Análisis de características en proyectos de *big data*: revisión sistemática de literatura

Mariel Liliana Ojeda
marielojeda@frba.utn.edu.ar
https://orcid.org/0009-0003-3529-8994
Universidad Tecnológica Nacional, Argentina

Cinthia Vegega
cinthia.vegega@gmail.com
https://orcid.org/0000-0002-5382-7875
Universidad Tecnológica Nacional, Argentina

María F. Pollo Cattaneo
flo.pollo@gmail.com
https://orcid.org/0000-0003-4197-3880
Universidad Tecnológica Nacional, Argentina

Recibido: 4 de octubre del 2024 / Aceptado: 2 de noviembre del 2024
doi: https://doi.org/10.26439/interfases2024.n020.7457

RESUMEN. En el desarrollo de proyectos de *big data* se identifican diversas problemáticas que pueden deberse a distintos factores, como la baja calidad de los datos utilizados con anomalías que pueden afectar la precisión de los resultados o la falta de claridad en los objetivos comerciales. Esta situación puede provocar errores en el proceso de toma de decisiones, retrasos en las entregas y hasta la cancelación del proyecto. En este contexto, el presente trabajo surge de la necesidad de recopilar investigaciones previas con el fin de conocer la importancia de la aplicación de una metodología de trabajo en proyectos de *big data*. Se realiza con el objetivo de identificar los enfoques de las metodologías más utilizadas y analizar las características propias de cada una, así como las características comunes o transversales, que permiten la combinación, o adaptación, de distintas metodologías en un mismo proyecto. La generación de grandes volúmenes de datos provenientes de diferentes fuentes y formatos aumenta el desafío de verificar la calidad, ya que pueden presentar anomalías que afecten así la precisión de los resultados obtenidos.

PALABRAS CLAVE: metodología / tecnología *big data* / gestión de datos empresariales

ANALYSIS OF FEATURES IN BIG DATA PROJECTS: A SYSTEMATIC LITERATURE REVIEW

ABSTRACT. In the implementation of big data projects, several problems are identified that may be due to different factors, such as the low quality of the data used with anomalies that may affect the accuracy of the results or the lack of clarity in the business objectives. This situation can lead to errors in the decision making process, delays in deliveries and even the cancellation of the project. In this context, the present work arises from the need to compile previous research in order to know the importance of the application of a working methodology in big data projects. The objective is to identify the approaches of the most used methodologies and to analyze the characteristics of each one, as well as the common or transversal characteristics that allow the combination, or adaptation, of different methodologies in the same project. The generation of large volumes of data from different sources and formats increases the challenge of verifying quality, as they may present anomalies that affect the accuracy of the results obtained.

KEYWORDS: methodology / big data technology / enterprise data management

INTRODUCCIÓN

Uno de los impactos de la globalización es la integración de los mercados. Un mercado globalizado exige a las organizaciones poseer una amplia visión del entorno en el cual actúa, con el afán de poder anticipar las oportunidades y amenazas que pueden emerger (Zúñiga et al., 2023). Como indica Dai et al. (2019), el análisis de la demanda del mercado, junto con los requisitos del cliente, puede ser utilizado para mejorar el diseño y la calidad de los productos.

Constantemente, los usuarios consumidores de bienes y servicios dejan registro de datos personales, financieros, entre otros, en distintas fuentes de almacenamiento. A lo largo de los últimos años, se hizo foco en los datos, los cuales eran considerados como un recurso estratégico de las organizaciones, ya que su proceso de tratamiento ha ayudado a organizaciones públicas y privadas a obtener conocimiento de gran valor (Krasteva & Ilieva, 2021).

Asimismo, el crecimiento continuo de los datos intensifica el desarrollo de proyectos de *big data* (Abdul Hamid et al., 2021), a partir de lo que comienzan a obtenerse beneficios, pero también se identifican nuevos desafíos (Shu et al., 2020). La generación de grandes volúmenes de datos provenientes de diferentes fuentes y formatos aumenta el desafío de verificar la calidad, ya que pueden presentar anomalías que afecten así la precisión de los resultados obtenidos (Caffetti et al., 2023).

