Hacía un marco de definición de requisitos para Big Data

Julio Moreno[1], Manuel A. Serrano[2], Eduardo Fernández-Medina[1]

[1] Grupo de investigación de Seguridad y Auditoría (GSyA), Universidad de Castilla-La Mancha, Ciudad Real, España
[2] Grupo de investigación Alarcos, Universidad de Castilla-la Mancha, Ciudad Real, España

{Julio.Moreno, Manuel.Serrano, Eduardo.FdezMedina}@uclm.es

Abstract. Big Data permite la toma de mejores decisiones por parte de las organizaciones y cada vez tiene una mayor importancia en el día a día de las compañías. Pero diseñar sistemas Big Data de forma que no haya conflicto con los objetivos de negocio y las diferentes regulaciones que lo afectan es un tema complejo. Como forma de abordar este problema, y debido a la falta de propuestas en este tema, proponemos un modelo de requisitos para Big Data, diferenciando el caso de requisitos de seguridad, por responder éstos a la existencia de amenazas y vulnerabilidades, además de ser obligatorio (o muy recomendado) cumplir con ciertas legislaciones o regulaciones en materia de seguridad.

Keywords: Modelado de requisitos, Big Data, Seguridad

1 Introducción

Organizaciones de cualquier ámbito son cada vez más conscientes de la importancia de Big Data [1]. Para todas, los datos son fundamentales para llevar a cabo sus actividades cotidianas, así como, ayudar a la gerencia en la toma de decisiones con la información extraída de dichos datos [2]. Big Data implica un cambio respecto a los sistemas tradicionales, fundamentalmente en tres aspectos: la cantidad de datos (volumen), el ratio de generación y transmisión de los datos (velocidad) y la heterogeneidad de los tipos de datos estructurados o no estructurados que puede manejar (variedad) [3]. Estas propiedades son conocidas como las tres Vs de Big Data [4]. Diversos autores han ido añadiendo nuevas Vs a este grupo inicial, como el valor o veracidad de los datos [5]. Con cada nueva tecnología surgen nuevos problemas, Big Data no es una excepción. Estos problemas no solo están relacionados con las Vs típicas de Big Data sino que también se encuentran en el ámbito de la privacidad y seguridad de los datos. Big Data no solo incrementa la escala de los problemas de privacidad y seguridad sino que también añade nuevos que deben ser gestionados mediante diferentes técnicas [6].

M. Genero, M. Kalinowski (Eds.): CIbSE 2018, Bogotá - Colombia, 2018

Aun con estas diferencias, un entorno Big Data no deja de ser un proyecto de Tecnologías de la Información (TI), cuyo desarrollo debe ser guiado por una serie de modelos o metodologías. Según una encuesta realizada a expertos del campo de las TI [7], uno de los principales retos que se encuentran a la hora de operar un entorno Big Data es su poca alineación con los objetivos de negocio de la compañía. Una forma de solucionar este problema es el uso de un marco de definición de requisitos cuyo primer paso sea tener como entrada el contexto en el que se va a encontrar el sistema Big Data, incluyendo los objetivos de negocio de la empresa y las diferentes regulaciones que lo afectan. Sin embargo, el conjunto de características y necesidades inherentes de un proyecto de creación de un entorno Big Data hace que sea complicado abarcar todos los aspectos posibles a considerar mediante el uso de metodologías tradicionales de definición de requisitos [8].

En la literatura, no existen muchos marcos o metodologías que traten el tema de la definición de requisitos en contextos Big Data. Diferentes autores proponen modelar requisitos utilizando un enfoque basado en *goals* del sistema. En [9], los autores exponen cómo usan los métodos GORE (Goal-Oriented Requirement Engineering), i* y KAOS (Knowledge Acquisition in autOmated Specification) para guiar la definición de requisitos. Si bien, toman en cuenta las características típicas de Big Data, no contemplan el contexto en el que se encuentra el entorno. Por otro lado la propuesta [10], también se encuentra basada en el método KAOS, sin embargo en este caso, los autores sí añaden una capa extra a su método que considera los objetivos de negocio de la compañía. El problema de esta propuesta es que se encuentra muy enfocada a la tecnología Apache Spark, por lo tanto, no es una solución genérica que pueda ser aplicada en cualquier proyecto Big Data. En cambio, en [11] los autores basan la definición de requisitos en los escenarios típicos de un entorno Big Data, es un enfoque interesante pero que se encuentra en un estado incompleto, ya que, solo contempla los requisitos relacionados con las fases de recogida y captura de datos. Finalmente en [12] se centran en cómo especificar requisitos de calidad en Big Data, para ello, se basan es las Vs básicas de Big Data en relación con características de calidad. Se trata de una propuesta interesante, pero que todavía se encuentra en una fase inmadura.

