. This relationship exists because threats attack certain assets (and not all of them) and have an impact on those assets that affect certain dimensions (and not all of them). For example, a DoS (Denial of service) threat could affect many asset types of the Infrastructure family (among others) but it would affect the '*Velocity*' dimension in Big Data because it would stop processing real-time information and therefore, the information analysis and storage services would cease to be available until the processing can be recovered.

This is the reason why a detailed analysis of the threats, of the assets type that are vulnerable to these threats, and of the dimensions that can be affected by the materialization of this threat on the asset, is necessary. A matrix of relation between these three elements of the risk analysis is defined in Table 3.

In this matrix, the different threats that can be identified in this type of systems can be seen, in which the different types of threats, both the different dimensions of Big Data and the affected asset types are integrated. If a threat does not fit into any of the cells of the matrix, it can be concluded that it is not a typical threat of a Big Data environment.

Table 3 shows the assets family that are part of a Big Data systems in the first row, which are affected by the threats considering the dimensions they affect. The first column shows the threats that can attack Big Data systems—taking into account the assets type (grouped by family to facilitate the analysis) —and each cell identifies the dimensions per asset family that are affected by a threat (all these elements previously defined in the pattern).

This relationship has been defined by conducting a broad study of both the threats and the dimensions without losing sight of the context, and the asset types mostly affected as a consequence of the materialization of these threats. With the experience of the research team in these subjects, together with the collaboration of security experts and with the help of

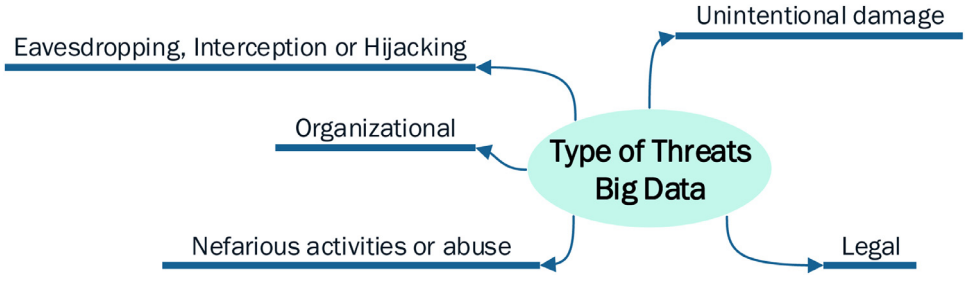


Fig. 11 – Types of threats for Big Data.

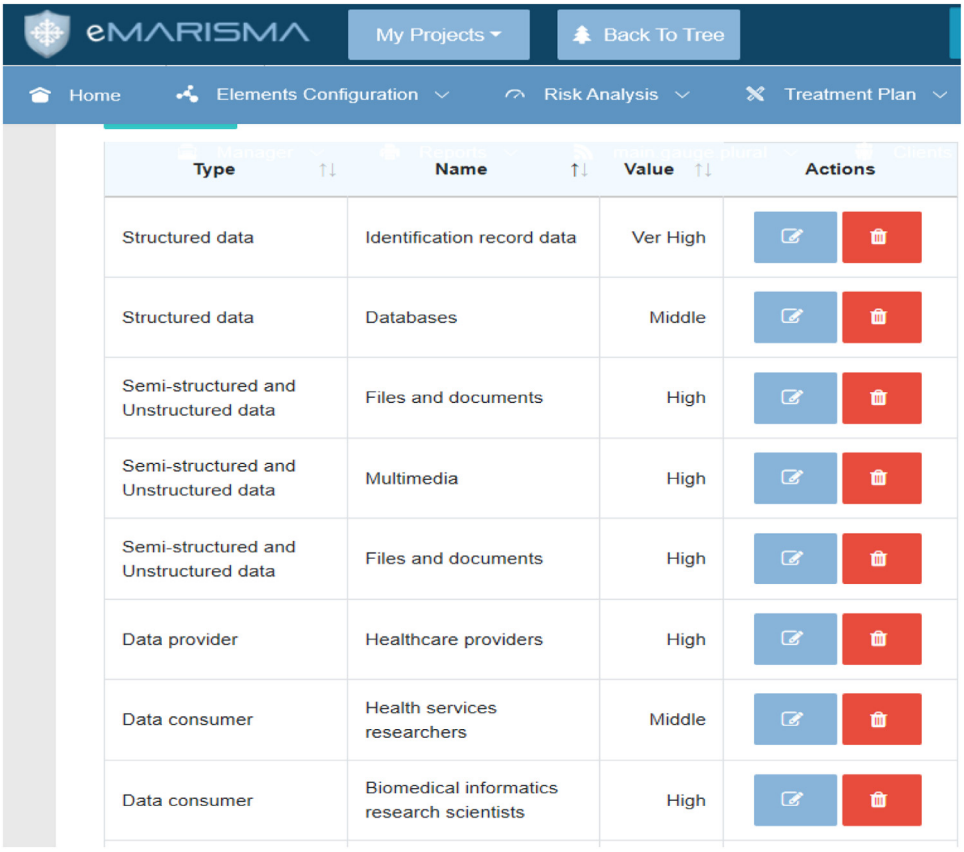
Table 2 – Types of threats and threats of the MARISMA-BiDa pattern.

THREATS TYPES	THREATS
Unintentional damage	Information leakage/sharing due to human error
	Leaks of data via Web applications (unsecure APIs)
	Loss of devices, storage media and documents
	Loss of (integrity of) sensitive information
	Loss of information in the cloud
	Damages resulting from penetration testing
	Damage caused by a third party
	Inadequate design and planning or incorrect adaptation
	Unintentional change of data in an information system
	Erroneous use or administration of devices and systems
	Using information from an unreliable source
	Destruction of records
Eavesdropping, interception or hijacking	Network Reconnaissance, Network traffic manipulation and Information gathering
	Intercepting compromising emissions
	War driving
	Interception of information
	Interfering radiation
	Replay of messages
	Man-in-the-middle/Session hijacking
Nefarious activities or abuse	Abuse of Information
	Social Engineering
	Malicious code/ software/ activity
	Abuse of authorizations
	Brute force
	Failed of business process
	Denial of service
	Targeted attacks (APTs etc.)
	Receive of unsolicited E-mail
	Remote activity (execution)
	Identity theft (Identity Fraud/ Account)
	Hoax
	Compromising confidential information (data breaches)
	Generation and use of rogue certificates
	Manipulation of hardware and software
	Manipulation of information
	Misuse of information/ information systems (including mobile apps)
	unauthorised activities
	unauthorised installation of software
	Misuse of audit tools
Legal	Violation of laws or regulations
	Failure to meet contractual requirements
	unauthorised use of ipr protected resources
	Abuse of personal data
	Judiciary decisions/ court orders
Organizational	Skill shortage

The particularity of Big Data systems, regarding threats that can affect the system, can be seen in the specific threats for each of the types defined. These specific threats oriented to Big Data (defined by ENISA) can be seen in the Table 2.

the ENISA reports, it has been possible to complete this relationship satisfactorily.

For example, one type of threat is the "interception of information" of the "Eavesdropping", where the interception of information in communications between Big Data applications can occur, and where the communication protocols are rarely secure between this type of applications (without the use of TLS and SSL). It is considered that this type of threat affects the “Veracity" and "Value" dimensions to a greater extent for the asset family “Data”. We consider that it affects the "Veracity" dimension because when intercepting the information on the network (e.g., it has been modified) it is not possible to be sure whether that information is accurate, which may lead to making the wrong decision. It also affects the “Value” dimension because if someone has been able to intercept the communication, they may have revealed the hidden information of the data, and consequently the value is lost. In addition, this threat also affects types of assets within the ‘Infrastructure’ and ‘Individuals and roles’ family in the ‘Veracity’ dimension, since the network and communication protocols have been intercepted. Hence, user sessions or unauthorised access could be obtained, so that they are no longer reliable. This type of study and reasoning for each of the type of threats and threats defined for our MARISMA-BiDa pattern has been conducted, the results of which are shown in Table 2.



