# MARISMA-BiDa: Gestión y Control del riesgo en Big Data. Caso de Estudio

David G. Rosado, Julio Moreno, Luis E. Sánchez, Antonio Santos-Olmo, Manuel A. Serrano, Eduardo Fernández-Medina

***Abstract*— En la actualidad, se genera una gran cantidad de información debido a la amplia hiperconectividad y sensorización del mundo que nos rodea. Esta información es considerada como uno de los activos más importantes para las empresas en todos los campos. El continuo crecimiento en la importancia y el volumen de datos ha creado un nuevo problema: no puede ser manejado por las técnicas de análisis tradicionales. Este problema se resolvió, por lo tanto, mediante la creación de un nuevo paradigma: Big Data. Sin embargo, Big Data originó nuevos problemas relacionados no sólo con el volumen o la variedad de los datos, sino también con la seguridad y privacidad de los datos. Al adoptar nuevas soluciones tecnológicas como Big Data, todos los riesgos deben ser identificados y gestionados. En este artículo se presenta un caso de estudio de la aplicación de una técnica de análisis y gestión de riesgos para entornos Big Data, guiada por una metodología de gestión de la seguridad (MARISMA) y soportada por un entorno tecnológico en la nube (eMARISMA). La propuesta, denominada MARISMA-BiDa es un patrón específico para Big Data que contiene los elementos necesarios para facilitar la aplicación de la metodología de análisis y gestión de riesgos MARISMA en un entorno específico y siguiendo los principales estándares y recomendaciones internacionales relacionados con Big Data (ISO/IEC, NIST, ENISA).**

***Index Terms*—Big Data, Análisis y Gestión de riesgos, Seguridad**

## I. INTRODUCCIÓN

UN número creciente de dispositivos, sensores y personas están conectados a la red global y esto cambia drásticamente la capacidad de generar, comunicar, compartir y acceder a los datos [1]. Los datos son esenciales para el desarrollo de sus actividades cotidianas, así como para ayudar a la dirección de las empresas a alcanzar sus objetivos y a tomar las mejores decisiones a partir de la información que se extrae de ellas [2]. Los desarrollos tecnológicos y nuevas aplicaciones continúan alimentando el debate sobre lo que define a Big Data y lo distingue de las anteriores formas de análisis de datos [3]. No existe un consenso real en cuanto a sus características clave, aunque la mayoría de las definiciones de Big Data se refieren a las tres Vs [4]. La primera V es para Volumen (el uso de grandes cantidades de datos), la segunda V es para Variety (el uso de diversas fuentes de datos que se almacenan en estructuras diversas o incluso de forma no estructurada) y la tercera significa Velocidad, o la velocidad del procesamiento de datos (los datos son a menudo en tiempo real). A lo largo del tiempo, varios autores han añadieron Vs adicionales a este trío, como Veracity [5], Variabilidad [6], [7], Valor [8], [9] y Virtual [10], [11].

La aplicación de Big Data ofrece beneficios significativos para los individuos y la sociedad, pero también plantea serias preocupaciones sobre varios riesgos de seguridad de la información como la seguridad de los datos, el gobierno y la privacidad [12]. Uno de los principales problemas en el uso de los sistemas Big Data es la seguridad. Los sistemas Big Data son complejos y heterogéneos, y la seguridad de todo el sistema debe ser abordada de manera integral. Además, la integración de diferentes tecnologías introduce nuevas cuestiones de seguridad que deben abordarse adecuadamente [1]. Big Data no fue diseñado con la seguridad en mente. Con estas montañas de datos, que informan a las empresas sobre las decisiones críticas de los clientes, los hábitos y otros innumerables detalles, surge la necesidad urgente de mantener esta valiosa información segura y protegida. Al fin y al cabo, se trata de información delicada, y con gran parte de ella existe un mayor riesgo de infracciones [13].

Los problemas de seguridad y privacidad se ven magnificados por la velocidad, el volumen y la variedad de los grandes datos, como las infraestructuras de nube a gran escala, la diversidad de fuentes y formatos de datos, la naturaleza de la transmisión de la adquisición de datos y la migración entre nubes de gran volumen. El uso de infraestructuras de nube a gran escala, con una diversidad de plataformas de software, distribuidas en grandes redes de ordenadores, también aumenta la superficie de ataque de todo el sistema. Por tanto, es muy importante contar con una serie de guías, metodologías y mecanismos para implementar de forma adecuada no solo el entorno Big Data, sino también su seguridad. Pero no solo eso, además, es ampliamente considerado que todo entorno global de gestión de seguridad de la información en la empresa, debe estar centrado en los riesgos [14]–[16]. Por lo que los riesgos de seguridad en Big Data, deben ser analizados y gestionados de manera adecuada, junto a los riesgos de otros tipos de activos de información [17].

Por otro lado, la mayoría de las organizaciones de hoy en

David G. Rosado, Julio Moreno, Luis E. Sánchez y Eduardo Fernández-Medina, Grupo de Investigación GSyA, Universidad de Castilla-la Mancha, Ciudad Real, España, david.grosado@uclm.es, julio.moreno@uclm.es, luisenrique@sanchezcrespo.org, eduardo.fdezmedina@uclm.es.

Antonio Santos-Olmo, Departamento I+D+i, Sicaman Nuevas Tecnologías y Marisma Shield, Tomelloso (Ciudad Real), España, asolmo@sicaman-nt.com.

Manuel A. Serrano, Grupo de Investigación Alarcos, Universidad de Castilla-la Mancha, Ciudad Real, España, Manuel.serrano@uclm.es.

día que utilizan tecnologías de la información tienen problemas con la seguridad de su sistema de información, y diferentes investigadores destacan que la gestión del riesgo es un proceso esencial en cualquier modelo de gestión empresarial [18], y que la información es un activo valioso que se espera que esté protegido [19].

Un análisis de riesgos es un proceso sistemático para estimar la magnitud de los riesgos a los que está expuesta una organización, para saber qué decisión tomar ante una posible eventualidad [20]. Para ello, se seleccionan e implementan salvaguardas para poder conocer, prevenir, impedir, reducir o controlar los riesgos identificados. Esto es lo que se entiende como gestión de riesgos.

De forma más técnica, el análisis de riesgos permite determinar cómo es, cuánto vale y cómo de protegidos se encuentran los activos. En coordinación con los objetivos, estrategia y política de la organización, las actividades de gestión de riesgos permiten elaborar un plan de seguridad que, implantado y operado, satisfaga los objetivos propuestos con el nivel de riesgo que acepta la dirección.

Actualmente se están realizando muchas investigaciones sobre análisis de riesgos, y muchas de ellas intentan comparar los métodos clásicos para ver cómo se podrían alinear [21]–[27]. Otros investigadores han realizado también algunos análisis comparativos de los principales estándares de riesgos con el objetivo de mejorar algunos de sus aspectos [28], o trabajos que relacionan los planes de contingencia con el análisis de riesgos [29]. Durante el estudio, se identificaron ciertas deficiencias como pueden ser dificultades para su aplicación en la práctica, no cuentan con herramientas adecuadas para su procesamiento (o en caso de existir, éstas no son muy usables), están pensadas para ser aplicadas en grandes compañías, no son sensibles al contexto, sin contar con capacidades de adaptación para entornos especiales, que requieran un especial tratamiento de sus riesgos, y su falta de dinamismo y asociatividad de sus riesgos. Parte de estas deficiencias se afrontaron mediante el desarrollo de una metodología denominada MARISMA (Methodology for the Analysis of Risks on Information System, using Meta-Pattern and Adaptability) que utilizan el concepto de meta-patrón para crear una estructura capaz de soportar los elementos asociados con un análisis de riesgos y sus relaciones, con el objetivo de crear estructuras evolutivas, dinámicas y capaces de adaptarse a las nuevas tecnologías [30], [31].

Dentro de estas nuevas tecnologías ha tomado especial importancia la necesidad de poder analizar los riesgos de seguridad dentro de los entornos de Big-Data, y por tanto se ha considerado relevante, afrontar el desarrollo de un nuevo patrón soportado sobre la metodología MARISMA, que permita realizar análisis de riesgos TIC en entornos de Big-Data, validándolo en un caso de estudio relacionado con el sector sanitario, donde los datos que se manejan son muy sensibles y el nivel de seguridad que se requiere es muy alto [32], [33].

El resto de este artículo está estructurado de la siguiente manera: La sección II presenta el marco de trabajo MARISMA donde se define el modelo conceptual de gestión de riesgos y se explica la metodología de análisis y gestión de riesgos utilizada. Además, se describe el meta-patrón definido y la herramienta de soporte utilizada. La sección III se define el patrón MARISMA-BiDa usando el meta-patrón para el contexto de Big Data definiendo los elementos específicos en estos entornos. En la sección IV se presenta un caso de estudio de registros médicos utilizando el patrón MARISMA-BiDa mostrando los resultados obtenidos y su implementación en la herramienta de soporte. Finalmente, la sección V muestra las conclusiones de nuestra investigación.

## II. MARCO DE TRABAJO MARISMA

En los últimos años se han detectado una serie de deficiencias en las principales propuestas de procesos y métodos de gestión de riesgos. Entre los principales problemas identificados en los métodos de Gestión de Riesgos, hay algunos que se pueden destacar: i) Alto costo y complejidad, ya que se realiza el análisis de riesgos, ii) Su falta de orientación hacia las pequeñas y medianas empresas (PYMES); iii) Los resultados de los conocimientos de gestión de riesgos de proyectos anteriores no suelen considerarse más fáciles para ejecutar nuevos procesos de gestión de riesgos, y iv) Los análisis de riesgos son casi siempre estáticos. En consecuencia, estas cuestiones ponían en tela de juicio su eficacia y ponían en peligro su valor para las organizaciones.

