Revista Technology Inside by CPIC – ISSN: 2215-5392 Vol.4 N°4-Agosto-Diciembre, 2019:1-10

Big data y ciencia de diseño: oportunidades de investigación

Luis Naranjo-Zeledón
San José, Costa Rica
luiscnaranjo@gmail.com

Fecha de recibido: 17 de noviembre de 2019

Fecha de aprobado: 19 de noviembre de 2019

Abstract—Este estudio explora las oportunidades de investigación que existen en big data y tecnologías relacionadas. Preguntas de investigación: (1) ¿Cuáles áreas de investigación están abiertas en big data? (2) ¿Cómo puede contribuir la ciencia de diseño a facilitar la investigación en estas áreas? El presente texto revisa literatura específica sobre big data y problemas abiertos, recurriendo a técnicas de búsqueda estricta, para descartar estudios que no aborden directamente el objeto de estudio. Este artículo aplica criterios de búsqueda excluyentes en repositorios académicos y hacer uso de la técnica de paper skimming, para luego seleccionar los artículos que debían ser leídos completos. Se concluye con la descripción de los problemas abiertos en big data, abordados desde distintas perspectivas. Resulta evidente que aún quedan problemas por resolver, lo cual puede motivar a estudiantes de posgrado a escoger esta línea de investigación. La metodología de investigación conocida como ciencia de diseño se ha usado en muy diversos hábitos, con muy buenos resultados y, sin duda, provee un marco de trabajo muy apropiado para investigar artefactos de big data dentro de contextos particulares, con procesos de validación adecuados.

Keywords—big data, problemas abiertos, ciencia de diseño, oportunidades de investigación.

Abstract — This study explores the research opportunities that exist in big data and related technologies. Research Questions: (1) What research areas are open in big data? (2) How can design science contribute to facilitating research in these areas? This text reviews specific literature on big data and open problems,

using strict search techniques to rule out studies that do not directly address the object of study. This article applies exclusive search criteria in academic repositories and makes use of the paper skimming technique, to then select the articles that should be read in full. It concludes with the description of open problems in big data, approached from different perspectives. It is evident that there are still problems to be solved, which can motivate graduate students to choose this line of research. The research methodology known as design science has been used in very diverse habits, with very good results and, without a doubt, it provides a very appropriate framework to investigate big data artifacts within particular contexts, with adequate validation processes.

Keywords — big data, open problems, design science, research opportunities.

I. INTRODUCCIÓN

Se suele atribuir a Roger Mougalas el concepto de big data, explicado en 2005, tal como se conoce en el contexto actual [1]. No obstante, ya lo había empleado anteriormente John Mashey para abordar el tema de la capacidad de almacenamiento y la rapidez de recuperación en discos [2].

Hoy en día, se encuentran definiciones de big data como la siguiente: “conjuntos de datos muy grandes producidos por personas que utilizan Internet, y que solo pueden almacenarse, comprenderse y utilizarse con la ayuda de herramientas y métodos especiales” [3]. Esta es otra definición habitual: “conjuntos de información que son demasiado grandes o demasiado complejos para manejar, analizar o usar con métodos estándar” [4].

Evidentemente, el tema de la complejidad asociada al tamaño, es recurrente en big data. Las áreas abiertas a investigación están asociadas al tema de la complejidad de los métodos y, en esta calidad, es que se sugiere en este artículo utilizar ciencia de diseño como metodología de investigación. La ciencia de diseño abarca tres ciclos: el de rigor, el de relevancia y el de diseño” [5]. La figura 1 muestra un esquema de estos ciclos, concebidos para construir artefactos y luego validarlos dentro de un contexto.