Type	Name	Value	Actions
Structured data	Identification record data	Ver High	
Structured data	Databases	Middle	
Semi-structured and Unstructured data	Files and documents	High	
Semi-structured and Unstructured data	Multimedia	High	
Semi-structured and Unstructured data	Files and documents	High	
Data provider	Healthcare providers	High	
Data consumer	Health services researchers	Middle	
Data consumer	Biomedical informatics research scientists	High	

Fig. 12 – Some assets for the case study added on the eMARISMA tool.

5. Case study: Electronic medical record data

To check the adaptability of the MARISMA-BiDa pattern in any given Big Data environment, we have used a typical Big Data scenario from the heathcare sector (based on a case study of the NIST) showing how defined elements for this scenario can easily be adapted to any other Big Data scenario using the same pattern. The case study selected is about electronic medical record data whose goal is "*Use advanced methods for normalizing patient, provider, facility and clinical concept identification within and among separate health care organizations to enhance models for defining and extracting clinical phenotypes from non-standard discrete and free-text clinical data using feature selection, information retrieval and machine learning decision-models. Leverage clinical phenotype data to support cohort selection, clinical outcomes research, and clinical decision support.*" (Case study 16 of (NIST, 2018a)).

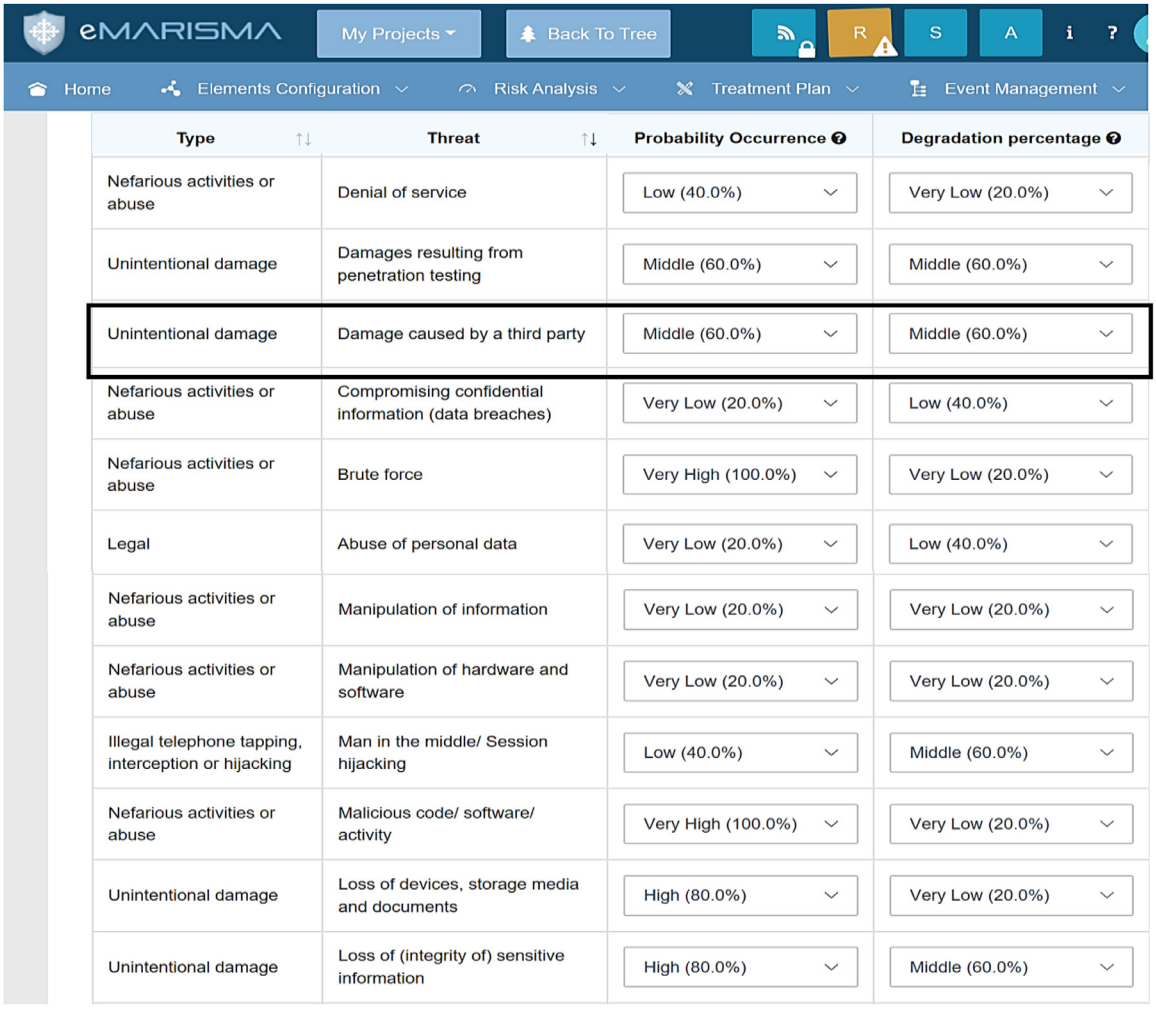
This Big Data system will be used for analysis and decision making in a medical environment as well as protecting, managing, and storing sensitive medical data that can be used in analytical processing for disease detection and decision making.

To perform a risk analysis of this system, the pattern defined for Big Data presented in the previous section (MARISMA-BiDa) will be used, and the eMARISMA tool will be applied to perform such risk analysis of the system. To do this, the first thing to identify is the set of assets that are part of the system (from of the assets defined in the pattern). Secondly, identify which dimensions will be affected. After that, the next step is to analyze the probabilities of occurrence of possible threats (from those defined in the pattern) that may occur in this system, as well as calculate the percentage of degradation that the materialization that the threat may have in the assets associated with those threats (association defined in the matrix of Table 2 of the pattern). This is part of the risk identification and analysis tasks of the MARISMA process shown in Fig. 3. All these elements will be defined and analyzed in the following subsections.

5.1. *Dimensions for case study*

While analyzing the case study, the first thing that must be established is the set of dimensions that influence the requirements and needs of the system to be studied. These dimensions are defined in the pattern (as shown in Fig. 10), and for the case study, the six dimensions are used for our system. They are described subsequently:

- Volume: more than 12 million patients, more than 4 billion discrete clinical observations and > 20 TB raw data.
- Variety: a broad variety of clinical datasets from multiple sources: free text provider notes; laboratory and emergency department encounters; chemistry, cardiology or hematology studies; blood bank and toxicology studies, etc.



Type	Threat	Probability Occurrence	Degradation percentage
Nefarious activities or abuse	Denial of service	Low (40.0%)	Very Low (20.0%)
Unintentional damage	Damages resulting from penetration testing	Middle (60.0%)	Middle (60.0%)
Unintentional damage	Damage caused by a third party	Middle (60.0%)	Middle (60.0%)
Nefarious activities or abuse	Compromising confidential information (data breaches)	Very Low (20.0%)	Low (40.0%)
Nefarious activities or abuse	Brute force	Very High (100.0%)	Very Low (20.0%)
Legal	Abuse of personal data	Very Low (20.0%)	Low (40.0%)
Nefarious activities or abuse	Manipulation of information	Very Low (20.0%)	Very Low (20.0%)
Nefarious activities or abuse	Manipulation of hardware and software	Very Low (20.0%)	Very Low (20.0%)
Illegal telephone tapping, interception or hijacking	Man in the middle/ Session hijacking	Low (40.0%)	Middle (60.0%)
Nefarious activities or abuse	Malicious code/ software/ activity	Very High (100.0%)	Very Low (20.0%)
Unintentional damage	Loss of devices, storage media and documents	High (80.0%)	Very Low (20.0%)
Unintentional damage	Loss of (integrity of) sensitive information	High (80.0%)	Middle (60.0%)

Fig. 13 – Threats with the occurrence probability and degradation percentage on the eMARISMA tool for our case study.