Para resolver estos problemas, se desarrolló una metodología llamada MARISMA, así como una herramienta de apoyo a dicha metodología (herramienta eMARISMA[1]). Esta metodología se centra en la reducción de los costes del proceso de gestión de riesgos y en la simplicidad de su aplicación; el modelo desarrollado permite el mayor nivel posible de automatización y reutilización con la mínima cantidad de información, recogida en un tiempo muy reducido [34].

En las siguientes subsecciones se muestran los detalles sobre esta metodología y esta herramienta.

### *A. Modelo Conceptual de Gestión de Riesgos*

Las organizaciones, independientemente de su tamaño, deben ser conscientes de la importancia de los riesgos de TI y de cómo deben gestionarse. La ISO 31000 [35] define la gestión de riesgos como un proceso organizacional que debe implicar la aplicación sistemática de políticas, procedimientos y prácticas a las actividades de comunicación y consultoría, estableciendo el contexto y evaluando, tratando, monitoreando y revisando el riesgo.

Todas estas actividades deben ser controladas por los actores internos y externos del proceso de gestión de riesgos [35], [36]. La Fig. 1. resume este proceso incluyendo todas las actividades y sus relaciones.

[1] www.emarisma.com

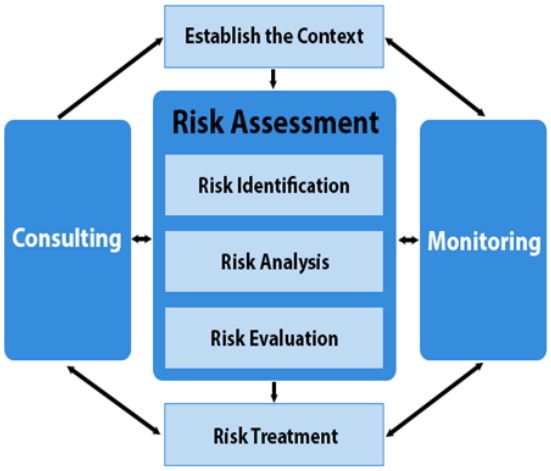


*Fig. 1 Proceso de Gestión de Riesgos basado en ISO 31000*

Siguiendo el proceso de gestión de riesgos anteriormente expuesto, y teniendo en cuenta todas las actividades, se ha considerado el desarrollo de MARISMA, ya que se trata de un marco que da soporte a todo el proceso de gestión de riesgos, centrándose principalmente en la actividad de tratamiento de riesgos. Se propone un modelo conceptual de gestión de riesgos en la Fig. 2, que representa los conceptos esenciales que deben ser capturados -en forma de información- para automatizar el proceso de análisis de riesgos a través de la herramienta eMARISMA. Este modelo conceptual sirve para establecer las relaciones entre todos los conceptos, y es posible almacenar toda la información que se genera en cualquier proceso de análisis de riesgos.

Como se muestra en la Fig. 1, el contexto de la organización es crucial para definir adecuadamente sus activos. Estos son los elementos más importantes de la gestión de riesgos y son esenciales para protegerlos. Conociendo los activos involucrados en la organización, se pueden identificar varias vulnerabilidades asociadas con ellos. Estas vulnerabilidades pueden ser mitigadas con el apoyo de controles de seguridad. Sin embargo, para abordar adecuadamente estas vulnerabilidades, es aconsejable abstraer los activos en un concepto más general, a saber, los tipos de activos. Los diferentes tipos de activos pueden estar en riesgo debido a amenazas específicas (clasificadas por categorías o tipos de amenazas), pero este riesgo no compromete a activos particulares en su totalidad, sino que lo hace en diferentes dimensiones (es decir, confidencialidad, integridad, etc.). Estas dimensiones dependen del contexto y se definen en la subsección 4.3. Es imperativo no olvidar la importancia de los requisitos de seguridad en la identificación y definición de los diferentes controles que tratan de mitigar las amenazas identificadas. El modelo definido se muestra en la Fig. 2 y se ha implementado en el marco de MARISMA.

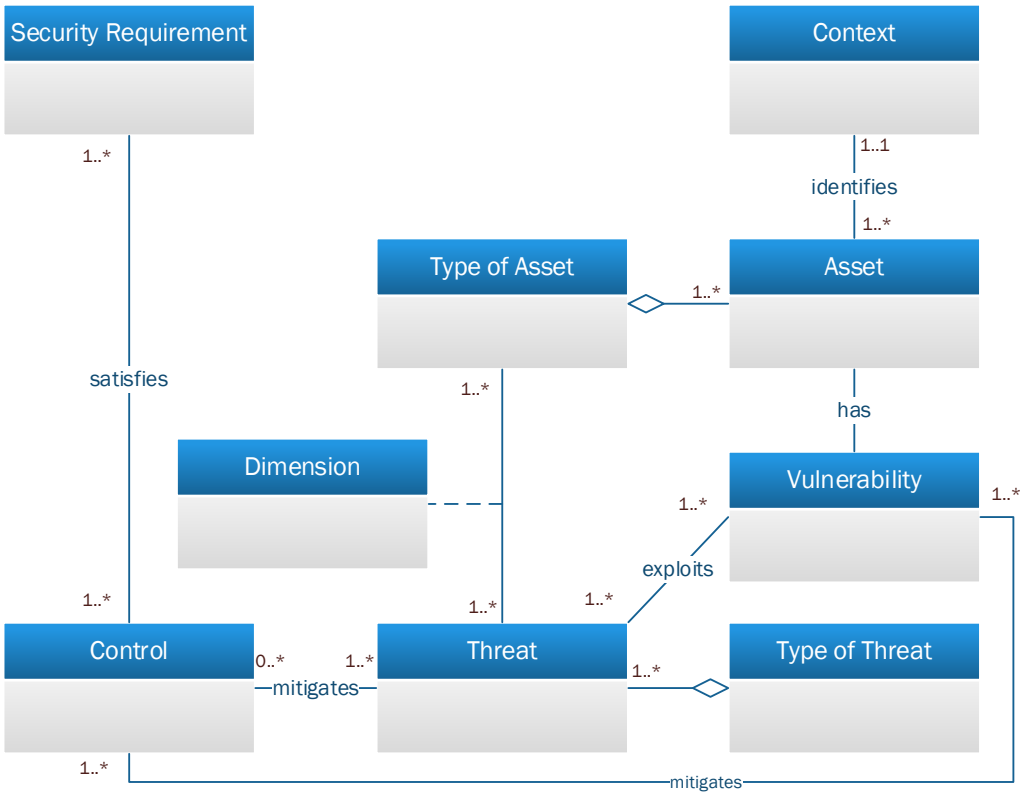


*Fig. 2. Modelo conceptual de Gestión de Riesgos*

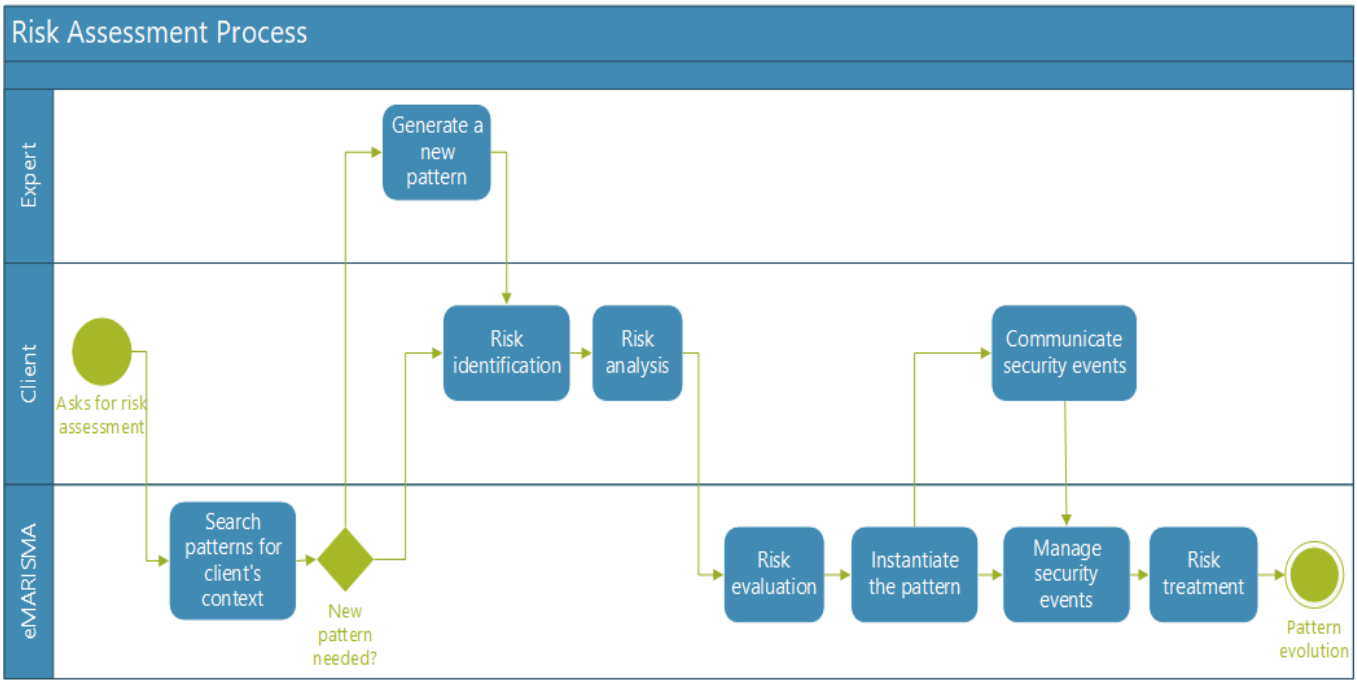


*Fig. 3. Visión general de los procesos en MARISMA*

### B. *Metodología MARISMA*

Dado que la gestión de riesgos puede considerarse como parte de la gestión de la seguridad, MARISMA intenta asociar las actividades de ambos procesos. Para lograr este objetivo se define el siguiente proceso en la Fig. 3. El éxito del proceso se basa en la reutilización de patrones, que son estructuras de conocimiento con características comunes para un contexto específico, por ejemplo, un patrón de riesgo para Big Data, un patrón de riesgo para sistemas críticos, o un patrón de riesgo para servicios web, etc.