En nuestro caso, estamos trabajando en la creación de un marco de definición de requisitos que contemple por un lado las características básicas del Big Data (las diferentes Vs) y por otro el contexto en el que se va a encontrar el sistema (incluyendo los objetivos de negocio, sus políticas y las regulaciones legales que le pueden afectar). Para ello, nos basamos en los marcos básicos de ingeniería de requisitos, adaptándolos al nuevo escenario de esta tecnología. Por ello, este artículo se organiza en varias secciones: un apartado sobre el contexto que consideramos para nuestra propuesta, en el cual, se explica la arquitectura para Big Data que propone el National Institute of Standards and Technologies (NIST). A continuación, se define nuestra propuesta de marco para la definición en Big Data basada en sus dimensiones. Finalmente, se incluye una sección de conclusiones y trabajo futuro.

2 Contexto

En los últimos años, el NIST ha definido una arquitectura de referencia para Big Data que cuenta con el consenso general de la industria y la comunidad científica [13]. La última versión del manuscrito fue lanzada en agosto de 2017. La arquitectura aúna diferentes ideas y características sobre cómo crear un ecosistema Big Data, las cuales, han sido extraídas de diferentes propuestas realizadas por organizaciones del sector como Oracle o IBM. Como resultado, se ha obtenido la arquitectura que se muestra en la Figura 1. La arquitectura se divide en cinco componentes que interaccionan entre sí y tienen diferentes objetivos.

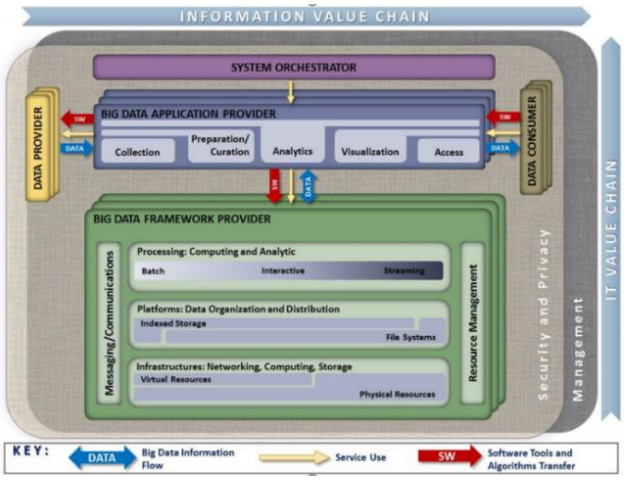

Figura 1. Arquitectura de referencia propuesta por el NIST

- **System Orchestrator (SO):** en relación con el tema de los requisitos, es el componente más importante, puesto que es el encargado de definir e integrar las actividades requeridas en el resto del ecosistema Big Data. Su principal objetivo es la configuración y gestión del resto de componentes.
- **Data Provider (DP):** este componente se encarga de alimentar al Big Data con nuevos datos. Para conseguirlo, dispone de una serie de interfaces que actúan como puerta entre el mundo exterior y el sistema Big Data.
- **Big Data Application Provider (BDAP):** el BDAP ejecuta una serie de servicios específicos siguiendo el ciclo de vida de los datos para cumplir con los requisitos establecidos en el SO. Su propósito principal es encapsular la lógica del negocio y las funcionalidades para ser ejecutados por la arquitectura.
- **Big Data Framework Provider (BDFP):** se puede considerar como la implementación de la lógica del sistema Big Data. Soporta las actividades definidas en el BDAP. En general, las implementaciones de Big Data son híbridos que combinan múltiples tecnologías.
- **Data Consumer (DC):** es similar al DP. Normalmente, el actor que interactúa con el componente es un usuario final y otro sistema. Similarmente al DP, está

compuesta por un conjunto de interfaces entre el usuario final y la información.

Tomando como base esta arquitectura de referencia, hemos realizado una ampliación sobre la misma, específicamente sobre el SO, para hacer un mayor hincapié en los requisitos. En la siguiente sección se define una primera aproximación a nuestra propuesta de marco de definición de requisitos para Big Data.

3 Marco de definición de requisitos para Big Data

Nuestra propuesta de marco de definición de requisitos para Big Data toma como primera consideración el marco típico de definición de requisitos definido en [14]. Este marco tiene tres componentes principales: el contexto del sistema, las actividades principales (documentación, elicitación y negociación) y los artefactos de requisito (*goals*, escenarios y requisitos orientados a la solución). Esta estructura general, la adaptaremos para que considere las características de Big Data.