- Velocity: between 500,000 and 1.5 million new real-time clinical transactions added per day.
- Variability: data from clinical systems evolve over time because the clinical and biological concept space is constantly evolving New scientific discoveries lead to new disease entities, new diagnostic modalities, and new disease management approaches.
- Veracity: data from each clinical source are commonly gathered using different methods and representations, yielding substantial heterogeneity. This leads to systematic errors and bias requiring robust methods for creating semantic interoperability.
- Value: Information retrieval methods to identify relevant clinical features, i.e., decision models used to identify a variety of clinical phenotypes such as diabetes, congestive heart failure, and pancreatic cancer.

5.2. *Assets for case study*

The MARISMA-BiDa pattern defines a broad set of Big Data oriented asset types that can be used to define the specific assets in our case study. In Table 4, in the first two columns the family of assets, and the type to which they belong can be seen, defined by the pattern. In the third column, once instantiated the pattern, the specific assets that have been obtained by analyzing and studying in depth the system, the resources and the context where the system is deployed can be observed.

The eMARISMA tool helps us define each of these assets by defining the type it belongs to. Once all the assets of the case study have been defined, the list of assets to protect from our system is obtained. The tool preloads the most typical assets used by companies in that sector, with the aim of thereby helping the user in the selection. Fig. 12 shows some of the assets added by the tool, in particular, the assets of the family '*Data*' and the assets of the family '*Individuals and roles*'. The tool also adds a field to indicate the value that the asset has for the company or organization, which will allow prioritization and accuracy improvement in the level of risk when taking into account the impact and degradation of assets due to threats.

Once the assets for our case study have been identified and added to the eMARISMA tool, it automatically performs the risk evaluation using all the relationships and matrices defined in our MARISMA-BiDa pattern. The tool then relates the asset defined in our case study with the threats and dimensions affected. It also shows the result of the current risk for

Table 3 – Matrix "Type of Asset - Threats - Dimensions" in Environments of Big Data.

FAMILIY OF ASSETS	Big Data Analytics						Data						Security privacy techniques						Infrastructure						Individuals and roles					
THREATS / DIMENSIONS	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Unintentional damage																														
Information leakage						■					■	■											■	■						
Leaks of data via Web app					■						■									■			■							
Loss of devices, documents						■						■												■						
Loss of sensitive information					■						■												■						■	
Loss of information in the cloud					■						■												■							
Damages from penetration testing						■						■												■						
Damage caused by a third party					■						■												■							
Inadequate design			■	■		■		■	■	■		■								■	■	■								
Unintentional change of data				■						■												■							■	
Erroneous use or administration	■		■				■		■											■		■								
Using info. from unreliable source					■						■												■							
Destruction of records	■	■					■	■												■	■									
Eavesdropping																														
Network Reconnaissance									■																					
Intercepting compromising emissions											■																			
War driving											■																			
Interception of information											■	■											■						■	
Interfering radiation											■																			
Replay of messages									■																					
Man-in-the-middleg											■	■																		
Nefarious																														
Abuse of Information	■					■	■					■	■					■	■					■						
Social Engineering					■						■						■						■							■
Malicious code/ software/ activity											■	■					■	■			■		■	■						
Abuse of authorizations					■						■						■	■					■						■	
Brute force			■						■						■						■									
Failed of business process					■	■					■	■																		
Denial of service			■						■			■			■			■			■			■						
Targeted attacks (APTs etc.)						■						■						■						■						
Receive of unsolicited E-mail					■						■						■						■						■	
Remote activity (execution)			■						■						■						■									
Identity theft					■						■	■					■						■						■	
Hoax					■						■						■						■							
Compromising confidential info.					■						■						■						■						■	
Generation and use rogue certificates					■	■					■	■					■						■							
Manipulation of hw&sw	■		■				■		■				■		■				■		■									
Manipulation of information						■						■						■						■					■	■
Misuse of information						■						■						■						■						■
Unauthorized activities					■						■						■						■					■		
Unauthorized installation of software						■						■						■						■						
Misuse of audit tools						■						■						■				■		■						
Legal																														
Violation of laws or regulations											■	■																	■	
Failure meet contractual requirem.								■		■																			■	
Unauthorized use of IPR							■				■	■																		
Abuse of personal data											■	■																	■	
Judiciary decisions/ court orders											■	■																		■
Organizational																														
Skill shortage																											■		■	■

DIMENSIONS: 1: Volume; 2: Variety; 3: Velocity; 4: Variability; 5: Veracity; 6: Value

Table 4 – Assets involved in the case use for the types of assets defined in the MARISMA-BiDa pattern.

MARISMA-BiDa PATTERN Family of Asset	Type of Asset	PATTERN INSTANTIATED Assets for the Case Study
Infrastructure	Software	Operating Systems, Server Software
	Hardware (physical/virtual)	Servers, Network, Media and storage devices
	Computing Infrastructure models	Batch
	Storage Infrastructure models	Database management systems (Teradata, PostgreSQL, MongoDB)
Data	Structured data	Identification record data, Databases
	Semi-structured and Unstructured data	Files and documents, Multimedia
Big Data Analytics	Data analytics algorithms and procedures	Metrics definitions, Models definitions, Data preparation procedures
	Analytical results	Graphic results & Visualizations
Security and privacy techniques	Infrastructure Security	Security policies
	Security of Data Management	Security of Data Storage and Logs
	Integrity and reactive security	End Point validation and filtering
	Data Privacy	Privacy for Data mining and analytics, Access Control
Individuals and roles	Data provider	Healthcare providers (physicians, nurses, public health officials)
	Data consumer	Biomedical informatics research scientists, Health services researchers

this set of assets and links the objectives and controls necessary to implement or acquire to protect this set of assets.

5.3. *Types of threats for case study*

Taking the set of threats defined in the MARISMA-BiDa pattern, for the case study we must analyze and identify which of these threats will impact the system taking into account the assets to protect. To do this, two percentage values must be assigned in two parameters for each of the threats that may influence the system to analyze in the tool. These parameters are: 1) the probability of occurrence of the threat, i.e. with what probability an attack can occur on a specific asset. 2) the degradation of the asset value, i.e. the damage caused by the threat to the asset by reducing the value of the asset. These parameters are measured in percentages ranging from [0..100] taken in 10 by 10 (see Fig. 13).