En este contexto, Thomas H. Davenport (2006) describe cómo las empresas pueden obtener una ventaja competitiva con el análisis de datos de manera eficaz, dado que (son un insumo básico y clave en la economía del conocimiento; aunque sin el correspondiente refinamiento, procesamiento y análisis, el dato por sí solo no es generador de valor (Ontiveros et al, 2018). Luego del análisis de los datos, se puede extraer información que permite predecir próximos eventos, ahí es cuando comienzan a jugar un rol importante (Dai et al., 2019).

Dado que el crecimiento del volumen, la variedad y la velocidad de la generación de datos requiere de avances tecnológicos que acompañen una mayor capacidad de captura, almacenamiento, procesamiento y análisis, surge *big data* como concepto para el tratamiento de grandes volúmenes. Esto incluye el conjunto de herramientas y técnicas destinadas a extraer todo el valor de los datos para enriquecer y complementar sistemas con capacidades predictivas (Manzano & Avalos, 2023).

Por ello, un ejemplo de implementación de *big data* son las empresas que se dedican al comercio electrónico, ya que utilizan la gran variedad de datos recogidos de sus clientes (Tardío et al., 2020). Eso se transforma en un nuevo activo que alimenta la economía de la información (Manzano & Avalos, 2023).

Considerando que un proyecto se define como un esfuerzo temporal que se lleva a cabo para crear un producto, servicio o resultado único, que tiene un comienzo y un

final definidos, así como un propósito claro, que generalmente busca cumplir con ciertos objetivos o resolver un problema específico (Project Management Institute [PMI], 2017), es necesario contar con un marco de trabajo que dé soporte a la estandarización de procesos, con el fin de minimizar riesgos, mejorar la coordinación y garantizar que se cumplan los objetivos del proyecto a partir del uso de una metodología. Por definición de PMI (2017), una metodología es un sistema de prácticas, técnicas, procedimientos y reglas utilizado por quienes trabajan en una disciplina.

No obstante, la gestión de este tipo de proyectos puede presentar dificultades en su implementación, las cuales generan complicaciones que terminan impidiendo que un proyecto pueda alcanzar a completarse, o bien que el mismo se termine cancelando. Reggio y Astesiano (2020) presentan un estudio de investigación que estima que entre el 60 y 85 % de los proyectos no finalizan exitosamente, sus principales causas son las siguientes: inicio de proyecto con falta de claridad en los objetivos comerciales (o bien puede ocurrir que distintos actores tengan en cuenta requisitos diferentes), falta de experiencia o de conocimiento técnico del equipo e incorrecta identificación del problema.

En los últimos años, se ha trabajado en aproximaciones metodológicas con el objetivo de proporcionar soluciones efectivas. Aunque la mayoría de estas propuestas se basan en el análisis de los requisitos derivados de las diez características de *big data* (las diez *v*) indicadas por Khan et al. (2018), se considera imprescindible describir un escenario como factor clave para la elección de las técnicas y herramientas más adecuadas. Sin embargo, a modo de ejemplo para los casos donde un proyecto se oriente al desarrollo de la arquitectura, de acuerdo con lo mencionado por Tardío et al. (2020), las metodologías existentes no guían el desarrollo con suficiente detalle o no son aplicables.

Por lo tanto, la implementación de una metodología de gestión permite organizar mejor un proyecto, obtener mejores resultados del *software* entregado al cliente y evitar los fracasos (Bahit, 2012). La misma se define de acuerdo con el enfoque del mismo y la problemática a tratar, y se considera que cada metodología cuenta con características propias y características que las asemejan entre sí. En este contexto, el objetivo de este trabajo es realizar una revisión sistemática de la literatura (RSL) sobre las metodologías aplicadas en proyectos de *big data*.

Por esta razón, resulta necesario mencionar la metodología Cross Industry Standard Process for Data Mining (CRISP-DM), por ejemplo, la cual consta de seis fases iterativas, donde según sea necesario, el equipo puede "regresar" a una fase anterior (Saltz & Hotz, 2020). Estas fases configuran el problema empresarial (comprensión del negocio), revisan los datos disponibles (comprensión de datos), desarrollan modelos analíticos (preparación y modelado de datos), evalúan los resultados frente a las necesidades empresariales (evaluación) e implementan el modelo (implementación). Todo el

ciclo está diseñado para ser iterativo y repetirse según sea necesario para mantener los modelos actualizados y eficaces (Ahmad et al., 2022).