El proceso tiene tres puntos de vista: El punto de vista del *Cliente* define todas las iteraciones y el trabajo que el cliente debe realizar durante la identificación y análisis de los riesgos del sistema (identificación de activos, amenazas, controles, incluyendo probabilidad e impacto, etc.). Esto se apoya en la herramienta eMARISMA que ayuda al cliente a calcular el nivel de riesgo y su evaluación para la toma de decisiones. El punto de vista de *eMARISMA* se encarga de gestionar automáticamente toda la información proporcionada por el cliente a través de patrones de conocimiento, gestionar todos los eventos de seguridad que pueden ocurrir en el sistema, y hacer recomendaciones sobre la mejor forma de tratarlos. El tercer punto de vista es el del *Experto*, cuya única misión en el proceso es generar un nuevo patrón inexistente que contenga las características comunes de un entorno empresarial similar, como un entorno Big Data.

A continuación, el sistema está preparado para gestionar el evento de seguridad que debe ser comunicado por el cliente. Estos eventos de seguridad generan conocimiento para adaptar los niveles asociados a los elementos de la gestión de riesgos, haciendo que el riesgo sea recalculado dinámicamente, y también adaptando los elementos asociados al patrón seleccionado permitiendo su evolución.

### C. *Meta-Patrón MARISMA*

Como se definió en [17], el meta-patrón MARISMA está formado por los elementos comunes y sus relaciones que cualquier análisis de riesgos debe tener, con el objetivo de reutilizar el conocimiento y la experiencia adquirida por los consultores al ejecutar el proceso de análisis de riesgos. Así se pudo identificar que todos los análisis de riesgos tenían controles, activos de información y amenazas, y que estos elementos estaban interrelacionados, como se muestra en nuestro modelo conceptual de la Fig. 2.

Por tanto, para diferenciar entre un patrón y un meta-patrón en nuestro marco de trabajo MARISMA, un patrón contiene, por lo tanto, los elementos necesarios para llevar a cabo un proceso de análisis y gestión de riesgos en un contexto específico. Un patrón se basa en un meta-patrón más general que contiene todos los elementos necesarios y sus relaciones para un análisis de riesgos sin tener en cuenta el contexto específico de cada organización o empresa. La organización o empresa es la encargada de crear el patrón específico que mejor se adapte al contexto (punto de vista del experto en la Fig. 3), instanciando el meta-patrón y tomando los elementos que considere necesarios para llevar a cabo el análisis de riesgos de su empresa u organización. Este meta-patrón se muestra en la Fig. 4 y ha sido implementado en la herramienta eMARISMA.

La definición de cada uno de los elementos que forman el meta-patrón ha sido descrito en detalle en [17], y se ha definido siguiendo nuestro modelo conceptual (mostrado en la Fig. 2) que define todos los conceptos para realizar una gestión de riesgos. De esta forma, el meta-patrón contiene los

elementos necesarios para representar los conceptos esenciales y sus relaciones en un análisis de riesgos.

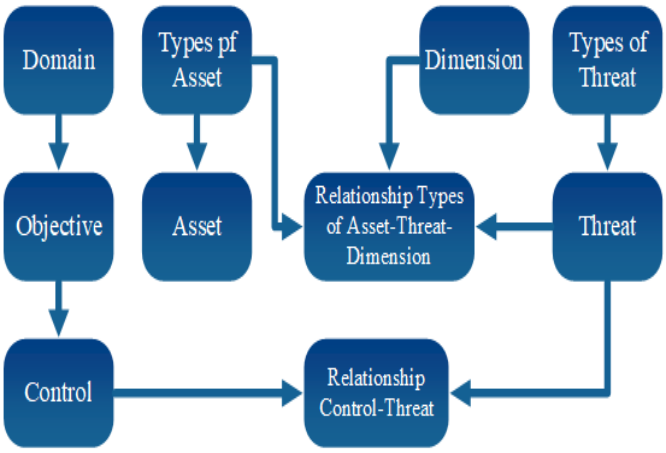


*Fig. 4 Meta-patrón en MARISMA*

Los elementos que aparecen definidos en la Fig. 4 tienen una relación directa con el modelo conceptual de la Fig. 2 de la siguiente forma:

- Dominio y Objetivo de Control: Estos elementos se definen a partir del concepto de *Context* de nuestro modelo conceptual.
- Control: En nuestro modelo conceptual hemos definido el concepto de *Control*.
- Tipos de Activos y Dimensiones: En el modelo conceptual, hay tres conceptos como *Asset*, *Types of Asset* y *Dimension* que los representan.
- Tipos de Amenazas y Amenazas: Estos conceptos están definidos en nuestro modelo conceptual como *Threat* y *Types of Threat.*
- La relación Control-Aamenaza: Con este elemento podemos gestionar el concepto de *Vulnerability* para nuestro modelo conceptual.
- La relación Tipos de Activos - Amenaza - Dimensión: Este elemento representa la relación entre *Types of Asset*, *Threat* y *Dimension* de nuestro modelo conceptual.

Estos elementos son necesarios y suficientes para iniciar el proceso de análisis de riesgos y pueden definir el resto de las relaciones de nuestro modelo conceptual como puede ser "*mitiga*" entre los conceptos de Control y Vulnerabilidad, y la relación "*tiene*" entre Activo y Vulnerabilidad. El concepto de "*Vulnerabilidad*" se obtendrá a través de las relaciones existentes del meta-patrón conociendo la falta de controles y amenazas por parte del elemento relación *control-amenaza*.

### *D. Herramienta eMARISMA*

Para apoyar la metodología MARISMA, se desarrolló una herramienta llamada eMARISMA como SaaS y utilizando tecnología Java, que soporta todos los procesos de la metodología, con bajo coste de mantenimiento, dinámico y monitorización en tiempo real.

eMARISMA dispone de diferentes zonas donde el usuario puede gestionar toda la información relacionada con la gestión de riesgos, la visualización y creación de patrones, la generación de relaciones entre conceptos, y el cálculo del riesgo y los eventos de seguridad. También dispone de una zona de visualización y cuadro de mando. En el caso de estudio podemos ver las capturas de pantalla de la herramienta. Las principales funciones de la herramienta eMARISMA son:

- Visualizar los diferentes patrones existentes y utilizarlos como base para crear otros patrones (por ejemplo, patrones sectoriales).
- Obtener un mapa detallado de la situación actual (identificación de riesgos) y un plan de recomendaciones sobre cómo mejorarla (evaluación de riesgos).
- El sistema realiza automáticamente una evaluación de riesgos y calcula el plan de tratamiento de riesgos más adecuado para que la empresa alcance un nivel de riesgo dentro de los límites definidos de una forma óptima.
- La herramienta representa un cuadro de mando con los niveles de seguridad que la empresa tiene en todo momento, para que el riesgo de la empresa pueda ser monitorizado en tiempo real.

## III. PATRÓN MARISMA-BIDA

La versatilidad de la metodología MARISMA permite el desarrollo de un marco de gestión de seguridad especializado en el contexto de Big Data. Para ello, se definiría un patrón especializado (llamado MARISMA-BiDa). Este es un marco genérico que puede aplicarse a cualquier ecosistema de Big Data.

Como hemos comentado anteriormente, el meta-patrón definido es genérico y válido para cualquier contexto, y debe ser el experto el que cree un patrón dependiente del contexto basado en este meta-patrón e instanciarlo con los conceptos específicos relacionados con la empresa. Así, por ejemplo, para un contexto de Big Data, el experto debe crear un patrón con dominios, objetivos y controles específicos, un conjunto de amenazas para Big Data, así como los activos más comunes en este tipo de entornos y sus dimensiones (por ejemplo, Volumen y Velocidad). El cliente utilizará este nuevo patrón e instanciará estos elementos con nombres específicos, por ejemplo, MongoBD como activo, robo de identidad como amenaza, etc.

Todos estos elementos instanciados han sido definidos en un trabajo previo [17] donde por la parte de los dominios, objetivos y controles se ha tomado como base la normativa ISO/IEC 27000; para las dimensiones se han considerado las diferentes Vs típicas de los sistemas Big Data [37]–[39]; para especificar los diferentes tipos de activos se han considerado principalmente la arquitectura de referencia para Big Data realizada por la organización NIST [40] junto con ENISA [41] y Cloud Security Alliance (CSA) [42]; para definir los diferentes tipos de amenazas que pueden afectar a un entorno Big Data se han seguido las recomendaciones dadas por la Agencia de Seguridad de las Redes y de la Información de la Unión Europea (ENISA) [43].