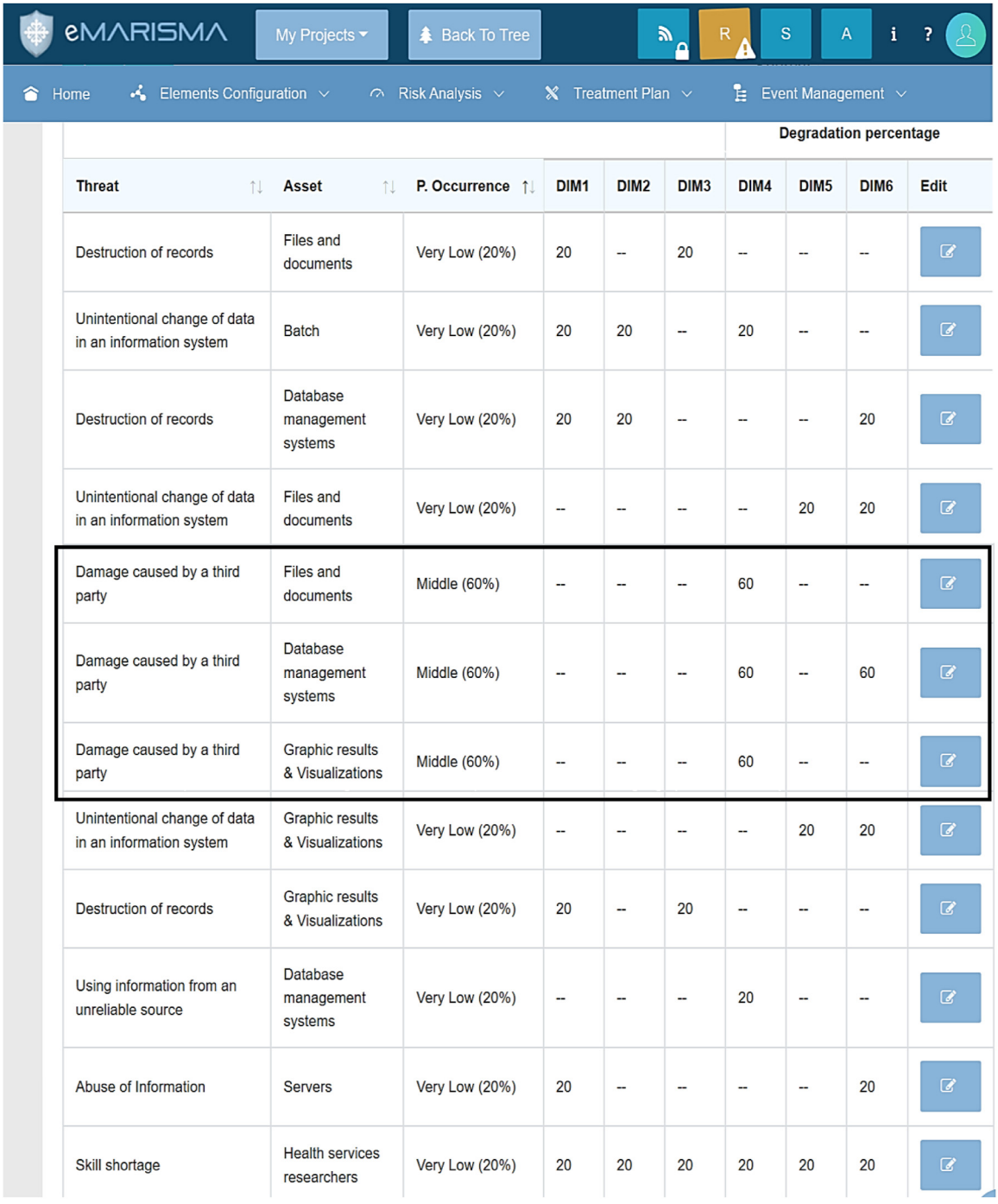
To facilitate the analysis and identification of the set of threats defined in the MARISMA-BiDa pattern, the tool preloads the list of threats with the most likely occurrence for that type of company, based on their experience and previous learning. It also loads the most probable value for the percentage of degradation, which, although it requires the pair (asset x threat), can be pre-configured at this stage as a shortcut to reduce the configuration costs of the subsequent stage. The eMARISMA tool allows us to change the values by default (for the occurrence probability and degradation percentage) according to our criteria, experience, or the environment of the system to be evaluated, as it may change over time since security events occur, as mentioned in Section 3. The risk assessment can be recalculated as security events take place, such as a specific attack on the system. As the tool is used by security expert users, they have a clear idea of what types of threats are most common in its environment and how it can degrade assets if the threat is successful. That way, they can freely change those values to others more in line with the context where the system is operating at their expert's discretion. If a threat is not likely to materialize in our system, we must indicate it with a value of 0% in each one of these two parameters.

For example, as Fig. 13 shows, the threat "*damage caused by third party*" can update the occurrence probability and the degradation percentage up to 60%. This percentage is obtained from the knowledge of the context in which the system is located. In this case, it is of utmost importance to consider that this is a context where many people are involved and want to have access to this type of sensitive data, or even external companies that are in charge of managing and securing this type of sensitive data. This can cause unintentional damage due to a misuse of such data by not following any security policy, or a misconfiguration in the database that stores them, or simply due to the accidental deletion of such data that would cause a significant damage with regards to medical data, including a violation of the law, since this type of sensitive data is not well protected. Therefore, the expert user can consider that the percentage of degradation is high (60%) and that the probability of occurring is greater than expected, for example, 60% as well. These values are changed in the tool because they are considered appropriate for the environment, and the risk level is evaluated with those introduced values, considering all the elements defined for our pattern (assets, dimensions, controls, etc.).

The security expert user can change or update the values when necessary due to the particularities of the scenario considered and for the extensive knowledge of the system and the environment.

5.4. *Relationships defined by the pattern*

In the MARISMA-BiDa pattern, the matrix that relates threats to the dimensions and the assets affected has been defined (see Table 3). In this stage, the pairs [*assets x threats*] gener-



			Degradation percentage						
Threat	Asset	P. Occurrence	DIM1	DIM2	DIM3	DIM4	DIM5	DIM6	Edit
Destruction of records	Files and documents	Very Low (20%)	20	--	20	--	--	--	
Unintentional change of data in an information system	Batch	Very Low (20%)	20	20	--	20	--	--	
Destruction of records	Database management systems	Very Low (20%)	20	20	--	--	--	20	
Unintentional change of data in an information system	Files and documents	Very Low (20%)	--	--	--	--	20	20	
Damage caused by a third party	Files and documents	Middle (60%)	--	--	--	60	--	--	
Damage caused by a third party	Database management systems	Middle (60%)	--	--	--	60	--	60	
Damage caused by a third party	Graphic results & Visualizations	Middle (60%)	--	--	--	60	--	--	
Unintentional change of data in an information system	Graphic results & Visualizations	Very Low (20%)	--	--	--	--	20	20	
Destruction of records	Graphic results & Visualizations	Very Low (20%)	20	--	20	--	--	--	
Using information from an unreliable source	Database management systems	Very Low (20%)	--	--	--	20	--	--	
Abuse of Information	Servers	Very Low (20%)	20	--	--	--	--	20	
Skill shortage	Health services researchers	Very Low (20%)	20	20	20	20	20	20	

Fig. 14 – Degradation percentage for the dimensions affected by assets and threats on the eMARISMA tool.

ated from the pattern configuration are loaded, and the most probable values established in the previous stage are loaded (by default) for the degradation percentages of the triad [*asset x threat x dimension*] defined by the matrix.

Once the most appropriate values for the threats have been assigned (by default or not), the eMARISMA tool applies the matrix '*Type of Assets-Threats-Dimensions*' to calculate the degradation percentage but, this time, including the affected dimensions for each threat and asset type defined in the MARISMA-BiDa pattern that can be seen in Table 3. As previously mentioned, these values can be modified for each dimension according to our criteria, experience and knowledge of the environment, although based on the current historical data of eMARISMA less than 15% of the values are actually modified. Fig. 14 depicts the matrix with the values on the eMARISMA tool.