Tal y como se comentó en la sección anterior, se ha realizado un ajuste del SO propuesto por NIST para que considere los requisitos como la parte central sobre la que gira la construcción del sistema Big Data. En la Figura 2 se muestra el diagrama de SO que proponemos [15]. Debido a las características de este componente, las actividades de seguridad se centran en los requisitos y cómo implementarlos y monitorizarlos. Dichos requisitos deben cumplir los *goals* del Big Data y estar alineados con los objetivos de negocio y las políticas de la compañía.

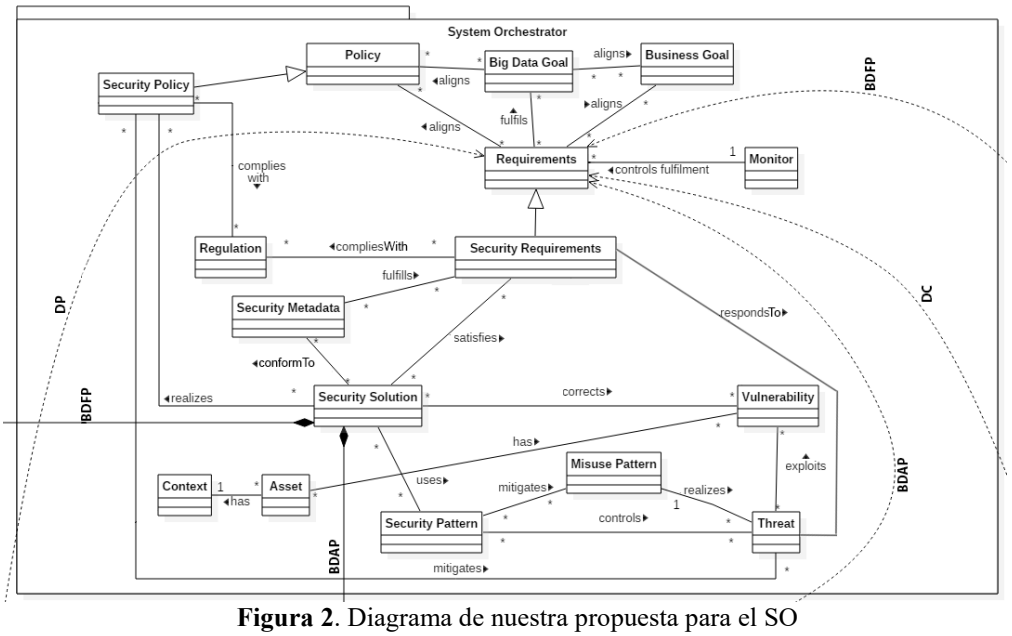

Figura 2. Diagrama de nuestra propuesta para el SO

Por otro lado, es necesario tener en cuenta el entorno ágil actual en el que se encuentran una gran cantidad de compañías relacionadas con Big Data. Para ello, definimos una iteración inicial en el que se define un escenario básico de alto nivel a partir del *goal* principal que se busca con el Big Data. Una vez creada dicha iteración inicial se procede a realizar una serie de iteraciones que siguen el mismo esquema. Del *goal* inicial se pasa a tener una serie de *subgoals* más específicos, los cuales, se pueden representar mediante un diagrama de grafos AND-OR de *goals* en el que se especifiquen las relaciones entre los mismos. Además, para la definición de requisitos en Big Data es necesario tomar en cuenta el contexto organizacional en el que se encuentre, incluyendo los objetivos de negocio y sus políticas. Por otro lado, también es importante tener en consideración las diferentes regulaciones legales que puede afectar a cómo son tratados los datos y la información que se genera. Al final de cada iteración se obtiene un conjunto de artefactos de requisitos que actuarán como entrada en la siguiente iteración. Cuando se considere que los requisitos se encuentran lo suficientemente detallados, se dará por finalizado el proceso. En la Figura 3 se muestra el esquema general nuestra propuesta.