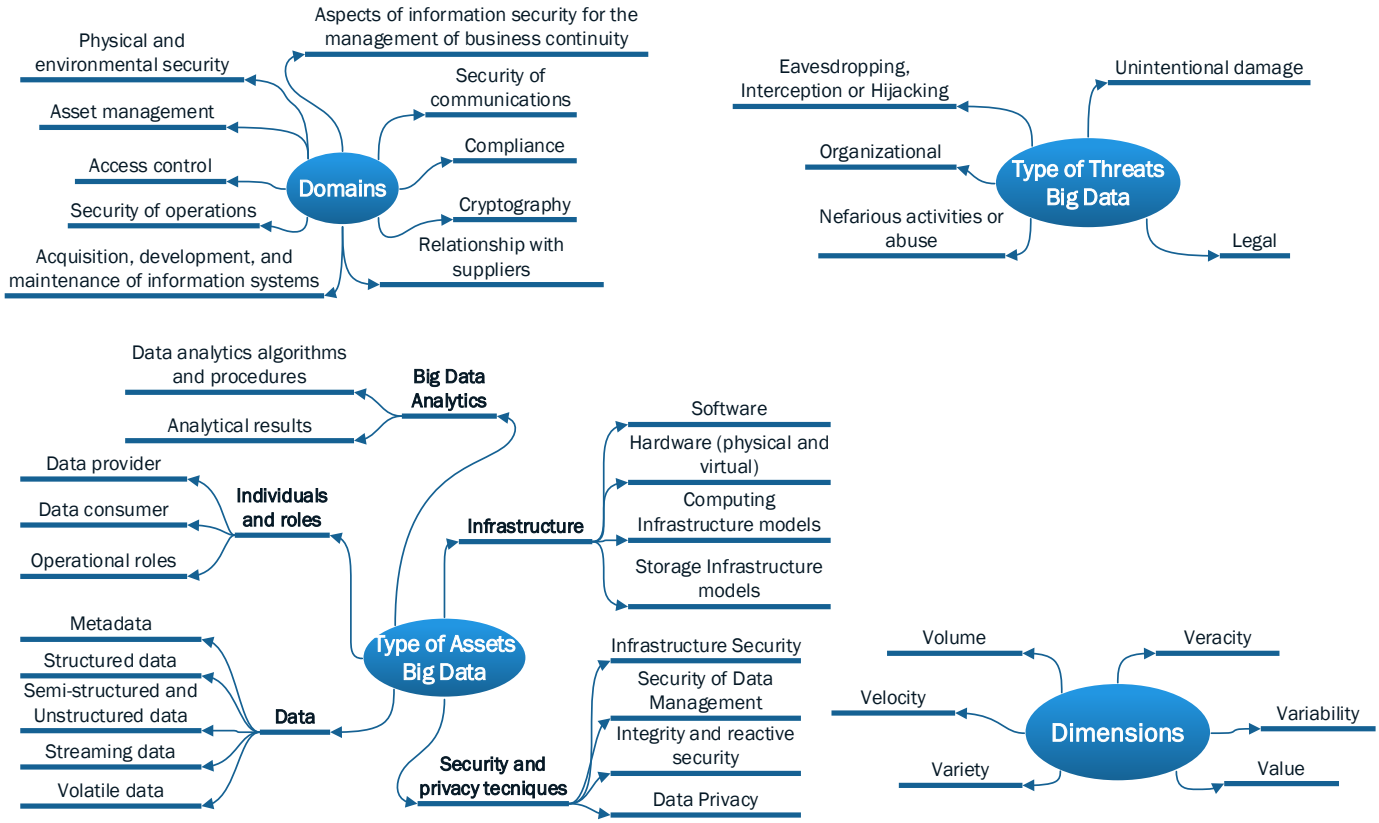


*Fig. 5 Elementos que forman parte del patrón MARISMA-BiDa*

La Fig. 5 muestra todos los elementos que forman parte del patrón de forma gráfica. Así se pueden ver el conjunto de dominios, los tipos de activos categorizados por grupos de activos, los tipos de amenazas y las dimensiones a tener en cuenta en el patrón para Big Data. Los detalles de cada uno de estos elementos y subelementos se pueden obtener en [17].

Para la definición de las diferentes amenazas que se pueden identificar en este tipo de sistemas, se ha definido una matriz en la que se integran tanto los diferentes tipos de amenazas como las diferentes dimensiones de Big Data. Si una amenaza no encaja en ninguna de las celdas de la matriz, se puede concluir que no es una amenaza típica de un entorno Big Data.

La Tabla 1 muestra un ejemplo de cómo se pueden identificar las diferentes amenazas para los entornos de Big Data relacionadas con una o varias dimensiones. Esta relación se ha definido mediante la realización de un amplio estudio tanto de las amenazas como de las dimensiones sin perder de vista el contexto y los activos más afectados como consecuencia de la materialización de estas amenazas. Con la experiencia del equipo de investigación en estos temas, junto con la colaboración de expertos en seguridad y con la ayuda de los informes de ENISA, ha sido posible completar satisfactoriamente esta relación.

Por ejemplo, un tipo de amenaza es la ”i*nterceptación de información”* de la “*escucha clandestina*”, donde puede producirse la interceptación de información en las comunicaciones entre aplicaciones Big Data, y donde los protocolos de comunicación son raramente seguros entre este tipo de aplicaciones (sin el uso de TLS y SSL). Se considera que este tipo de amenaza afecta en mayor medida a las dimensiones *“Veracidad”* y *“Valor”*. Consideramos que afecta a la dimensión de la "*veracidad*" porque al interceptar la información en la red (por ejemplo, ha sido modificada) no es posible estar seguro de que esa información sea exacta, lo que puede llevar a tomar una decisión equivocada. También afecta a la dimensión "*valor*" porque si alguien ha sido capaz de interceptar la comunicación, puede haber revelado la información oculta de los datos y, en consecuencia, el valor se pierde. Este tipo de estudio y razonamiento para cada uno de los tipos de amenazas y amenazas definidas para nuestro patrón MARISMA-BiDa ha sido realizado, cuyos resultados se muestran en la Tabla 1.

Este patrón así definido servirá para ser instanciado en un nuevo proceso de análisis de riesgos con la herramienta eMARISMA en un entorno Big Data como es el caso de estudio que se presenta a continuación.

| | Unintentional damage | Eavesdropping | Nefarious | Legal | Organizational |
|---|---|---|---|---|---|
| **Volume** | Configuration error<br>Destruction of records | | Abuse of Information<br>Manipulation of hardware and software | Violation of laws or legislation | Skill shortage |
| **Variety** | Destruction of records<br>Inadequate design | | Misuse of audit tools | | Skill shortage |
| **Velocity** | Inadequate design<br>Erroneous use | Network Reconnaissance<br>Replay of messages | DoS/DDoS<br>Malicious code<br>Remote activity<br>Manipulation of hardware and software<br>Misuse of audit tools<br>Brute force<br>Generation and use of rogue certificates | | Skill shortage |
| **Variability** | Unintentional change of data<br>Inadequate design | | | | Skill shortage |
| **Veracity** | Leaks of data via Web applications<br>Loss of sensitive information<br>Loss of information in the cloud<br>Damage caused by a third party<br>Using information from an unreliable source | Intercepting compromising emissions<br>War driving<br>Interfering radiation<br>Man-in-the-middle<br>Interception of information | Code injection<br>Social Engineering<br>Abuse of authorizations<br>Receive of unsolicited E-mail<br>Identity theft<br>Hoax<br>Compromising confidential information<br>Generation and use of rogue certificates<br>Unauthorized activities | Failure to meet contractual requirements<br>Unauthorized use of IPR protected resources | Skill shortage |
| **Value** | Information leakage<br>Loss of devices, storage media and documents<br>Unintentional change of data<br>Damages resulting from penetration testing<br>Inadequate design | Man-in-the-middle<br>Interception of information | Malicious code<br>Abuse of Information<br>Failed of business process<br>Targeted attacks<br>Manipulation of information<br>Misuse of information<br>Unauthorized installation of software/Manipulation of Algorithms | Judiciary decisions<br>Abuse of personal data | Skill shortage |
| **Affected Assets** | • Data<br>• Infrastructure<br>• Big Data analytics | • Data | • Infrastructure<br>• Data<br>• Big Data analytics<br>• Security and privacy techniques | • Data | • Individuals and Roles |

*Tabla 1. Matriz de Identificación de "Tipos de Activos – Amenazas - Dimensiones" para entornos Big Data.*

## IV. Caso de estudio: Datos historia clínica electrónica

NIST define diferentes casos de uso para un amplio conjunto de dominios de aplicación, y uno de ellos ha sido seleccionado por su interesante aplicación en el patrón MARISMA-BiDa a un caso particular en un entorno Big Data. El caso de estudio seleccionado se refiere a los datos de registros médicos electrónicos cuyo objetivo es "*Utilizar métodos avanzados para normalizar la identificación de pacientes, proveedores, instalaciones y conceptos clínicos dentro y entre organizaciones de atención de la salud separadas para mejorar los modelos de definición y extracción de fenotipos clínicos a partir de datos clínicos de texto libre y discreto no estándar mediante la selección de características, la recuperación de información y los modelos de toma de decisiones de aprendizaje automático. Aprovechar los datos del fenotipo clínico para apoyar la selección de cohortes, la investigación de resultados clínicos y el apoyo a la toma de decisiones clínicas"* (Caso de Uso 16 de [44]).

Este caso de estudio influye en todas las dimensiones definidas por el patrón MARISMA-BiDa (como se muestra en la Fig. 6), por ejemplo:

- *Volumen (DIM1)*: más de 12 millones de pacientes, más de 4.000 millones de observaciones clínicas y más de 20 TB de datos brutos.
- *Velocidad (DIM2)*: entre 500.000 y 1,5 millones de nuevas transacciones clínicas en tiempo real añadidas al

día.