In line with the previous example, for the threat '*damage caused by third party*' with a probability of occurrence of 60%, it can be appreciated that this threat affects a set of assets whose degradation depends on the affected dimension. Thus, this threat affects the assets '*Files and Documents*', '*Graphic results & Visualizations*', and '*Database management systems*' out

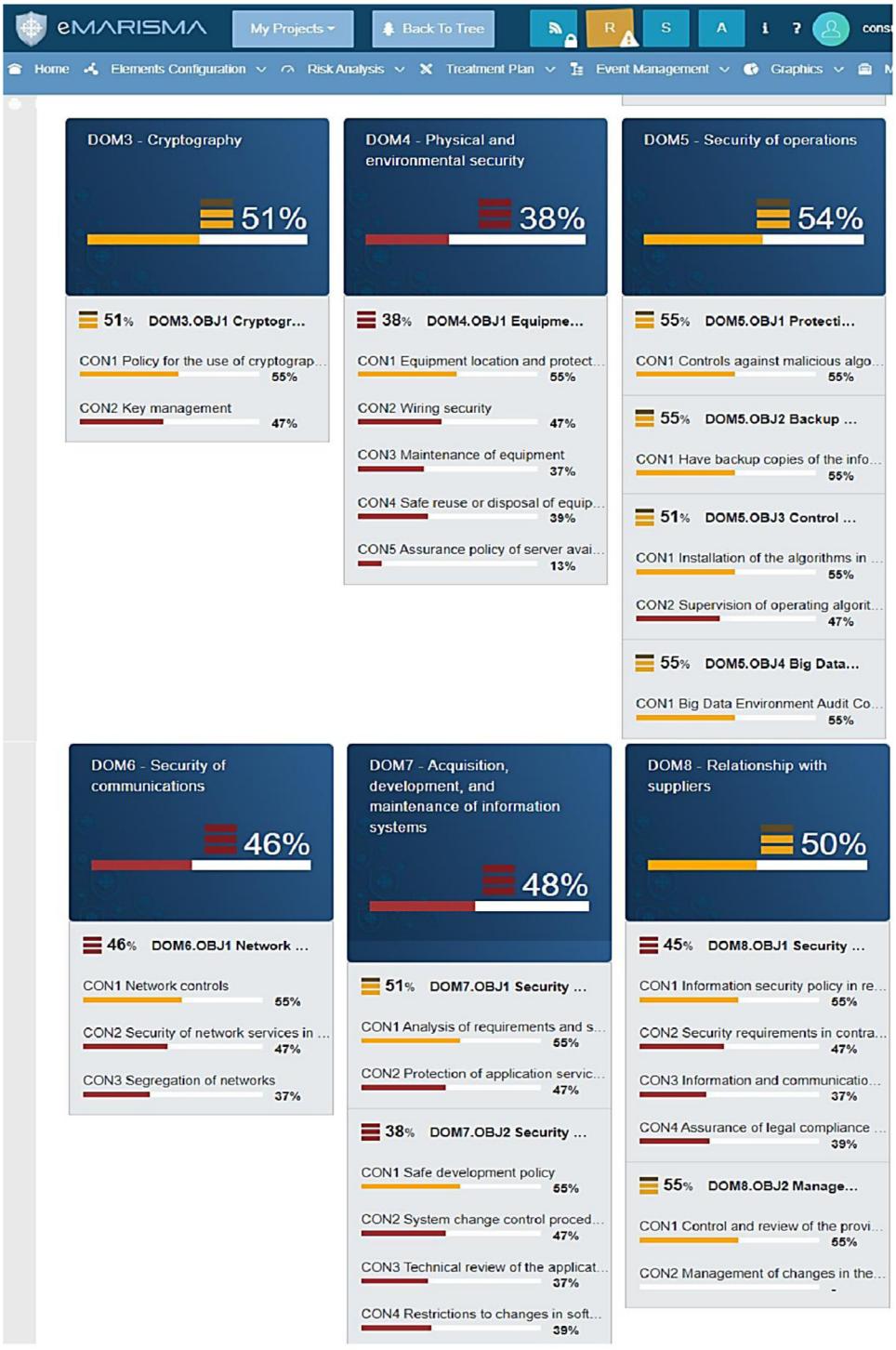


Fig. 15 – Dashboard generated by the eMARISMA tool for our case study.

of three different asset types (Data, Big Data Analytics and Infrastructure respectively), whose degradation of these assets is influenced by the '*Veracity*' dimension (DIM4). In addition, for the asset '*Database management systems*', another dimension that influences its degradation is '*Value*' (DIM6). This dependence is established in the MARISMA-BiDa pattern and it can be seen in Table 3. The initial values for the affected dimensions are the same as the value assigned for the degradation of this threat. The security expert user can change this percentage for each dimension, even for those that are not initially related should the user considers it necessary to add it for the specific particularities of the system to be analyzed.

5.5. *Results of risk analysis with eMARISMA*

Once the specific elements of the case study have been identified and added, based on the MARISMA-BiDa pattern, the eMARISMA tool automatically performs the risk analysis, ob-

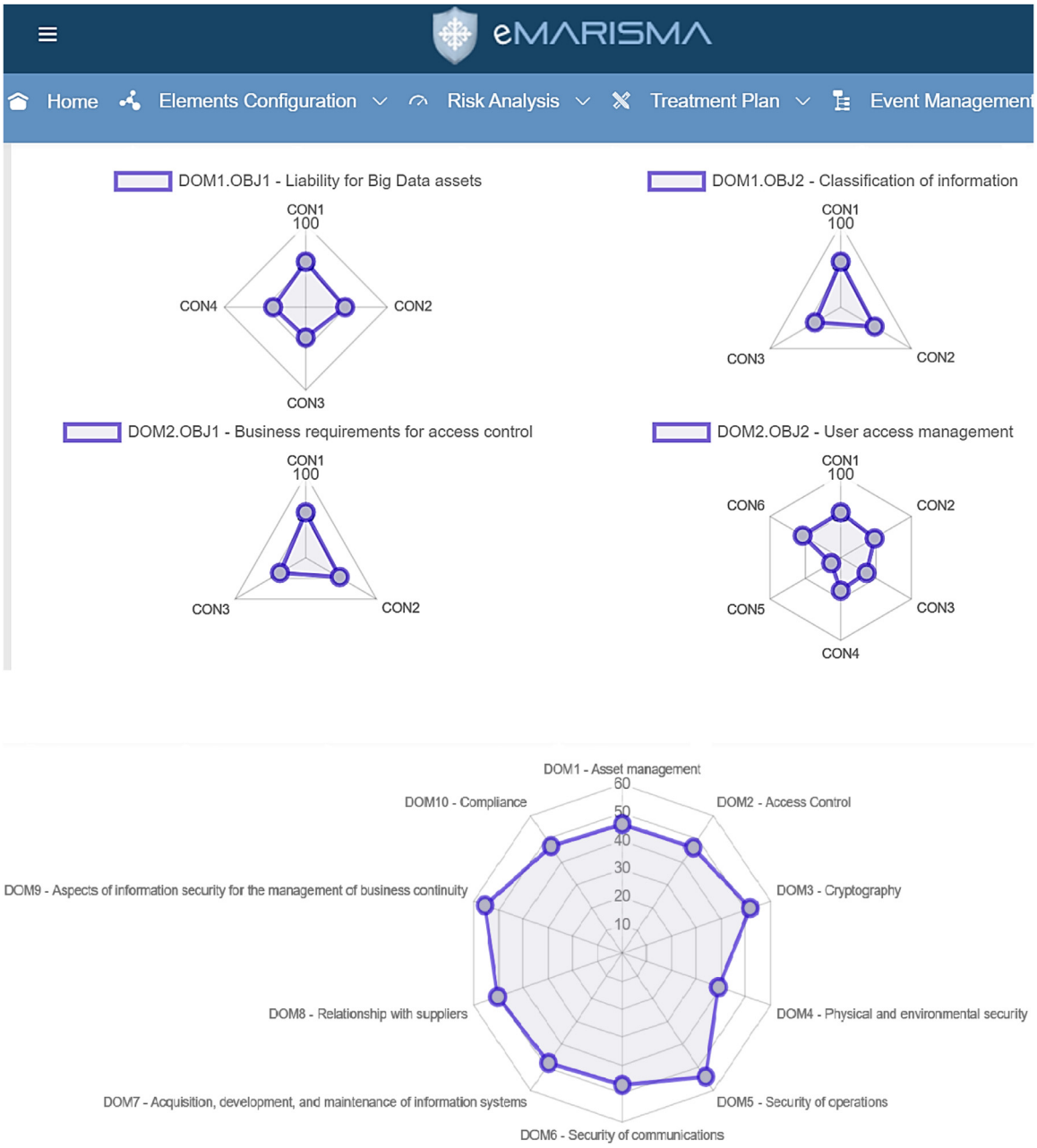


Fig. 16 – Kiviat diagrams generated by the eMARISMA tool for the objectives and audit achieved on our case study.