Otro marco metodológico a mencionar es Sample, Explore, Modify, Model, and Assess (SEMMA), el cual consta de cinco fases (Saltz & Hotz, 2020). Las características principales de este marco son la extracción de datos para muestreo aleatorio y la exploración de tendencias de datos. Se inicia con una muestra de datos estadísticamente representativa que utiliza estrategias de muestreo. El muestreo es un método utilizado para seleccionar un subconjunto de un grupo que permita obtener conclusiones estadísticas y aproximar las características de toda la población (Ahmad et al., 2022).

El presente trabajo se organiza de la siguiente manera: la primera sección describe la metodología utilizada y el proceso de revisión de literatura, mientras que la segunda sección muestra la interpretación de los resultados. Finalmente, en la tercera sección, se presentan las conclusiones y futuras líneas de trabajo.

1. METODOLOGÍA

De acuerdo con la definición de Kitchenham y Charters (2007), una revisión sistemática de la literatura (RSL) es un medio para identificar, evaluar e interpretar toda la investigación disponible que sea relevante para una pregunta de investigación, un área temática o un fenómeno de interés en particular.

En esta RSL, se siguen las instrucciones y recomendaciones de Kitchenham y Charters, las cuales aplican un método basado en sus aportes, con el fin de evaluar e interpretar el trabajo de investigadores, académicos y profesionales en el campo elegido. El objetivo es la búsqueda y hallazgo de artículos científicos que desarrollen la implementación de metodologías en proyectos de *big data* que permitan la generación de propuestas de futuras líneas de investigación.

A continuación, se describe el protocolo de revisión a utilizar:

1.1. Definición de preguntas de investigación

1.2. Definición de fuentes de datos

1.3. Establecimiento de cadenas de búsqueda

1.4. Ejecución de consultas

1.5. Proceso de selección

1.1. DEFINICIÓN DE PREGUNTAS DE INVESTIGACIÓN

Dado que el eje principal de esta investigación es conocer la importancia de la aplicación de una metodología de trabajo en proyectos de *big data*, se plantea una serie de interrogantes para dar curso al desarrollo de la presente RSL.

En primer lugar, se definen en la Tabla 1 las preguntas de investigación, junto a la motivación de cada una de ellas, las cuales van a dar lugar al desarrollo del trabajo actual. La definición y establecimiento de las preguntas de investigación orienta la búsqueda y posterior análisis de información, ayudando a cumplir con el objetivo del estudio. Se considera necesaria la RSL buscando dar respuesta a las siguientes preguntas:

Tabla 1

Preguntas de investigación

Ref	Preguntas	Motivación
PI1	¿Cuál es el foco de la metodología utilizada aplicada a proyectos de *big data*?	Analizar cómo las metodologías abordan las necesidades de los proyectos.
PI2	¿Cuáles son las características propias de cada una de las metodologías aplicadas a proyectos de *big data*?	Identificar las características únicas de cada metodología para evaluar las fortalezas de las mismas.
PI3	¿Cuáles son las características comunes entre las distintas metodologías aplicadas a proyectos de *big data*?	Determinar si una metodología puede combinarse en un nuevo aspecto metodológico.

1.2. Definición de fuentes de datos

El segundo paso es definir las fuentes de datos que van a ser consultadas. Se decide realizar una búsqueda automática en los repositorios descritos en la Tabla 2 con el fin de recuperar estudios alineados al propósito del presente estudio.

Tabla 2

Fuentes de datos

Bibliotecas/repositorios	Opciones
Biblioteca digital de ACM	Artículos de congresos, artículos de revistas
IEEE Xplore	Artículos de congresos, artículos de revistas
arXiv	Artículos de congresos, artículos de revistas
Sistema Nacional de Repositorios Digitales	Artículos de congresos, artículos de revistas

1.3. Establecimiento de cadenas de búsqueda

Con el fin de recuperar estudios relevantes que respondan las preguntas de investigación planteadas inicialmente, se definen los términos para realizar las búsquedas, en idioma inglés y español: “metodología”, “metodológica”, “método”, “methodology”, “methodologies”, “methodological”, “method”, “big data”, “data science” y “análisis de datos”. Para las búsquedas se utilizan los términos aplicados en los filtros: título, resumen y palabras clave.