- *Variedad (DIM3)*: una amplia variedad de conjuntos de datos clínicos de múltiples fuentes: notas del proveedor de texto libre; informes de laboratorios y servicios de urgencias; estudios de química, cardiología o hematología; estudios de bancos de sangre y toxicología, etc.
- *Veracidad (DIM4)*: los datos de cada fuente clínica se recopilan comúnmente utilizando diferentes métodos y representaciones, lo que produce una heterogeneidad sustancial. Esto conduce a errores y sesgos sistemáticos que requieren métodos robustos para crear la interoperabilidad semántica.
- *Variabilidad (DIM5)*: los datos de los sistemas clínicos evolucionan con el tiempo porque el espacio conceptual clínico y biológico está en constante evolución. Los nuevos descubrimientos científicos conducen a nuevas entidades de la enfermedad, nuevas modalidades de diagnóstico y nuevos enfoques de manejo de la enfermedad.
- *Valor (DIM6)*: Métodos de recuperación de información para identificar características clínicas relevantes, es decir, modelos de decisión utilizados para identificar una variedad de fenotipos clínicos como la diabetes, la insuficiencia cardíaca congestiva y el cáncer de páncreas.

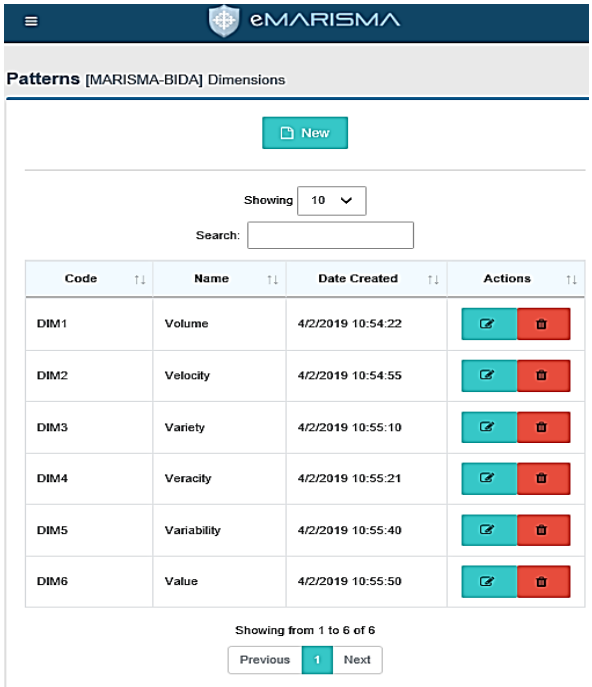


*Fig. 6 Dimensions defined on the eMARISMA tool*

Los pasos del proceso MARISMA comienzan con la identificación de los activos involucrados en el sistema del tipo de activos definidos por el patrón MARISMA-BiDa. Este conjunto de activos involucrados se muestra en la Tabla 2, y en la Fig. 7 se muestran los activos añadidos a la herramienta eMARISMA.

Una vez que los activos para nuestro caso de estudio han sido identificados y añadidos a la herramienta eMARISMA, se realiza automáticamente la evaluación de riesgos utilizando todas las relaciones y matrices definidas en nuestro patrón MARISMA-BiDa. La herramienta relaciona entonces el activo definido en nuestro caso con las amenazas y dimensiones afectadas. También muestra el resultado del riesgo actual para este conjunto de activos y vincula los objetivos con los controles necesarios para proteger este conjunto de activos.

Para el conjunto de amenazas identificadas en el patrón MARISMA-BiDa, la herramienta eMARISMA toma valores por defecto basados en la experiencia de análisis de riesgo previo, así como para la probabilidad de ocurrencia y el porcentaje de degradación, los cuales pueden ser modificados de acuerdo a nuestros criterios, experiencia y entorno del sistema a evaluar.

| **MARISMA-BiDa PATTERN** | | **PATTERN INSTANTIATED** |
|---|---|---|
| **Asset group** | **Asset Type** | **Assets of Case use** |
| **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 |

*Tabla 2. Activos involucrados en el caso de estudio para los tipos de activos definidos en el patrón MARISMA-BiDa*

La herramienta eMARISMA, con los datos introducidos en el patrón, inicializa todos los valores a valores por defecto, que pueden cambiar con el tiempo a medida que se producen eventos de seguridad, ya que, como se menciona en la sección II.B, la evaluación de riesgos puede recalcularse a medida que se producen eventos de seguridad, como un ataque específico

al sistema. La herramienta muestra los valores por defecto tanto para la probabilidad de ocurrencia de las amenazas como para el porcentaje en que degradan los criterios de riesgo de los activos. Como la herramienta es utilizada por usuarios expertos en seguridad, tienen una idea clara de qué tipos de amenazas son las más comunes en su entorno y cómo puede degradar los activos si la amenaza tiene éxito. De esta manera, pueden cambiar libremente esos valores por otros más acordes con el contexto en el que opera el sistema a discreción de sus expertos. Sin embargo, en una primera evaluación no es necesario cambiar ningún valor, ya que la herramienta realiza una evaluación preliminar con los valores por defecto y los modifica cuando se producen ataques o intentos de ataque.

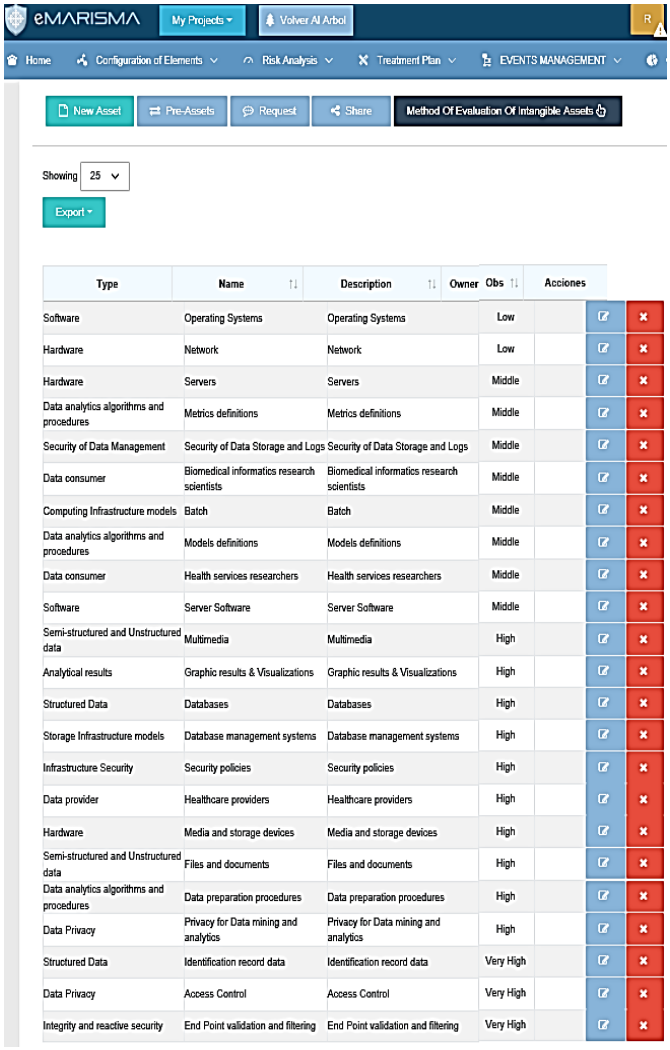


*Fig. 7 Assets for the case use added on the eMARISMA tool*

Por ejemplo, como muestra la Fig. 8, la amenaza "*daños causados por terceros*" puede actualizar la probabilidad de ocurrencia y el porcentaje de degradación hasta en un 60%. Este porcentaje se obtiene a partir del conocimiento del contexto en el que se encuentra el sistema. En este caso, es de suma importancia considerar que se trata de un contexto en el que muchas personas están involucradas y desean tener acceso a este tipo de datos sensibles, o incluso empresas externas que se encargan de gestionar y proteger este tipo de datos sensibles. Esto puede causar daños involuntarios debido al mal uso de dichos datos al no seguir ninguna política de seguridad, o a una mala configuración en la base de datos que los almacena, o simplemente debido a la eliminación accidental de tales datos que causaría un daño significativo con respecto a los datos médicos, incluyendo una violación de la ley, ya que este tipo de datos sensibles no están bien protegidos. Por lo tanto, el usuario experto puede considerar que el porcentaje de degradación es alto (60%) y que la probabilidad de que ocurra es mayor de lo esperado, por ejemplo, 60% también. Estos valores se modifican en la herramienta porque se consideran apropiados para el entorno, y el nivel de riesgo se evalúa con los valores introducidos, considerando todos los elementos definidos para nuestro patrón (activos, dimensiones, controles, etc.).

El usuario experto puede modificar o actualizar los valores necesarios debido a las particularidades del escenario considerado y por el amplio conocimiento del sistema y del entorno.

Una vez seleccionados los valores más adecuados para las amenazas, la herramienta eMARISMA aplica la matriz "*Tipo de Activos-Amenazas-Dimensiones*" para calcular el porcentaje de degradación pero, esta vez, incluyendo las dimensiones afectadas para cada amenaza y tipo de activo definido en el patrón MARISMA-BiDa que se puede ver en la Tabla 3. Como se ha mencionado anteriormente, estos valores pueden ser modificados para cada dimensión de acuerdo a nuestros criterios, experiencia y conocimiento del entorno. La Fig. 9 muestra la matriz con los valores de la herramienta eMARISMA.

Finalmente, la Fig. 10 y la Fig. 11 muestran los informes creados por la herramienta. Representa diagramas kiviat y mapa de riesgos para los niveles de seguridad que la organización de salud tiene en cada momento, por lo que puede ser considerado como un tablero de control para monitorear el riesgo de la organización en tiempo real. De esta forma, permite a la alta dirección de la empresa tomar decisiones en función de los resultados obtenidos y de su apetito de riesgo.