taining a system risk level and a level of coverage of domains and objectives. To do this, and once the assets to be protected from the case study are added in the tool, an initial audit of the system within the organization must be conducted taking into account the domains, control objectives, and controls defined by the MARISMA-BiDa pattern. This audit is performed as a checklist, and also supported by the tool. The objective of this initial audit is to know the current state of security of the system, knowing the set of controls that are applicable to protect the assets, which of those applicable controls are implemented in the system, and also this audit indicates the set of controls that are not applicable to the context in which the system is deployed. All things considered, the tool gives a real level of risk and control coverage of the system within the organization, once the risk analysis process is completed (considering the rest of the elements of the pattern, dimensions, threats, relationships, probability and degradation).

Fig. 15 shows a dashboard to monitor the risk of the organization in real time for each domain—defined for the pattern and once the initial audit has been conducted. This figure shows the level of compliance for each of the domains, control objectives and controls that make up the pattern on which the risk analysis is being executed. This dashboard is organized by domains, and within each domain, the objectives taking part can be seen, as well as for each control objective, the set of controls that are applicable to the health organization can be observed. This dashboard allows navigation until reaching the control level, where we can see the evolution that the controls

have followed over time. For all these elements, the percentage of actual coverage is shown.

Thus, for example, for domain 6 '*Security of operations*' there is coverage of 56%, which coincides with the only control objective defined for this domain, which is '*Network security management*'. This level of coverage comes from the average of the control coverage in the system for the three possible controls available for this control objective which are: '*Network controls*' with coverage of 55%, '*Security of network services in Big Data*' with coverage of 47% and '*Segregation of networks*' with coverage of 37%.

Fig. 16 also represents the level of coverage with kiviat diagrams and risks maps for the security levels that the health organization has at every moment. The figure shows an individualized and global vision of the level of compliance with the controls for each of the domains that make up the pattern. The tool shows the coverage that the controls have from three points of view: from the point of view of the general audit (bottom of Fig. 16), from the point of view of each control objective (top of Fig. 16), and from the point of view of each domain. Hence, it allows the upper management of the company to make decisions depending on the results shown and their risk appetite.

6. Conclusions and future work

A risk analysis and management process can be divided into several stages including setting the context, identifying risks, analyzing risks in terms of probability and impact, assessing risks, and finally risks treatment.

This paper shows how the MARISMA methodology is used (supported by the eMARISMA tool), to generate a security management and analysis pattern focused on Big Data aspects, which allows dynamic management of the risk associated with elements of a Big Data environment in a company. MARISMA defines a meta-pattern, which is a generic structure with the main elements of a risk analysis and management process. It facilitates its reuse and its facility to create sectorial patterns, which depend on the context where the system runs. This meta-pattern allows to incorporate security standards or frameworks for the risk analysis of specific sectors or contexts.

The use of the eMARISMA tool helps to automate the process, define context-specific patterns, and perform automatic risk assessment and assists in making risk decisions. Its usability allows it to be used by both large companies and SMEs. In fact, it is being applied to a variety of clients in different sectors, creating various patterns adapted to the context, such as critical infrastructure or the maritime sector.

In this paper, MARISMA has been applied in a case study, whose application allowed to refine it and improve it with that experience. These refinements have been mainly focused on adjusting the main concepts of the MARISMA-BiDa pattern, reaffirming the most relevant concepts of those already identified, and finding some others based on experience.

As future work, the evolution of the eMARISMA tool is contemplated to be a learning system in the cloud. Everything related to the management of security incidents will also be incorporated, in a way that in the event of an attack, it is incorporated as an incident. It is also automatically associated with the controls and assets to be protected and it recommends making an effort to improve a set of specific controls. It will allow to incorporate the security incidents that affect one of the systems, in all those systems elated or that may be affected through the pattern.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been funded by the ECLIPSE project (Ministerio de Ciencia, Innovación y Universidades, y Fondo Europeo de Desarrollo Regional FEDER, RTI2018–094283-B-C31), the GENESIS project (Consejería de Educación, Cultura y Deportes de la Junta de Comunidades de Castilla La Mancha, y Fondo Europeo de Desarrollo Regional FEDER, SBPLY/17/180501/000202). We thank the support of the companies Sicaman Nuevas Tecnologías S.L. (www.sicaman-nt.com) and Marisma Shield S.L. (www.emarisma.com) that have facilitated the validation of case studies and the use of the eMARISMA tool.

REFERENCES

Abbass W, Baina A, Bellafkih M. Using EBIOS for risk management in critical information infrastructure. In: 2015 5th World Congress on Information and Communication Technologies (WICT). IEEE; 2015. p. 107–12.

Acevedo N, Satizábal C. Risk management and prevention methodologies: a comparison. Sistemas y Telemática 2016;14:39–58.

Agrawal V. A Comparative Study on Information Security Risk Analysis Methods. J. Comput. (Taipei) 2017;12:57–67.

Akinrolabu O, Nurse JRC, Martin AP, New S. Cyber risk assessment in cloud provider environments: current models and future needs. Comput. Secur. 2019;87.

Akoka J, Comyn-Wattiau I, Laoufi N. Research on Big Data – a systematic mapping study. Comput. Standards Interfaces 2017;54:105–15.

Al-ahmad W, Mohammad B. Addressing information security risks by adopting standards. Int. J. Inf. Secur. Sci. 2013;2:28–43.

Alberts C, Dorofee A, Stevens J, Woody C. Introduction to the OCTAVE Approach. Carnegie-Mellon Univ Pittsburgh Pa Software Engineering Inst; 2003.

Ale B. Risk analysis and big data. Safety Reliab. 2016;36:153–65.

Armstrong K. Big data: a revolution that will transform how we live, work, and think. information. Commun. Soc. 2014;17(10):1300–2.

Aviad A, Wecel K, Abramowicz W. Semantic Risk Assessment for Cybersecurity. International Conference on Cyber Warfare and Security: Academic Conferences International Limited 2018 513-X.

Axelos. ITIL | IT Service Management | ITSM | AXELOS. 2019.

Barrientos AM, Areiza KA. Integration of a Safety Management System Withan Information Quality Management System. Universidad EAFIT; 2005.