Se construye una cadena de búsqueda que incluye el conjunto de términos indicados anteriormente y se utilizan los operadores booleanos AND y OR para relacionarlos. Por

consiguiente, para la recopilación de los diferentes artículos científicos, se emplea la siguiente cadena de búsqueda:

("Título": metodología OR metodológica OR método OR methodology OR methodologies OR methodological OR method) AND ("Título": "big data" OR "data science" OR "análisis de datos") OR ("Resumen": metodología OR metodológica OR método OR methodology OR methodologies OR methodological OR method) AND ("Resumen": "big data" OR "data science" OR "análisis de datos") AND ("Palabras clave": metodología OR metodológica OR método OR methodology OR methodologies OR methodological OR method) AND ("Palabras clave": "big data" OR "data science" OR "análisis de datos")

1.4. Ejecución de consultas

Con el objetivo de realizar la selección de estudios más representativos en forma adecuada, se definen los criterios de inclusión y exclusión.

Las variables utilizadas son el periodo de tiempo de publicación de artículos, el idioma, la duplicidad de hallazgos y que apliquen con la cadena establecida. Los criterios definidos se encuentran listados a continuación en la Tabla 3.

Tabla 3

Preguntas de investigación

Criterios de inclusión	Criterios de exclusión
Artículos publicados en el periodo: 1/1/2020-30/4/2024	Artículos que no cumplan con el periodo de publicación establecido como criterio de inclusión
Idioma: español o inglés	Idioma: distintos del español o inglés
Artículos que contengan las cadenas definidas en el título, palabras clave o en el resumen	Duplicidad de hallazgo entre repositorios

La primera ejecución se realiza en los cuatro repositorios con los términos de la cadena en ambos idiomas, español e inglés. En la Tabla 4, se visualiza el resultado de la ejecución:

Tabla 4

Resultados de la primera ejecución

Repositorio	Cantidad de resultados
IEEE	802
ACM	219
Sistema Nacional de Repositorios Digitales	12
arXiv	25
Total	1058

Para la segunda etapa, se considera una serie de revisiones desde lo general a lo particular, con el fin de depurar los resultados, que definen este proceso de selección de estudios en cinco fases descritas en el siguiente punto.

1.5. Proceso de selección de estudios

Como parte del proceso de selección de estudios, se define una serie de fases que van a actuar como filtros para obtener el conjunto de estudios primarios.

Las fases se describen a continuación:

- Fase 1: Revisión de artículos con la aplicación de la cadena de búsqueda en los gestores bibliográficos
- Fase 2: Eliminación de estudios duplicados para evitar repetición de información
- Fase 3: Lectura del título, las palabras clave y el resumen del artículo para evaluar su relevancia
- Fase 4: Lectura de la introducción, los resultados y las conclusiones del artículo para obtener una comprensión más profunda del estudio
- Fase 5: Lectura completa del artículo para asegurar su relevancia para la investigación

Al momento de ejecutar la quinta fase, la cantidad de resultados para iniciar la lectura completa de los artículos se visualiza en la Tabla 5.

En esta instancia, la lectura determina el grado de relación del artículo con el tema objeto de la investigación. En esta etapa son descartados 15 artículos por no ser relevantes para las preguntas de investigación planteadas, por lo que finalmente 14 artículos forman parte del resultado de esta investigación.

Tabla 5

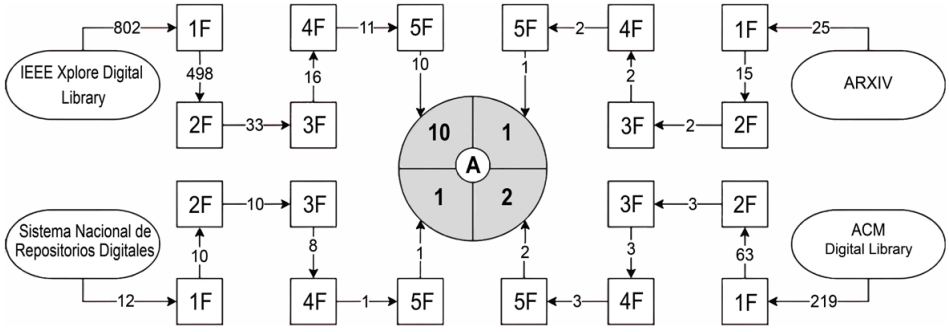
Cadena definitiva por repositorio

Repositorio	Periodo	Cantidad de resultados
IEEE	2020-2024	16
ACM	2020-2024	3
Sistema Nacional de Repositorios Digitales	2020-2024	8
arXiv	2020-2024	2
Total		29

El proceso de selección de estudios primarios se visualiza en la Figura 1, donde se reflejan las fases de filtros aplicados y se especifica en cada instancia la cantidad de estudios resultantes. Al finalizar la lectura se identificará el conjunto de artículos primarios que darán respuesta a las preguntas de investigación planteadas inicialmente.