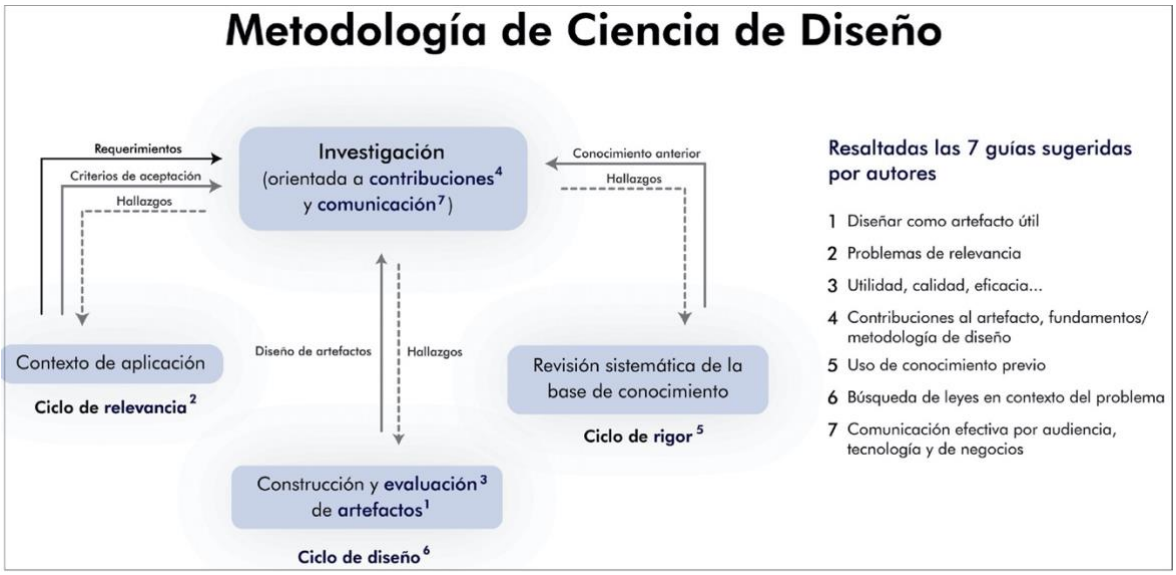

Fig. 1: Los tres ciclos de la ciencia de diseño. Tomado de [6].

El fenómeno de la complejidad, inherente al tratamiento de big data, es abordado de manera muy adecuada por la ciencia de diseño [7].

II. REVISIÓN DE LITERATURA

A. Big data y la ciencia de diseño

Algunos autores [8] se han cuestionado si el análisis de big data afecta la evaluación en investigación de ciencia de diseño (ICD) y, de ser así, si ello conduce a un nuevo género de ICD, concluyendo que el análisis de big data debe influir en la forma en que se evalúa, pero no conduce a la creación de un nuevo género de ICD.

La ciencia de diseño se ha usado para desarrollar y evaluar métodos novedosos de clasificación mediante la construcción de redes de movilidad laboral, utilizando big data de currículos en línea recopilados de redes sociales profesionales [9].

En Arabia Saudita se desarrolló una aplicación de salud móvil basada en big data, utilizando un marco de ciencia de diseño para prevenir y gestionar los problemas de salud relacionados con la peregrinación anual [10].

En un estudio se ha explicado la evolución del desarrollo de servicio en la nube durante la era de big data,

desde la perspectiva de la ciencia del diseño y sus tres ciclos. Se propuso un modelo de ecosistema de big data y las relaciones con el desarrollo de servicios en la nube [11].

También se ha abordado el tema de la sostenibilidad, con una relación de datos dinámica y compleja entre ecosistemas humanos y ambientales distribuidos geográficamente. Esto ha motivado el desarrollo de sistemas de información holísticos para gestionar diferentes ecosistemas dentro de un contexto de sostenibilidad, por medio de una interpretación y análisis empírico, con métodos cualitativos y cuantitativos. El enfoque de ciencia de diseño empleado tuvo como objetivo utilizar observaciones a partir de volúmenes de big data no estructurada [12].

B. Problemas abiertos en big data

En [13], los autores insisten en tener una caracterización de la calidad de los datos. Para explotar los datos web, estas caracterizaciones se materializan confiabilidad y procedencia. También señalan aspectos de la calidad de big data mediante ejemplos de calidad de datos de sensores.

En un estudio se insiste en una necesidad muy grande para aportar poder analítico a los datos, en lugar de sencillamente integrarlos en los sistemas de análisis ya existentes [14].

En cuanto a telecomunicaciones, también se han identificado varios problemas en el uso de big data [15], tales como zonas activas de tráfico espacio-temporal y secuencias de transferencia con problemas de rendimiento. En el caso de los sistemas de minería de datos industriales con palabras reales, la garantía de privacidad diferencial propicia una pérdida de precisión del 15% al 30%. La privacidad y la confidencialidad son críticas para la confiabilidad de las empresas de telecomunicaciones debido a lo sensible de los datos del usuario, como registros de facturación, números de llamadas e información de trayectoria.

En este sentido, los autores de ese estudio hacen ver que las empresas de telecomunicaciones están llegando a un punto en el cual recopilan más datos de los cuales podrían explotar. Esto introduce una carga financiera significativa sobre el operador, tomando en cuenta que el almacenamiento de datos en nubes públicas, donde existan economías de escala disponibles, no es una opción debido a razones de privacidad. Por otro lado, se impone un alto costo computacional para acceder y procesar los datos recopilados. La visión de almacenar infinitamente todos los datos por IoT gradualmente se volverá demasiado costoso y poco práctico.