Benjelloun F-Z, Lahcen AA. Big Data Security. Web Services 2018;2012:25–38.
Bergvall J, Svensson L. In: Risk Analysis Review. Linköping, Sweden: Linköpings Universitet; 2015. p. 1–36.
CCTA B. CRAMM. CCTA risk analysis and management method). Insight Consult. 2003.
Chen F. In: An Investigation and Evaluation of Risk Assessment Methods in Information systems. Goteborg: Chalmers University of Technology; 2015. p. 1–83.
Chen M, Mao S, Liu Y. Big Data: a survey. Mobile Netw. Appl. 2014;19(2):171–209.
Choi T-M, Lambert JH. Advances in risk analysis with Big Data. Risk Anal. 2017;37:1435–42.
Damiani E. Toward big data risk analysis. In: IEEE International Conference on Big Data (Big Data). IEEE; 2015. p. 1905–9.
Disterer G. ISO/IEC 27000, 27001 and 27002 for information security management. J. Inf. Secur. 2013;04:92–100.
Dong X, Li R, He H, Zhou W, Xue Z, Wu H. Secure sensitive data sharing on a big data platform. Tsinghua Sci. Technol. 2015;20(1):72–80.
Fenz S, Ekelhart A. Verification, validation, and evaluation in information security risk management. IEEE Secur. Privacy Mag. 2011;9:58–65.
Fredriksen R, Kristiansen M, Gran BA, Stølen K, Opperud TA, Dimitrakos T. The CORAS framework for a model-based risk management process. LNCS 2002;2434:94–105.
Garcia FYH, Moreta LML. Maturity model for the risk analysis of information assets based on methodologies MAGERIT, OCTAVE y MEHARI; focused on shipping companies.. In: 2018 7th International Conference On Software Process Improvement (CIMPS): IEEE; 2018. p. 29–39.
Haiwen S, Xiaofang X. Threat evaluation method of warships formation air defense based on AR(p)-DITOPSIS#br#. J. Syst. Eng. Electron. 2019;30(2):297–307.
Hashem IAT, Yaqoob I, Anuar NB, Mokhtar S, Gani A, Ullah Khan S. The rise of "big data" on cloud computing: review and open research issues. Inf. Syst. 2015;47:98–115.
Hashim NA, Abidin ZZ, Zakaria NA, Ahmad R, Puvanasvaran A. Risk assessment method for insider threats in cyber security: a review. Risk 2018;9(11).
He C, Lu K. Risk management in SMEs with financial and non-financial indicators using business intelligence methods. Management 2018;16:18.
ISACA. COBIT | Control Objectives for Information Technologies | ISACA. ISACA 2019.
. In: Information Technology - Security techniques - Systems Security Engineering - Capability Maturity Model® (SSE-CMM®).. International Organization for Standardization ISO: ISO/IEC; 2008. p. 132.
ISO/IEC 27005. Information technology - security techniques - Information security risk management. Inf. Secur. Risk Manage. 2018:80.
ISO/IEC TR 15443-1. Information technology – Security techniques – a framework for IT security assurance – Part 1: Overview and framework 2012.
Kelemen R, Biskup M, Redep NB. The conceptual risk management model — a case study of Varazdin County. In: 2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). IEEE; 2016. p. 1539–45.
Khan MAU, Uddin MF, Gupta N. Seven V's of Big Data understanding Big Data to extract value. In: Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education. IEEE; 2014. p. 1–5.
Korman M, Sommestad T, Hallberg J, Bengtsson J, Ekstedt M. Overview of enterprise information needs in information security risk assessment. In: 2014 IEEE 18th International Enterprise Distributed Object Computing Conference. IEEE; 2014. p. 42–51.
Kouns J, Minoli D. Information Technology Risk Management in Enterprise environments: A review of Industry Practices and a Practical Guide to Risk Management Teams. Wiley Online Library; 2010.
Lund MS, Solhaug B, Stølen K. The CORAS Method. SourceForge; 2003.
Macedo F, Neto F. Models For Assessing Information Security Risk. Instituto Superior Técnico da Universidade Técnica de Lisboa; 2009.
MEHARI. MEHARI (MEthod for Harmonized Analysis of RIsk) 2010.
Microsoft. Microsoft Big Data Solution Brief 2014.
Moreno J, Fernandez EB, Serrano MA, Fernández-Medina E. Secure development of big data ecosystems. IEEE Access 2019;7:96604–19.
Moreno J, Serrano MA, Fernandez-Medina E, Fernandez EB. In: 20th International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data (DOLAP). Towards a security reference architecture for big data; 2018.
Mukama J. Risk Analysis as a Security Metric For Industrial Control Systems. Master's thesis in Computer Systems and Networks. Chalmers University of Technology; 2016.
Murthy P, Bharadwaj A, Subrahmanyam P, Roy A, Rajan S. In: Big Data Taxonomy. Cloud Security Alliance: Cloud Security Alliance; 2014. p. 33 September.
NATO. Improving Common Security Risk Analysis. Research and Technology Organisation of NATO; 2008. RTO Technical Report TR-IST-049.
. In: NIST Big Data Interoperability Framework, 6. Gaithersburg, MD: Reference Architecture. NIST Special Publication 1500-6r1; 2015. p. 62.
NIST. NIST Big Data Interoperability Framework: volume 3, use cases and general requirements. NIST Spec. Publ. 2018a:1500–13r1.
NIST. Special Publication 800-37 Risk management framework for information systems and organizations a system life cycle approach for security and privacy 2018b.
Nurse JRC, Creese S, De Roure D. Security risk assessment in internet of things systems. IT Prof. 2017;19:20–6.
Oppliger R, Pernul G, Katsikas S. New Frontiers: assessing and managing security risks. Computer (Long Beach Calif) 2017;50:48–51.
Pan L. In: Application of a Financial Quantitative Risk Model to Information Security Risk Assessment. School of Mathematics and Information Security Royal Holloway. University of London; 2018. p. 14–157.
Pan L, Tomlinson A. A systematic review of information security risk assessment. Int. J. Safety Secur. Eng. 2016;6:270–281.
Pandey SK. A comparative study of risk assessment methodologies for information systems. Bull. Electr. Eng. Informatics 2012;1(2):111–22.
Paryasto M, Alamsyah A, Rahardjo B, Kuspriyanto. Big-data security management issues. In: 2014 2nd International Conference on Information and Communication Technology (ICoICT). IEEE; 2014. p. 59–63.
Patgiri R. Taxonomy of Big Data: A Survey. Distributed, Parallel, and Cluster Computing 2018.
Patgiri R, Ahmed A. Big Data: the V's of the Game Changer Paradigm. In: 2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS). IEEE; 2016. p. 17–24.
Petrescu AG, Postole MA, Ciobanasu M. In: The International Experience in Security Risk Analysis Methods. Network Security and Its Impact On Business Strategy. IGI Global; 2019. p. 157–69.