Figura 1

Selección de estudios primarios



En la Tabla 6, se muestran los estudios primarios junto con el repositorio desde el cual se obtuvieron.

Tabla 6

Detalle de estudios primarios

Artículo	Repositorio
Jin, W., Yang, J., & Fang, Y. (2020). Application methodology of big data for emergency management. En *2020 IEEE 11th International Conference on Software Engineering and Service Science (ICSESS)* (pp. 326-330). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/ICSESS49938.2020.9237653	IEEE Xplore
Abdul Hamid, K., Abu Bakar, M., Jalar, A., & Hakim Badarisman, A. (2021). Incorporation of big data in methodology of identifying corrosion factors in the semiconductor package. En *2021 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)* (pp. 1-4). https://doi.org/10.1109/ICECCE52056.2021.9514240	IEEE Xplore
Kavakli, E., Sakellariou, R., Eleftheriou, I., & Mascolo, J. (2020). Towards a multi-perspective methodology for big data requirements. En *2020 IEEE International Conference on Big Data* (pp. 5719-5720). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/BigData50022.2020.9378406	IEEE Xplore
Song, X., Zhang, H., Akerkar, R., Huang, H., Guo, S., Zhong, L., Ji, Y., Opdahl, A. L., Purohit, H., Skupin, A., Pottathil, A., & Culotta, A. (2020). Big data and emergency management: concepts, methodologies, and applications. *IEEE Transactions on Big Data*, *8*(2), 397-419. https://doi.org/10.1109/TBDATA.2020.2972871	IEEE Xplore
Tardío, R., Maté, A., & Trujillo, J. (2020). An iterative methodology for defining big data analytics architectures. *IEEE Access*, *8*, 210597-210616. https://doi.org/10.1109/ACCESS.2020.3039455	IEEE Xplore

(continúa)

(continuación)

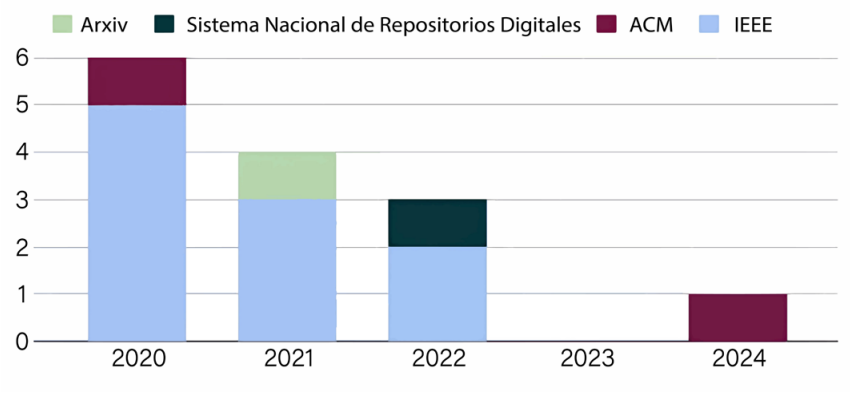
Artículo	Repositorio
Saltz, J., & Hotz, N. (2020). Identifying the most common frameworks data science teams use to structure and coordinate their projects. En *2020 IEEE International Conference on Big Data* (pp. 2038-2042). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/BigData50022.2020.9377813	IEEE Xplore
Dastgerdi, A., & Gandomani, T. (2021). On the appropriate methodologies for data science projects. En *2021 International Conference on Information Technology* (pp. 667-673). Institute of Electrical and Electronic Engineers. https://doi.org/10.1109/ICIT52682.2021.9491712	IEEE Xplore
Krasteva, I., & Ilieva, S. (2021). Adopting agile software development methodologies in big data projects – a systematic literature review of experience reports. En *2020 IEEE International Conference on Big Data* (pp. 2028-2033). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/BigData50022.2020.9378118	IEEE Xplore
Funde, S., & Swain, G. (2022). Big data privacy and security using abundant data recovery techniques and data obliviousness methodologies. *IEEE Access*, *10*, 105458-205484. https://doi.org/10.1109/ACCESS.2022.3211304	IEEE Xplore
Ahmad, Z., Yaacob, S., Ibrahim, R., & Farahwani, W. (2022). The review for visual analytics methodology. En *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications* (pp. 1-10). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/HORA55278.2022.9800100	IEEE Xplore
Caffetti, Y. A., Eckert, K., Ruidías, H. J., & Vera Laceiras, M. S. (2023). Data cleansing en entornos big data: mapeo sistemático de la literatura. En S. Rodríguez, M. Giménez y M. A. Molina (Comps.), *XXVIII Congreso Argentino de Ciencias de la Computación – CACIC 2022* (pp. 75-79). Editorial de la Universidad Nacional de La Rioja. https://repositoriosdigitales.mincyt.gob.ar/vufind/Record/SEDICI_1d437c59c0d397280848f3cfd422df97	Sistema Nacional de Repositorios Digitales
García-Gil, D., García, S., Xiong, N., & Herrera, F. (2021). Smart data driven decision trees ensemble methodology for imbalanced big data. *Cognitive Computation*, *16*, 1572-1588. https://doi.org/10.48550/arXiv.2001.05759	arXiv
Shu, W., Sun, W., & Li, Y. (2020). The development trend of design methodology under the influence of artificial intelligence and big data. En I*CDLT '20: Proceedings of the 2020 4th International Conference on Deep Learning Technologies* (pp. 104-108). Association for Computing Machinery. https://doi.org/10.1145/3417188.3417214	ACM Digital Library
Markopoulos, D., Tsolakidis, A., Karanikolas, N., Marinagi, A., & Skourlas, C. (2024). Applying soft system methodology for a clearer understanding of the future intensive care units. En *PCI '23: Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics* (pp. 163-170). Association for Computing Machinery. https://doi.org/10.1145/3635059.3635084	ACM Digital Library