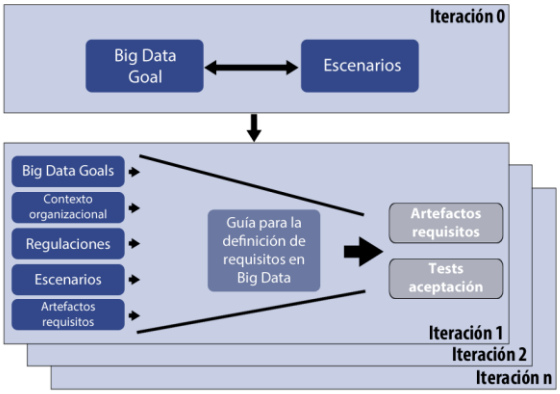

Figura 3. Esquema general de nuestra arquitectura

La guía de definición de requisitos se refiere a un artefacto que ayuda en el descubrimiento y clasificación de requisitos. Para ello, se tienen en cuenta las Vs básicas de Big Data: volumen, velocidad, variedad, valor y veracidad. Por otro lado, también se debe valorar una serie de características inherentes a los entornos Big Data, como son la analítica, seguridad y privacidad, funcionalidades o el hardware. Estas características deberán modificarse en función del contexto en el que vaya a desarrollar el sistema. La Figura 4 expresa una aproximación a cómo se realizaría esta guía con las diferentes dimensiones que considera. En las intersecciones que se producen es donde se deberán definir los diferentes artefactos de requisitos orientados a solución. Estos artefactos no tienen que ser definidos solo mediante lenguaje natural, sino que en iteraciones más avanzadas pueden ser diagramas o modelos.

Figura 4. Guía para la definición de requisitos en Big Data

A modo de ejemplo, en la Tabla I se incluye un caso de uso de la guía, en el cual, se simula la necesidad de tener que definir los requisitos para un Big Data que tiene como *goal* principal "Predecir conflictos raciales en los distintos barrios de la ciudad". La tabla que se muestra pertenecería a una iteración intermedia, en la cual, ya se encuentran definidas las fuentes de datos que se van a utilizar y los diferentes *subgoals* del sistema. Como se puede observar, no es necesario completar todas las intersecciones; su objetivo no es ese, sino apoyar en la definición de requisitos sin olvidar aquellos relacionados con Big Data.

Tabla I. Ejemplo de uso de la guía de definición de requisitos

	Volumen	**Velocidad**	**Variedad**	**Valor**	**Veracidad**
Analítico	Tener una mínima cantidad de datos	Los datos se tienen que analizar en tiempo real	Integración de las diferentes fuentes de datos	Obtener porcentaje de probabilidad de ocurrencia de conflicto	
Privacidad y seguridad	Incapacidad de inferir datos personales		Proteger privacidad de datos personales en fuentes de datos	Control de acceso a los resultados obtenidos	Comprobar la autenticidad de las fuentes de datos
Funciona-lidad		Observar mapa interactivo con zonas con posible conflicto	Uso de diferentes fuentes de datos		
Hardware	Servidores suficientes para tratar con dicha cantidad de datos	Servidores suficientemente potentes para tratar el tiempo real			

Finalmente, como resultado final de las iteraciones se obtiene una serie de artefactos de requisitos orientados a la solución como modelos de las fuentes de datos, modelo de integración de las fuentes de datos, diferentes diagramas UML (como diagramas de escenarios y de comportamiento), el modelo de proceso ETL, y los tests

de aceptación. El proceso de definición de requisitos no concluye aquí, sino que normalmente en contextos de Big Data y debido a la heterogeneidad de las fuentes de datos de que se dispone, es necesario abordar una fase en la que se integran los distintos modelos de las fuentes de datos.

3.1 Requisitos de seguridad

Tal y como se muestra en la Figura 2, en nuestro marco los requisitos de seguridad pueden ser satisfechos mediante el uso diferentes soluciones de seguridad que siguen las políticas de seguridad y tienen el principal objetivo de abordar las amenazas y reducir vulnerabilidades. Estas soluciones de seguridad serán implementadas en otros componentes del sistema Big Data como el BDAP o el BDFP. Sin embargo, estas soluciones de seguridad no son fáciles de implementar; por ello, nuestro modelo usa patrones de seguridad como guía. Un patrón de seguridad es una solución a un problema recurrente que indica cómo defenderse de una amenaza, o un conjunto de amenazas de un modo conciso y reusable [16]. Por ello, la definición de requisitos de seguridad en nuestra propuesta sigue un esquema similar al visto, pero debe tener un trato especial, ya que, se encuentra condicionado por una serie de legislaciones, normativas, amenazas y vulnerabilidades.