Un artículo concentrado en los problemas de calidad de datos, agrupa en tres grandes categorías los problemas abiertos de big data [16]:

- Heterogeneidad: los sistemas de recomendación en Internet están generando datos de muy variada naturaleza, para lo cual se sugiere usar Pig, Hive y Mahout, que son parte de Hadoop y HDFS, como herramientas de análisis y gestión de datos precisos. Oozie y EMR con Flume y Zookeeper se pueden utilizar para manejar el volumen y la veracidad de los datos.

- Cuello de botella del algoritmo de minería: la mayoría de algoritmos de minería de datos para análisis de big data están diseñados para computación paralela, pero la mayoría de los algoritmos tradicionales de minería de datos no están diseñados para computación paralela. Por lo tanto, no son particularmente útiles para la minería de big data. Un algoritmo tradicional de minería de datos debe transformarse correspondientemente e implementarse en una plataforma de computación en la nube. Aunque los algoritmos tradicionales de minería de datos se pueden usar para analizar problemas de big data, hasta ahora, no hay mucha investigación respecto a este tema.

- Privacidad: los datos abiertos contienen mucha información personal y no se puede garantizar que esta no sea accesible para otras

personas y organizaciones. Incluso con datos de entrada anónimos, quizá el sistema pueda recuperar o inferir información personal de los resultados del análisis de big data. El análisis de datos con información confidencial puede traer muchos problemas. En cuanto a privacidad, se han propuesto enfoques como: el cifrado, el control de acceso, el anonimato, las transformaciones y la privacidad diferencial. Los autores del artículo hacen notas que el factor clave es como aplicar estos enfoques en el área de big data.

En el aspecto más cercano a las ciencias de la computación, [17] se estudian aspectos de pre-procesamiento en contextos imprecisos, tomando como base la teoría de conjuntos aproximados. En las últimas décadas, la cantidad de datos ha aumentado a un ritmo nunca antes visto, caracterizando los datos por su volumen, variedad, velocidad y veracidad. Con base en ello, se ha vuelto difícil adquirir rápidamente la información más útil a partir de big data. Es necesario, por lo tanto, realizar pre procesamiento de datos como un primer paso. A pesar de que existen técnicas para esta tarea, la mayoría de los métodos de vanguardia requieren información adicional como primer paso y no son capaces de lidiar con el aspecto de veracidad, ni con sus requisitos computacionales. Existen vacíos en investigación en esta área, que se han tratado de abordar principalmente en teoría de conjuntos aproximados y heurísticas de búsqueda aleatoria para optimización.

III. METODOLOGÍA

La metodología utilizada para constituir este artículo de revisión es la búsqueda directa en repositorios académicos, mediante la conformación de cadenas de búsqueda estrictas. Se ha usado el repositorio Google Scholar, primero para medir la cardinalidad del objeto de estudio y luego se ha determinado que los artículos candidatos son suficientes, sin necesidad de utilizar ténicas de backward snowballing o forward snowballing, tal como las describe Wohlin [18]. Se utilizaron las siguientes cadenas de búsqueda:

- intitle: “big data intitle:” design science”

- intitle: “big data intitle:” open problems”

Por tratarse de un artículo de revisión, no se han seguido todos los pasos de una revisión sistemática de literatura completa. En todo caso, para abordar problemas abiertos se ha restringido la búsqueda a artículos del año 2015 en adelante.

IV. RESULTADOS Y CONCLUSIONES

Las dos preguntas de investigación han sido contestadas a lo largo de este trabajo. Recapitulamos la primera pregunta se refiere a las categorías de problemas abiertos, finalmente es evidente que sobresalen los relacionados con la complejidad del manejo de big data, en cuanto a la calidad de los datos, la cantidad de datos generados y transmitidos a partir de IOT y sistemas de telecomunicaciones, así como el manejo de los temas asociados a la privacidad.

La segunda pregunta cuestiona cómo puede ser útil la ciencia de diseño para la investigación en estas áreas. En cuanto a esta escogencia por parte de los autores referenciados, algunos la justifican debido a su fuerte énfasis en evaluación del diseño dentro de un contexto, lo cual coincide con las fuertes necesidades de evaluación que impone el diseño de cualquier artefacto de big data.

Otros autores, no obstante, simplemente hacen referencia al énfasis en la conceptualización de artefactos, o bien al despliegue de soluciones a través de diferentes dominios. Estos dominios equivalen a los contextos, habituales en la terminología utilizada en la ciencia de diseño.

A pesar de lo anterior, los problemas abiertos no arrojaron resultados en los cuales se utilice la ciencia de diseño como posible metodología de base, para referencia de los investigadores. Esta situación, es precisamente la que ha motivado a escribir este artículo, como una manera de proponer a estudiantes de posgrado

un punto de partida en la búsqueda de temas innovadores y marcos metodológicos apropiados para su abordaje.