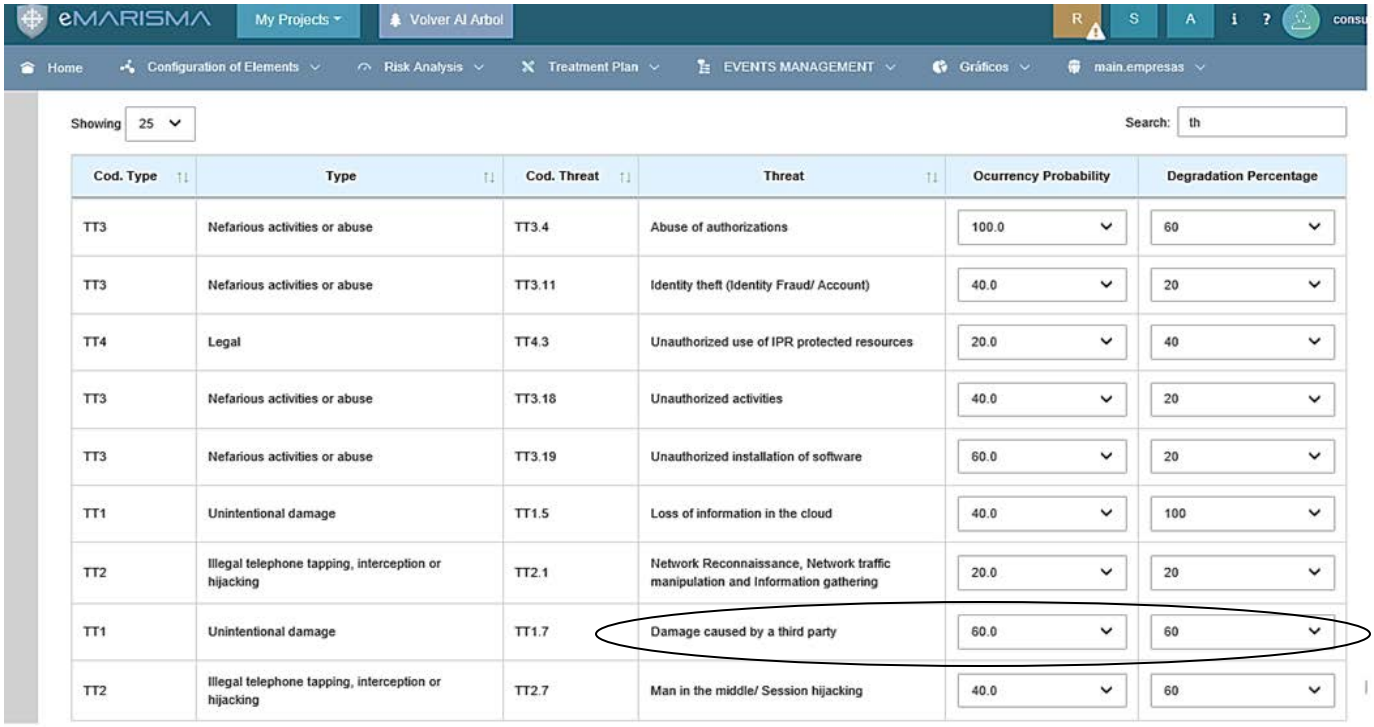


*Fig. 8. Amenazas con la probabilidad de ocurrencia y el porcentaje de degradación en la herramienta eMARISMA para el caso de estudio.*

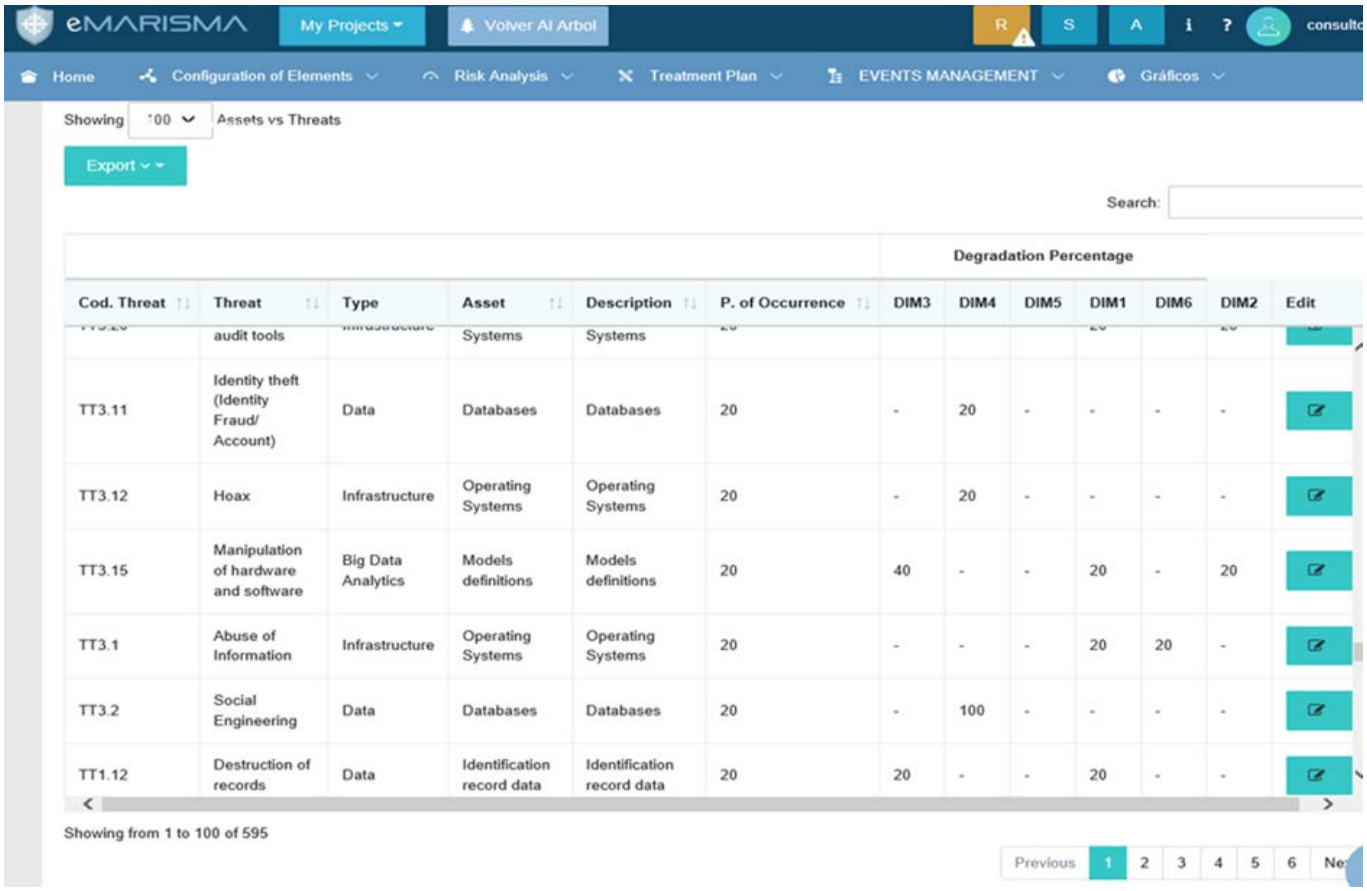


*Fig. 9. Porcentaje de degradación para las dimensiones afectadas por los activos y amenazas de eMARISMA para nuestro caso de estudio.*