4 Conclusiones y trabajo futuro

En este artículo se presenta una primera aproximación a la creación de un marco de definición de requisitos para entornos Big Data. Nuestra propuesta se basa en metodologías y técnicas de definición de requisitos ampliamente aceptadas por la comunidad científica, además de la arquitectura Big Data propuesta por el NIST. Para ello, hemos tomado en cuenta nociones típicas de entornos ágiles como las iteraciones. Nuestro marco hace énfasis en el entorno que rodea al Big Data que se desea construir incluyendo las necesidades de la organización y las regulaciones legales que lo pueden afectar. Por otro lado, a la hora de definir requisitos es importante considerar las propiedades Vs típicas de Big Data además de las características inherentes del mismo como la seguridad y privacidad o la analítica. Para ayudar en el descubrimiento y definición de requisitos hemos generado una guía en la que estas propiedades se toman en consideración. Nuestra propuesta hace un especial hincapié en los requisitos de seguridad y privacidad debido a la importancia que tienen en un entorno de este tipo.

Este trabajo surge como respuesta a la necesidad sobre cómo definir de forma efectiva requisitos para el contexto Big Data. Aun así se encuentra todavía en una fase preliminar, con mucho trabajo futuro por delante incluyendo la definición formal y refinamiento tanto del proceso, como de los modelos que se obtienen como resultado de utilizar nuestro marco o la realización de casos de estudio para probar su eficacia.

Agradecimientos. Este trabajo ha sido financiado por el proyecto SEQUOIA (Ministerio de Economía y Competitividad y el Fondo Europeo de Desarrollo Regional FEDER, TIN2015-63502-C3-1-R).

Referencias

1. Akoka, J., Comyn-Wattiau, I., Laoufi, N.: Research on Big Data – A systematic mapping study. Computer Standards & Interfaces. 54, 105–115 (2017).
2. Mayer-Schönberger, V., Cukier, K.: Big Data: A Revolution that Will Transform how We Live, Work, and Think. Houghton Mifflin Harcourt (2013).
3. Chen, M., Mao, S., Liu, Y.: Big data: A survey. Mobile Networks and Applications. 19, 171–209 (2014).
4. S. Sagiroglu, D. Sinanc: Big data: A review. Collaboration Technologies and Systems (CTS), 2013 International Conference on. 42–47 (2013).
5. Ali-ud-din Khan, M., Uddin, M.F., Gupta, N.: Seven V's of Big Data understanding Big Data to extract value. In: American Society for Engineering Education (ASEE Zone 1), 2014 Zone 1 Conference of the. pp. 1–5. IEEE (2014).
6. Sharma, P.P., Navdeti, C.P.: Securing big data hadoop: a review of security issues, threats and solution. Int. J. Comput. Sci. Inf. Technol. 5, (2014).
7. Kelly, J., and J. Kaskade. "CIOS & BIG DATA what your IT team wants you to know." DOI= http://blog.infochimps.com/2013/01/24/cios-big-data (2013).
8. L. Liu: Security and Privacy Requirements Engineering Revisited in the Big Data Era. In: 2016 IEEE 24th International Requirements Engineering Conference Workshops (REW). pp. 55–55 (2016).
9. H. Eridaputra, B. Hendradjaya, W. Danar Sunindyo: Modeling the requirements for big data application using goal oriented approach. In: 2014 International Conference on Data and Software Engineering (ICODSE). pp. 1–6 (2014).
10. G. Park, L. Chung, L. Zhao, S. Supakkul: A Goal-Oriented Big Data Analytics Framework for Aligning with Business. In: 2017 IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService). pp. 31–40 (2017).
11. Al-Najran, N., Dahanayake, A.: A Requirements Specification Framework for Big Data Collection and Capture. In: Morzy, T., Valduriez, P., and Bellatreche, L. (eds.) New Trends in Databases and Information Systems: ADBIS 2015 Short Papers and Workshops, BigDap, DCSA, GID, MEBIS, OAIS, SW4CH, WISARD, Poitiers, France, September 8-11, 2015. Proceedings. pp. 12–19. Springer International Publishing, Cham (2015).
12. Noorwali, I., Arruda, D., Madhavji, N.H.: Understanding quality requirements in the context of big data systems. In: Proceedings of the 2nd International Workshop on BIG Data Software Engineering. pp. 76–79. ACM, Austin, Texas (2016).
13. NBD-WG, NIST: NIST Big Data Reference Architecture, https://bigdatawg.nist.gov/_uploadfiles/M0639_v1_9796711131.docx.
14. Pohl, K.: Requirements Engineering: Fundamentals, Principles, and Techniques. Springer Publishing Company, Incorporated (2010).
15. Moreno, J., Serrano, M.A., Fernández-Medina, Fernandez, E.B.: Towards a Security Reference Architecture for Big Data. (2017). In: International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data. Vienna, Austria (2018).
16. Fernandez, E.B.: Security patterns in practice: designing secure architectures using software patterns. John Wiley & Sons (2013).