Se suman a las dos preguntas de investigación, a raíz de la revisión de la literatura una serie adicional de observaciones interesantes, como se procede a explicar a continuación. La revisión de la literatura disponible indica pocas fuentes arbitradas que hablen sobre el objeto de estudio; pero sí muestran, con mucha claridad, la clase de problemas abiertos que existen en big data.

Se consiguen identificar algunas propuestas de tratamiento de big data que ya toman la ciencia de diseño como marco de referencia para la investigación. Las fuentes consultadas están disponibles de manera abierta en lo relativo a big data abordado mediante la metodología de ciencia de diseño. Llama la atención que la segunda cadena de búsqueda recuperó artículos en repositorios de pago, a los cuales el autor tiene acceso.

La tendencia a que los problemas abiertos sean expuestos en repositorios de pago puede explicarse debido a la oportunidad financiera que representa para las casas editoriales el cobro de temas de gran actualidad. Desde un punto de vista formal, queda para otras investigaciones analizar esta situación, sin embargo, para efectos de este artículo parece corroborar que las búsquedas tal como se diseñaron para este estudio arrojaron los resultados esperados.

V. REFERENCES

[1] M. v. Rijmenam, “A short history of big data” 2015.

[2] J. R. Mashey, “Big data and the next wave of infras-tress” in Computer Science Division Seminar, University of California, Berkeley, 1997.

[3] “Significado de big data en el Diccionario Cambridge inglés,” 2019. [Online]. Available: https://dictionary.cambridge.org/es/diccionario/ ingles/big-data

[4] “Significado de big data en el diccionario Oxford inglés,” 2019.

[Online]. Available: https://www.oxfordlearnersdictionaries .com/us/ definition/english/big- data

[5] R. J. Wieringa, Design science methodology for information systems and software engineering. Springer, 2014.

[6] S. Robles-Sandoval, H. Va´squez-Carvajal, and L. Naranjo-Zeledón, “Adaptación de la metodología de ciencia de diseño en el desarrollo de luminarias.” [Online]. Available: https:// revistas.ulatina.ac.cr/index.php/ tecnologiavital/ article/view/252

[7] T. G. Gill and W. Murphy, “Task complexity and design science” in 9th Int. Conference on Education and Information Systems, Technologies and Applications EISTA, 2011, pp. 19–22.

[8] A. Elragal and M. Haddara, “Design science research: Evaluation in the lens of big data analytics” Systems, vol. 7, no. 2, p. 27, 2019.

[9] X. Xu, H. Qian, C. Ge, and Z. Lin, “Industry classification with online resume big data: A design science approach” Information & Management, p. 103182, 2019.

[10] I. Alharbi, B. Alyoubi, M. R. Hoque, and N. Almazmomi, “Big data based m-health application to prevent health hazards: a design science framework” Telemedicine and e-Health, vol. 25, no. 4, pp. 326–331, 2019.

[11] C.-H. Liu, S.-C. Chen, and P.-H. Hsieh, “How big data ecosystem changes cloud services: A design science perspective” Open Journal of Social Sciences, vol. 3, no. 07, p. 74, 2015.

[12] S. L. Nimmagadda, T. Reiners, and G. Burke, “Big data guided de-sign science information system (dsis) development for sustainability management and accounting” Procedia computer science, vol. 112, pp. 1871–1880, 2017.

[13] M. Scannapieco and L. Berti, “Quality of web data and quality of big data: Open problems,” in Data and Information Quality. Springer, 2016, pp. 421–449.

[14] A. Cuzzocrea, "Data warehousing and olap over big data: a survey of the state-of- the-art, open problems and future challenges," International Journal of Business Process Integration and Management, vol. 7, no. 4, pp. 372–377, 2015.

[15] C. Costa and D. Zeinalipour-Yazti, "Telco big data research and open problems" in 2019 IEEE 35th International Conference on Data Engi- neering (ICDE). IEEE, 2019, pp. 2056– 2059.

[16] P. Zhang, F. Xiong, J. Gao, and J. Wang, "Data quality in big data processing: Issues, solutions and open problems" in 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (Smart- World/ SCALCOM/UIC/ATC/CBDCom/IOP/SCI). IEEE, 2017, pp. 1–7.

[17] Z. C. Dagdia, "Optimized framework based on rough set theory for big data pre-processing in certain and imprecise contexts"–Marie Sklodowska-curie project: Open problems'," in Recent Trends in Knowl- edge Compilation, 2018.

[18] C. Wohlin, "Guidelines for snowballing in systematic literature studies and a replication in software engineering" in Proceedings of the 18th international conference on evaluation and assessment in software engineering. Citeseer, 2014, p. 38.