Priya S, Navdeti C. Securing Big Data Hadoop: a review of security issues, threats and solution. Int. J. Comput. Sci. Inf. Technol. 2015;5(2):1 -.
Radanliev P, De Roure D, Cannady S, Montalvo RM, Nicolescu R, Huth M. Economic impact of IoT cyber risk - analysing past and present to predict the future developments in IoT risk analysis and IoT cyber insurance. In: Living in the Internet of Things: Cybersecurity of the IoT - 2018. Institution of Engineering and Technology;; 2018. p. 1–9.
Rajbhandari L. Consideration of opportunity and human factor: required paradigm shift for information security risk management. In: 2013 European Intelligence and Security Informatics Conference: IEEE; 2013. p. 147–50.
Reinsel D, Gantz J, Rydning J. Data Age 2025: don 't focus on Big Data; focus on the data that's Big Data age 2025. IDC White Paper; Sponsored by Seagate 2017:1–25.
Rekleitis E. Big Data Threat Landscape and Good Practice Guide. European Union Agency For Network And Information Security; 2016.
Rossebø JE, Wolthuis R, Fransen F, Björkman G, Medeiros N. An enhanced risk-assessment methodology for smart grids. Computer (Long Beach Calif) 2017;50(4):62–71.
Rot A. Enterprise information technology security: risk management perspective. In: Proceedings of the World Congress on Engineering and Computer Science; 2009. p. 20–2.
Sagiroglu S, Sinanc D. Big data: a review. In: International Conference on Collaboration Technologies and Systems (CTS): IEEE; 2013. p. 42–7.
Sánchez LE, Ruiz C, Fernández-Medina E, Piattini M. Managing the asset risk of SMEs. In: 2010 International Conference on Availability, Reliability and Security. IEEE; 2010. p. 422–429.
Santos-Olmo A, Sánchez L, Rosado D, Fernández-Medina E, Piattini M. Applying the action-research method to develop a methodology to reduce the installation and maintenance times of information security management systems. Future Internet 2016;8:36.
Shamala P, Ahmad R, Yusoff M. A conceptual framework of info structure for information security risk assessment (ISRA). J. Inf. Secur. Appl. 2013;18:45–52.
Shameli-Sendi A, Aghababaei-Barzegar R, Cheriet M. Taxonomy of information security risk assessment (ISRA). Comput. Secur. 2016;57:14–30.
Shukla N, Kumar S. A comparative study on information security risk analysis practices. IJCA Special Issue on Issues and Challenges in Networking. Intell. Comput. Technol. ICNICT 2012(3):28–33.
Spanish Higher Council for Government. PAe - MAGERIT v.3: Methodology of analysis and risk management information systems 2012.
Stergiopoulos G, Gritzalis D, Kouktzoglou V. Using formal distributions for threat likelihood estimation in cloud-enabled IT risk assessment. Comput. Netw. 2018;134:23–45.
Sun Z, Strang K, Li R. Big Data with Ten Big Characteristics 2019: 56–61.
Syalim A, Hori Y, Sakurai K. Comparison of risk analysis methods: mehari, Magerit, NIST800-30 and Microsoft's Security Management Guide. In: 2009 International Conference on Availability, Reliability and Security: IEEE; 2009. p. 726–731.
Tatiana O, Mikhail U, Arina N, Natalia S. Model of enterprise's information security management. IV International research conference "Information technologies in Science, Management, Social sphere and Medicine" (ITSMSSM 2017). Atlantis Press; 2017.
Tubío Figueira P, López Bravo C, Rivas López JL. Improving information security risk analysis by including threat-occurrence predictive models. Comput. Secur. 2020;88.
Wang H, Jiang X, Kambourakis G. Special issue on security, privacy and trust in network-based big data. Inf. Sci. (Ny) 2015;318:48–50.
Wangen G. Information security risk assessment: a method comparison. Computer (Long Beach Calif) 2017;50:52–61.
Wangen G, Snekkenes E. In: Proceeding of Norwegian Information Security Conference/Norsk informasjonssikkerhetskonferanse-NISK 2013-Stavanger, 18th-20th November 2013. A taxonomy of challenges in information security risk management. Akademika Forlag; 2013.
Zambon E, Etalle S, Wieringa RJ, Hartel P. Model-based qualitative risk assessment for availability of IT infrastructures. Softw. Syst. Model. 2011;10:553–80.
Zhang D. Big Data security and privacy protection. In: 8th International Conference on Management and Computer Science (ICMCS 2018). Paris, France: Atlantis Press; 2018. p. 279–80.
Zio E. The future of risk assessment. Reliab. Eng. Syst. Safety 2018;177:176–90.
Zissis D, Lekkas D. Addressing cloud computing security issues. Future Gener. Comput. Syst. 2012;28(3):583–92.
ISO 31000, 2018. ISO 31000:2018 Risk Management - Guidelines. Geneva, Switzerland.

David G. Rosado has an MSc and PhD. in Computer Science from the University of Málaga (Spain) and from the University of Castilla-La Mancha (Spain), respectively. Associate Professor at the Escuela Superior de Informática of the Castilla-La Mancha University in Ciudad Real (Spain). His-research activities are focused on security for Information Systems and Cloud Computing. He has published several papers in national and international conferences on these subjects, and he is co-editor of a book and chapter books. Author of several manuscripts in national and international journals (Information Software Technology, System Architecture, Network and Computer Applications, etc.). He is member of Program Committee of several conferences and workshops nationals and internationals such as ICEIS, ICCGI, CISIS, SBP, IAS, SDM, SECRYPT, COSE and international journals such as Internet Research, JNCA, KNOSYS, JKSU, and so on. He is a member of the GSyA research group of the Information Systems and Technologies Department at the University of Castilla-La Mancha, in Ciudad Real, Spain. His-email is david.grosado@uclm.es.

Julio Moreno is M.Sc. and Ph.D. in Computer Science by the University of Castilla-La Mancha. His-research interests are Data Security and Privacy and Security Architectures for Big Data ecosystems. He is a member of the GSyA research group of the Information Systems and Technologies Department at the University of Castilla-La Mancha, in Ciudad Real, Spain. His-e-mail is julio.moreno@uclm.es.

Luis E. Sánchez holds a PhD in Computer Science from the University of Castilla-La Mancha (Spain), a MSc in Computer Science from the Polytechnic University of Madrid (Spain), and holds a degree in Computer Science from the University of Granada (Spain). He is Certified Information System Auditor by ISACA and Leader Auditor of ISO27001 by IRCA. He is Assistant Professor at the University of the Armed Forces of Ecuador. He participates at the GSyA research group of the Department of Information Technologies and Systems at the Castilla- La Mancha University and he is a researcher of Biological Neurocomputing and Cyberdefense within the PROMETEO project. He was Assistant Professor of the Technologies and Information Systems Department of the University of Castilla-La Mancha. He has directed more than 50 projects in multinational companies. He has more than 60 national and international papers and conference on Software Engineering and Teaching. He belongs to various professional and research associations (COIILCLM, ALI, ASIA, TUVRheinland, ISACA, eSec INTECO, SC27 AENOR …). His-email is LuisEnrique@sanchezcrespo.org.

Antonio Santos-Olmo is M.*Sc* and PhD. in Computer Science by the University of Castilla-La Mancha. He is an Assistant Professor at the Escuela Superior de Informática of the University of Castilla- La Mancha in Ciudad Real (Spain). M.*Sc* in Information Systems Audit from the Polytechnic University of Madrid, and Certified Information System Auditor by ISACA. He is the Director of Software Factory departments of the company Sicaman Nuevas Tecnologías S.L. His-research activities are management security system, security metrics, data mining, data cleaning, and business intelligence. He participates in the GSyA research group of the Department of Computer Science at the University of Castilla- La Mancha, in Ciudad Real (Spain). His-email is asolmo@sicaman-nt.com

Manuel A. Serrano is M.*Sc*. and Ph.D. in Computer Science by the University of Castilla-La Mancha. Associate Professor at the Escuela Superior de Informática of the Castilla-La Mancha University in Ciudad Real. Regarding his research interests, he is working on cyber security (specially in Big Data and IoT), data quality, software quality and measurement and business intelligence. His-scientific production is large, having published more than fifty papers in high level journals and conferences. He has participated in more than 20 research projects, have conduct several invited speeches and have work in several transfer project with companies. He has been teaching for near two decades at the University, especially in Software Engineering and Programming subjects. He has supervised several final degree theses, final master works and PhD theses. His-e-mail is manuel.serrano@uclm.es.

Eduardo Fernández-Medina holds a PhD. and an MSc. in Computer Science from the University of Castilla-La Mancha. He is a Full Professor at the Escuela Superior de Informática of the University of Castilla-La Mancha in Ciudad Real (Spain) (Computer Science Department, University of Castilla La Mancha, Ciudad Real, Spain), his research activity being in the field of security in information systems, and particularly in security in big data, Cloud Computing and cyber-physical systems. Fernández-Medina is co-editor of several books and chapter books on these subjects and has published several dozens of papers in national and international conferences (BPM, UML, ER, ESORICS, TRUSTBUS, etc.). He is author of more than fifty manuscripts in international journals (Decision Support Systems, Information Systems, ACM Sigmod Record, Information Software Technology, Computers & Security, Computer Standards and Interfaces, etc.). He leads the GSyA research group of the Department of Computer Science at the University of Castilla-La Mancha, in Ciudad Real, Spain and belongs to various professional and research associations (ATI, AEC, AENOR, etc.). His-email is eduardo.fdezmedina@uclm.es