2. DISCUSIÓN DE RESULTADOS

En base a los datos extraídos, se realiza un análisis que permite dar cuenta de las publicaciones por año, a partir del cual es posible determinar que si bien las publicaciones alcanzan un pico en el año 2021 decrecen en los años posteriores. Ello ha generado una menor cantidad de investigaciones que evidencien la importancia de la aplicación de una metodología de trabajo en proyectos de *big data*. En la Figura 2, se muestra la cantidad de publicaciones realizadas por año, la cual considera el rango 2020-2024 incluido en los criterios de inclusión.

Figura 2

Artículos publicados por año



Como se describe en el protocolo aplicado, la ejecución de la RSL retorna un listado de artículos filtrados hasta la selección final, que conforma un conjunto de artículos relevantes para responder las preguntas de investigación planteadas como parte del presente trabajo.

Los estudios primarios se analizan teniendo en cuenta lo siguiente:

- El tipo de enfoque que busca resolver el problema que dio origen al proyecto.
- Las características comunes entre las distintas metodologías halladas.
- Las características propias de cada una de las metodologías aplicadas.

Para cada uno de los puntos mencionados anteriormente se define una categorización para la posterior clasificación de los artículos, que se presenta en la Tabla 7.

Tabla 7

Definición de categorización de artículos

Dimensión	Categoría
PI1/Tipo de enfoque	Los tipos de enfoque son los siguientes: análisis de problema, aplicación de metodologías, arquitectura, calidad de datos, gestión de emergencias, proceso de obtención de requisitos, reducción de riesgos, seguridad de los datos, toma de decisiones, uso pedagógico.
PI2/Características propias	Las características propias son las siguientes: análisis predictivo de eventos, combinación de métodos, esquema lógico, modelo conceptual, modelos de innovación, métodos de recuperación de datos, obtención de requisitos, proceso de entrega de valor, proceso de limpieza de datos, retroalimentación, toma de decisiones.
PI3/Características comunes	Las características comunes son las siguientes: análisis de datos, análisis del contexto, guía del proceso, metodología iterativa, perspectiva multidisciplinaria, resolución de problema complejo.

Se analizan los trabajos desarrollados en 14 artículos de investigación, los cuales fueron considerados como "estudios primarios". El resultado de la RSL muestra particularidades en cada metodología aplicada, que permiten su implementación cuando el proyecto cuenta con determinadas características.

La aplicación de metodologías se enfoca principalmente en mitigar los riesgos en los que se puede incurrir por no llevar adecuadamente el análisis de la problemática a resolver (García-Gil et al., 2021; Shu et al., 2020), por niveles bajos en la calidad de los datos (Caffetti et al., 2023; García-Gil et al., 2021). Por ende, ello puede generar errores en el proceso de toma de decisiones (Markopoulos et al., 2024; Tardío et al., 2020; Jin et al., 2020).