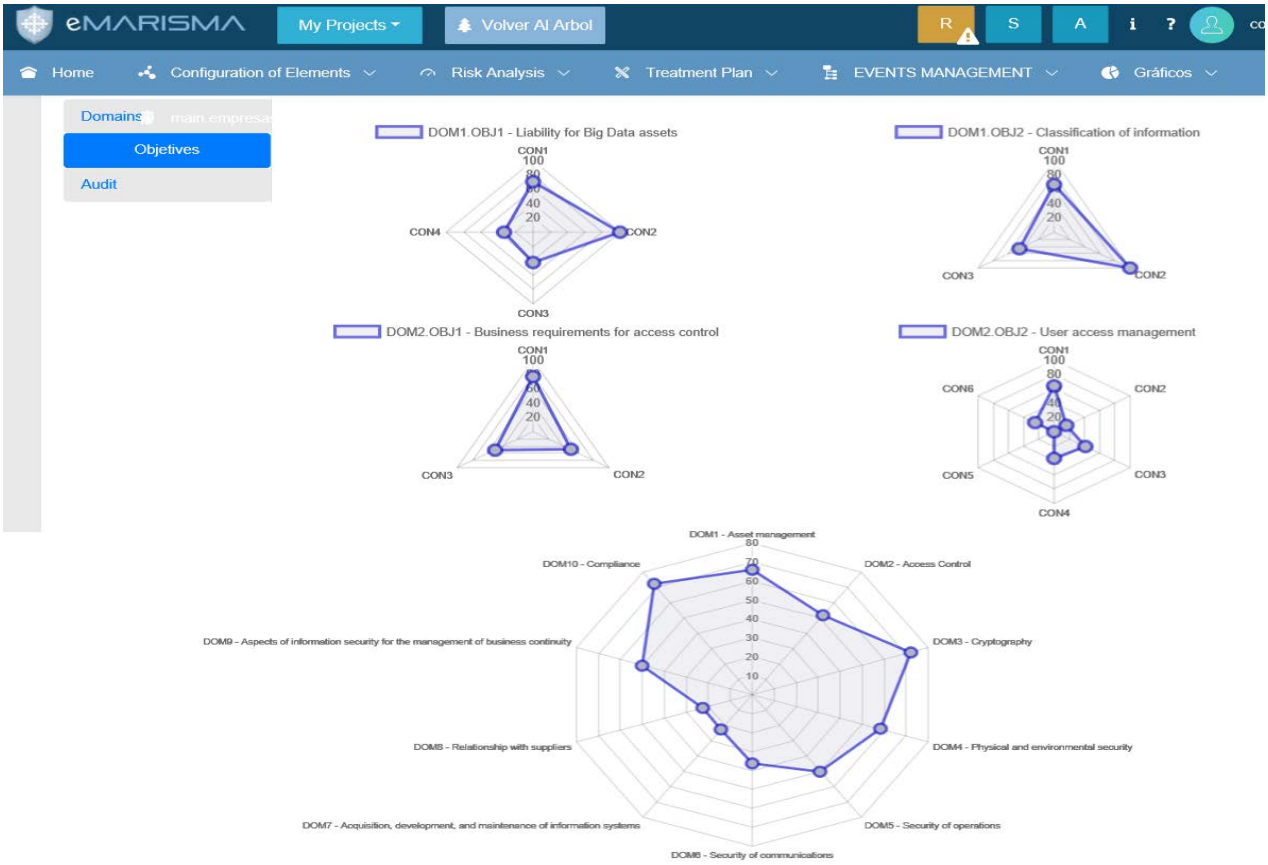


*Fig. 10. Diagramas Kiviat generados por la herramienta eMARISMA para los objetivos y auditoría alcanzados en nuestro caso de estudio.*

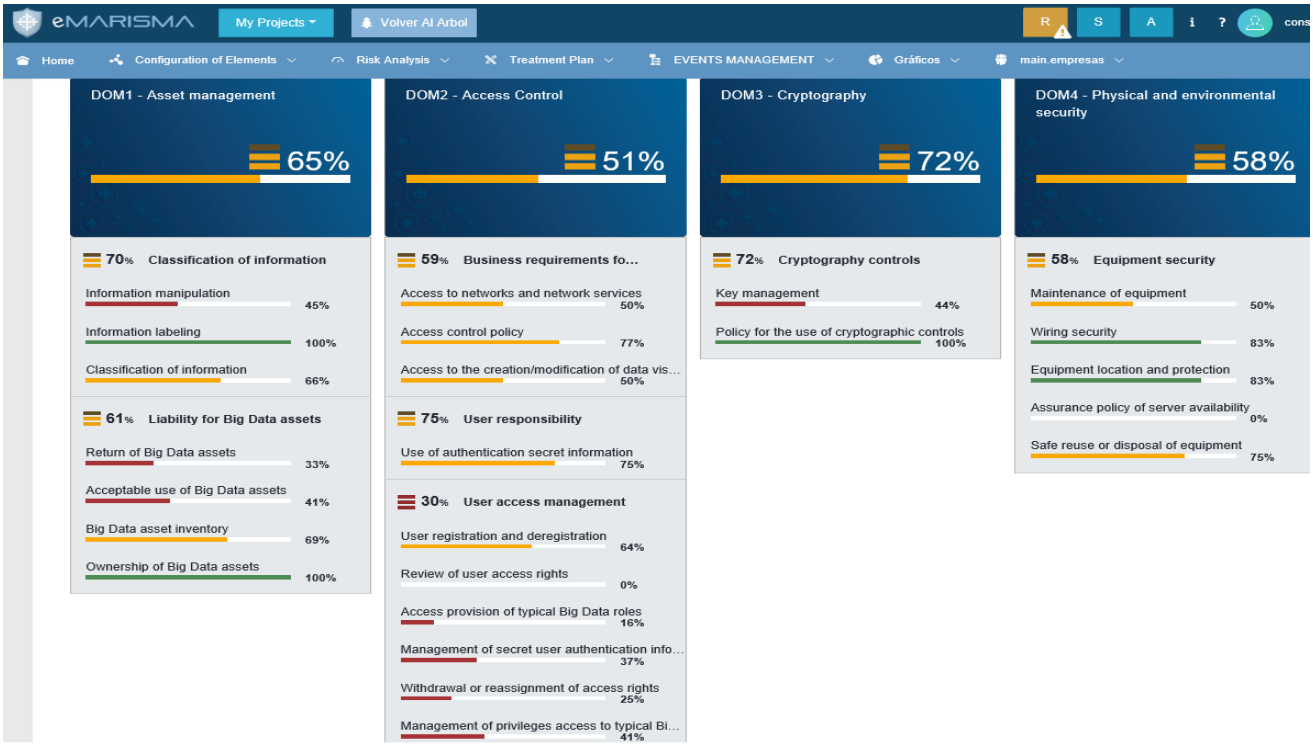


*Fig. 11. Cuadro de mando generado por la herramienta eMARISMA para nuestro caso de estudio.*

## V. Conclusiones y trabajo futuro

Un proceso de evaluación y gestión de riesgos puede dividirse en varias etapas, incluyendo el establecimiento del contexto, la identificación de riesgos, el análisis de riesgos en términos de probabilidad e impacto, la evaluación de riesgos y, finalmente, el tratamiento de riesgos.

Este trabajo muestra cómo se utiliza la metodología MARISMA (apoyada por la herramienta eMARISMA), para generar un patrón de gestión y análisis de seguridad enfocado en aspectos de Big Data, que permita una gestión dinámica del riesgo asociado a los elementos de un entorno de Big Data en una empresa.

Esta propuesta se ha aplicado en un caso de estudio, cuya aplicación ha permitido afinarla y validarla con esa experiencia. Estos refinamientos se han centrado principalmente en ajustar los principales conceptos del patrón MARISMA-BiDa, reafirmando los conceptos más relevantes de los ya identificados, y encontrando otros basados en la experiencia. Como trabajo futuro, se contempla la evolución de la herramienta eMARISMA como un sistema de aprendizaje en la nube. Permitirá incorporar los incidentes de seguridad que afecten a uno de los sistemas, en todos aquellos sistemas que estén relacionados o puedan verse afectados.

## Agradecimientos

Este trabajo ha sido financiado por el proyecto ECLIPSE (Ministerio de Economía, Industria y Competitividad de España y el Fondo Europeo de Desarrollo Regional FEDER, RTI2018-094283-B-C31), y el proyecto GENESIS (Consejería de Educación, Cultura y Deportes de la Dirección General de Universidades, Investigación e Innovación de la Junta de Comunidades de Castilla-La Mancha, España, SBPLY/17/180501/000202). Agradecemos la ayuda de las compañías Sicaman Nuevas Tecnologías SL (www.sicaman-nt.com) y Marisma Shield SL (www.emarisma.com), que han facilitado el uso de la herramienta eMARISMA.