Los artículos resultantes muestran que una metodología no aplica a todos los casos, ya que el contexto en el cual se desarrolla el proyecto es determinante y cada metodología tiene características propias que hacen que se pueda aplicar con mayor éxito en un universo de casos y en otros no. Entre las principales características que se identifican en los artículos, se destacan la flexibilidad para permitir la combinación de diferentes métodos de trabajo (García-Gil et al., 2021; Kavakli et al., 2020) y el enfoque en dar soporte al proceso de toma de decisiones (Markopoulos et al., 2024; Song et al., 2020).

Se observan características propias entre las metodologías utilizadas, así como se identifican características comunes que permiten hallar la metodología que aplique mejor con los requerimientos del proyecto. Entre las características comunes destacadas, pueden mencionarse el análisis de datos (García-Gil et al., 2021; Krasteva & Ilieva, 2021), el trabajo desde una perspectiva multidisciplinaria (Shu et al., 2020; Song et al., 2020; Ahmad et al., 2022), y la implementación de un proceso de fases iterativas (Caffetti et al., 2023; Tardío et al., 2020; Jin et al., 2020).

A continuación, se contesta cada una de las preguntas de investigación:

PI1: ¿Cuál es el foco de la metodología utilizada aplicada a proyectos de big data?

Shu et al. (2020) realizan un análisis de la situación actual en el ámbito del diseño de innovación de productos. En ese sentido, plantean que la innovación y el desarrollo tradicional de nuevos productos requieren que las empresas realicen enormes inversiones de recursos humanos y materiales para hacer frente a cambios rápidos del mercado; sus dificultades de diseño son cada vez más difíciles. Ante ello, cualquier empresa debe enfrentar los riesgos en el proceso de desarrollo de productos, incluso para grandes empresas. Se establece la idea de diseño innovador impulsado por algoritmos para establecer un modelo diferente del diseño de innovación de productos tradicionales y promover el desarrollo de teorías y métodos de desarrollo de innovación basados en datos.

Por otra parte, Jin et al. (2020) basan su modelado en un análisis previo de la demanda real. Se sostiene que el desglose y claridad con que se deben presentar los elementos centrales que componen el objetivo del desarrollo permitirán determinar el modelado y, a partir de ahí, seleccionar el algoritmo a aplicar.

De la misma forma en que Jin et al. (2020) plantean el enfoque en el análisis del problema, Abdul Hamid et al. (2021) indican que la identificación de las probables causas raíz es la piedra angular de la resolución de problemas. Así, esta representa la base de la mejora continua donde la lección aprendida es una parte integral de la misma.

Continuando con el enfoque que se plantea de abordar adecuadamente las características de los requisitos tempranos, Kavakli et al. (2020) desarrollan la interacción entre las intenciones del negocio y la funcionalidad del sistema, con el objetivo de alinear las necesidades empresariales y los requisitos, tanto del usuario como del sistema, ya que luego se conceptualizan en objetivos fundamentales para conseguir valor.

Es necesario determinar el contexto en el cual se desarrolla el proyecto de Krasteva e Ilieva (2021), que proporciona información relevante para mitigar los casos que no permiten que un proyecto alcance a completarse.

Tardío et al. (2020) identifican la falta de conocimiento profesional en el uso de la tecnología *big data* como el principal problema actual que impide a las empresas su adopción con éxito. De este modo, presentan una guía para la selección de las herramientas y técnicas que posteriormente darán paso a la construcción de una correcta arquitectura del sistema.

La calidad de datos se ve referenciada también por Caffetti et al. (2023), donde se plantea la necesidad de una adecuada limpieza de los datos, ya que, a grandes volúmenes provenientes de diferentes fuentes y formatos, aumenta el desafío de verificar la calidad de los mismos, que pueden ser imprecisos, presentar anomalías o no ser adecuados

para el análisis o procesamiento, que afecten así la precisión de los resultados obtenidos. La ausencia de datos, valores ficticios o predeterminados, ruido, datos erróneos, datos inconsistentes, datos crípticos, claves primarias duplicadas, identificadores no únicos, campos multipropósito y violación de reglas comerciales son problemas que pueden ser contrarrestados por una correcta limpieza de datos, lo que permita garantizar las fases siguientes de análisis.