## References

[1] ENISA, “Good Practices and Recommendations on the,” 2015.
[2] K. Armstrong, “Big data: a revolution that will transform how we live, work, and think,” Information, Commun. Soc., vol. 17, no. 10, pp. 1300–1302, Nov. 2014.
[3] D. Broeders, E. Schrijvers, B. van der Sloot, R. van Brakel, J. de Hoog, and E. Hirsch Ballin, “Big Data and security policies: Towards a framework for regulating the phases of analytics and use of Big Data,” Comput. Law Secur. Rev., vol. 33, no. 3, pp. 309–323, Jun. 2017.
[4] D. Laney, “3D data management: Controlling data volume, velocity and variety,” META Gr. Res. note, vol. 6, no. 70, p. 1, 2001.
[5] S. Klous, “Sustainable harvesting of the Big Data potential,” Explor. Boundaries Big Data, p. 27, 2016.
[6] B. Hopkins and B. Evelson, “Expand your Digital Horizon with Big Data,” Forrester, vol. 30, 2011.
[7] F. J. Alexander, A. Hoisie, and A. Szalay, “Big Data,” Comput. Sci. Eng., vol. 13, no. 6, pp. 10–13, Nov. 2011.
[8] J. Dijcks, “Oracle: Big data for the enterprise,” Oracle White Pap., no. June, p. 16, 2012.
[9] E. Dumbill, “Making sense f big data (editorial),” Big Data, vol. 1, no. 1, 2013.
[10] P. Zikoupoulos and C. Eaton, Understanding big data: Analytics for Enterprise Class Hadoop and Streaming, vol. 11, no. 1. McGraw-Hill Osborne Media, 2016.
[11] R. Akerkar et al., “Understanding and mapping big data,” no. March. D1, 2015.
[12] S. V. Bharathi, “Prioritizing and Ranking the Big Data Information Security Risk Spectrum,” Glob. J. Flex. Syst. Manag., vol. 18, no. 3, pp. 183–201, Sep. 2017.
[13] B. Goswami and P. K. Chandra, “Risk Assessment and Analysis for Big Data,” Int. J. Adv. Res. Comput. Sci. Softw. Eng., vol. 5, no. 11, 2015.
[14] A. M. Barrientos and K. A. Areiza, “Integration of a safety management system withan information quality management system.,” Universidad EAFIT, 2005.
[15] R. Fredriksen, M. Kristiansen, B. A. Gran, K. Stølen, T. A. Opperud, and T. Dimitrakos, “The CORAS Framework for a Model-Based Risk Management Process,” LNCS 2434, 2002, pp. 94–105.
[16] G. Disterer, “ISO/IEC 27000, 27001 and 27002 for Information Security Management,” J. Inf. Secur., vol. 04, no. 02, pp. 92–100, 2013.
[17] J. Moreno, L. E. Sánchez, A. S. Olmo, D. G. Rosad, M. A. Serrano, and E. F. Medina, “Marisma-BiDa: Entorno Integrado de Análisis y Gestión de Riesgos en Big Data,” in Actas de las Cuartas Jornadas Nacionales de Investigación en Ciberseguridad, 2018, pp. 159–165.
[18] L. Ortiz Restrepo, V. Duque, and F. Javier, “Gestión de riesgos en eTOM. Un análisis comparativo con los estándares de riesgo corporativo,” Rev. Logos, Cienc. Tecnol., vol. 9, no. 1, pp. 85–99, 2017.
[19] L. D. Bodin, L. A. Gordon, and M. P. Loeb, “Information security and risk management,” Commun. ACM, vol. 51, no. 4, pp. 64–68, Apr. 2008.
[20] MAGERIT, “MAGERIT v.3 : Metodología de Análisis y Gestión de Riesgos de los Sistemas de Información,” 2012.
[21] G. Wangen, “Information Security Risk Assessment: A Method Comparison,” Computer (Long. Beach. Calif)., vol. 50, no. 4, pp. 52–61, Apr. 2017.
[22] C. Fakrane and B. Regragui, “INTERACTIONS AND COMPARISON OF IT RISK ANALYSIS METHODS,” in 2018 4th International Conference on Cloud Computing Technologies and Applications (Cloudtech), 2018, pp. 1–7.
[23] F. Jeannot, “Méthodologies d’évaluation et gestion de risques en sécurité,” Montréal, Canada, Mai 2018, R518, v1.0, 2018.
[24] A. Benavides, “Modelo de Sistema de Gestión de Seguridad de la Información Basado en la Norma NTC ISO/IEC 27001 para Instituciones Públicas de Educación Básica de la Comuna Universidad de la Ciudad de Pereira,” no. 6, pp. 67–72, 2017.
[25] Werner George Bornman, “Information Security Risk Management: a Holistic Framework,” University of Johannesburg, 2004.
[26] A. Refsdal, B. Solhaug, and K. Stølen, “Cyber-risk management,” in Cyber-Risk Management, Springer, 2015, pp. 33–47.
[27] R. Zudin, “Analysis of information risk management methods,” Univ. Jyväskylä, 2014.
[28] J. P. Carrillo Sánchez, “Guía y análisis de gestión de riesgos en la adquisición e implantación de equipamiento y servicios de tecnologías de información y comunicaciones para proyectos de alcance nacional,” Quito: EPN, 2012, 2012.
[29] D. C. Pacheco Pozo, “Propuesta de un plan de contigencia de TI para la empresa LOGICIEL,” Quito, 2016., 2016.
[30] L. E. Sánchez, a S. O. Parra, D. G. Rosado, and M. Piattini, “Managing Security and its Maturity in Small and Medium-sized Enterprises,” J. Univers. Comput. Sci., vol. 15, no. 15, pp. 3038–3058, 2009.
[31] A. Santos Olmo Parra, L. E. Sanchez Crespo, E. Alvarez, M. Huerta, and E. Fernandez Medina Paton, “Methodology for Dynamic Analysis and Risk Management on ISO27001,” IEEE Lat. Am. Trans., vol. 14, no. 6, pp. 2897–2911, Jun. 2016.
[32] J. Pirrone and M. Huerta, “Security Mechanism for Medical Record Exchange Using Hippocratic Protocol,” in IFMBE Proceedings, vol. 68, no. 1, Springer Verlag, 2019, pp. 401–404.
[33] T. Vivas, A. Zambrano, and M. Huerta, “Mechanisms of security based on digital certificates applied in a telemedicine network,” in 2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2008, pp. 1817–1820.
[34] A. Santos-Olmo, L. Sánchez, D. Rosado, E. Fernández-Medina, and M. Piattini, “Applying the Action-Research Method to Develop a Methodology to Reduce the Installation and Maintenance Times of Information Security Management Systems,” Futur. Internet, vol. 8, no. 3, p. 36, Jul. 2016.
[35] ISO 31000, “ISO 31000:2018 Risk Management,” ISO, 2018.

[36] R. Kelemen, M. Biskup, and N. B. Redep, “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), 2016, pp. 1539–1545.
[37] M. A. Khan, M. F. Uddin, and N. Gupta, “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, 2014, pp. 1–5.
[38] Z. Sun, K. Strang, and R. Li, Big Data with Ten Big Characteristics. 2019.
[39] M. Chen, S. Mao, and Y. Liu, “Big Data: A Survey,” Mob. Networks Appl., vol. 19, no. 2, pp. 171–209, Apr. 2014.
[40] NIST, “NIST Big Data Interoperability Framework: Volume 1, Definitions, version 2,” NIST Special Publication 1500-1r1, Jun-2018. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-4r1.pdf.
[41] E. Rekleitis, “Big Data Threat Landscape and Good Practice Guide,” Eur. Union Agency Netw. Inf. Secur., no. January, 2016.
[42] P. Murthy, A. Bharadwaj, P. Subrahmanyam, A. Roy, and S. Rajan, “Big Data Taxonomy,” Cloud Security Alliance, no. September. Cloud Security Alliance, September, p. 33, 2014.
[43] ENISA, “Threat Landscape and Good Practice Guide for Software Defined Networks/5G - SDN Threat Landscape,” Jan-2016. .
[44] NIST, “NIST Big Data Interoperability Framework: volume 3, use cases and general requirements,” NIST Special Publication 1500-3r1, Jun-2018. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-3r1.pdf.

**David G. Rosado** tiene un Máster y es doctor en Informática por la Universidad de Málaga y por la Universidad de Castilla-La Mancha, respectivamente. Es profesor titular en la Escuela Superior de Informática de la Universidad de Castilla-La Mancha en Ciudad Real. Su actividad investigadora se centra en seguridad de sistemas de información, Cloud Computing y Big Data. Sobre estos temas, él ha publicado numerosos artículos en conderencias de ámbito nacional e internacional, también es editor y coeditor de varios libros. Es autor de varias publicaciones en revistas nacionales e internacionales (Information Software Technology, System Architecture, Network and Computer Applications, etc.). Él es miembro del comité de programa de numerosas conferencias y workshops nacionales e internacionales. Es miembro del grupo de investigación GSyA del departamento de sistemas de información y tecnologías de la Universidad de Castilla-La Mancha. Su correo electrónico es david.grosado@uclm.es.

**Julio Moreno** es Máster y estudiante de doctorado en Informática por la Universidad de Castilla-La Mancha. Su investigación se centra en la seguridad y privacidad de los datos, así como en la creación de un marco de gobierno para Big Data. Es miembro del grupo de investigación GSyA del departamento de sistemas de información y tecnologías de la Universidad de Castilla-La Mancha. Su correo electrónico es julio.moreno@uclm.es.

**Luis Enrique Sánchez** es Doctor y Máster en Informática y es profesor ayudante doctor de la Universidad de Castilla-La Mancha (Ciudad Real, España), Master en Auditoría de Sistemas de Información por la Universidad Politécnica de Madrid, y Auditor Certificado de Sistemas de Información por ISACA. Es Director de los departamentos de Servicios Profesionales y de I+D de la empresa Sicaman Nuevas Tecnologías S.L. Sus actividades de investigación son sistemas de seguridad de gestión, métricas de seguridad, minería de datos, limpieza de datos e inteligencia de negocios. Participa en el grupo de investigación GSyA del departamento de sistemas de información y tecnologías de la Universidad de Castilla-La Mancha, en Ciudad Real (España). Su correo electrónico es luisenrique@sanchezcrespo.org.

**Antonio Santos-Olmo** es Licenciado en Informática y es profesor asociado de la Escuela Superior de Informática de la Universidad de Castilla-La Mancha en Ciudad Real (España), Master en Auditoría de Sistemas de Información por la Universidad Politécnica de Madrid, y Auditor Certificado de Sistemas de Información por ISACA. Es Director de los departamentos de Software Factory de la empresa Sicaman Nuevas Tecnologías S.L. Sus actividades de investigación son sistemas de seguridad de gestión, métricas de seguridad, minería de datos, limpieza de datos e inteligencia de negocio. Participa en el grupo de investigación GSyA del Departamento de de sistemas de información y tecnologías de la Universidad de Castilla-La Mancha, en Ciudad Real (España). Su correo electrónico es asolmo@sicaman-nt.com.

**Manuel A. Serrano** es Máster y doctor en Informática por la Universidad de Castilla-La Mancha. Es profesor titular en la Escuela Superior de Informática de la Universidad de Castilla-La Mancha en Ciudad Real. Su investigación se centra en la calidad de sofotware de los datos, la medición de software y la calidad y medición de almacenes de datos y Big Data. Su correo electrónico es manuel.serrano@uclm.es.

**Eduardo Fernández-Medina** es Máster y doctor en Informática por la Universidad de Castilla-La Mancha. Es profesor catedrático en la Escuela Superior de Informática de la Universidad de Castilla-La Mancha en Ciudad Real (España). Su actividad investigadora se centra en el campo de la seguridad de sistemas de información, en particular en seguridad en Big Data, Cloud Computing y sistemas ciberfísicos. En estas temáticas, él es coeditor de varios libros y capítulos de libros, y ha publicado numerosos artículos en conferencias nacionales e internacionales (BPM, UML, ER, ESORICS, TRUSTBUS, etc.). Es autor de más de cincuenta publicaciones en revistas internacionales (Decision Support Systems, Information Systems, ACM Sigmod Record, Information Software Technology, Computer & Security, Computer Standards and Interfaces, etc.). Él lidera el grupo de investigación GSyA del departamento de sistemas de información y tecnologías de la Universidad de Castilla-La Mancha y pertenece a varias asociaciones profesionales y de investigación (ATI, AEC, AENOR, etc.). Su correo electrónico es eduardo.fdezmedina@uclm.es.