Otro enfoque identificado en los artículos encontrados es el de la seguridad con el fin de preservar la privacidad de la información que se está gestionando, ya que los piratas informáticos explotan las plataformas de *big data* para lanzar ataques contra las organizaciones. Frente a ello, Funde y Swain (2022) presentan diferentes técnicas para preservar la privacidad de los datos con enfoques criptográficos y no criptográficos.

Si bien se considera cada vez más a los datos como un recurso estratégico para la organización, con el fin de obtener una ventaja competitiva es necesario aprovecharlos mediante análisis (Saltz & Hotz, 2020), y se ha observado que existen desafíos importantes al tratar de aprovechar los datos estratégicamente.

El último enfoque identificado está relacionado con la dinámica de equipo, ya que, a medida que aumenta el volumen de datos, aumenta la cantidad de recursos que tienen que intervenir en un proyecto, de modo que se deja de trabajar individuos en forma aislada para volverse un equipo de trabajo (Dastgerdi & Gandomani, 2021). Es en este punto donde surge el riesgo de que el rendimiento del equipo no sea óptimo. Por lo tanto, la cuestión principal es cómo se puede garantizar que el equipo trabaje de forma eficiente y eficaz.

PI2: ¿Cuáles son las características propias de cada una de las metodologías aplicadas a proyectos de big data?

Dentro de las principales propiedades de cada una de las metodologías identificadas, se encuentra el trabajo realizado por Shu et al. (2020), donde la implementación de algoritmos en el modelado permite sistemas innovadores que logran alcanzar mayor inteligencia, paralelismo y escalabilidad que las técnicas tradicionales. La propuesta de Shu et al. (2020) se centra en el desarrollo de un algoritmo inteligente basado en un modelo básico de demanda de usuario, que combina métodos cualitativos y cuantitativos, con el uso de técnicas como redes neuronales; parte de la recuperación y el análisis correspondiente de información para optimizar el proceso de innovación de productos. La puesta en práctica del modelado de algoritmos eficaces para el análisis inteligente es mencionada por Jin et al. (2020) como muestra de gran aporte a la obtención de un resultado exitoso. En este último artículo también se destaca la retroalimentación entre etapas (Jin et al., 2020), lo que genera la mejora continua del proceso.

La identificación de problemas es la base de la mejora continua y, ante el incremento de su complejidad, resulta oportuno adoptar un enfoque alternativo para su

resolución. Una posibilidad de hacer frente a la resolución de problemas complejos es la combinación de diferentes marcos de trabajo, lo que es tratado por Abdul Hamid et al. (2021) como una estrategia eficaz y efectiva para resolver un problema complejo. En la misma línea, García-Gil et al. (2021) mencionan la combinación de diferentes métodos de preprocesamiento de datos para mejorar la calidad de los mismos.

Otra característica que se destaca por Kavakli et al. (2020) es la priorización de los requisitos de *big data* desde diferentes perspectivas y en etapas tempranas. Esto permite una mejor alineación entre los objetivos del negocio y el comportamiento que debe poseer el sistema a desarrollar, lo cual resulta con alta relevancia al garantizar una trazabilidad entre el negocio y el rendimiento del sistema.

El modelado de los requisitos también es una característica a destacar como un punto de alta importancia, donde Tardío et al. (2020) indican que incurrir en errores en la toma de requisitos puede conducir al fracaso del proyecto. El trabajo desarrollado por Tardío et al. (2020) propone una metodología iterativa compuesta de cinco fases, donde a su vez cada una de ellas contiene múltiples subprocesos y se basa en el análisis de requisitos no funcionales derivados de las características de *big data*. Estos requisitos se utilizan como entrada para que el algoritmo propuesto genere una estructura inicial que se va a ir refinando sucesivamente durante las fases.

PI3: ¿Cuáles son las características comunes entre las distintas metodologías aplicadas a proyectos de big data?

Shu et al. (2020) destacan el análisis interdisciplinario hecho por profesionales de campos relacionados con la tecnología de la información, la inteligencia artificial y el *big data*, así como por profesionales vinculados con el conocimiento sociológico en los campos de la psicología, el comportamiento, la estética y la filosofía, los cuales también influyen en el desarrollo de una investigación. Contar con un equipo multidisciplinario permite el cruzamiento de diferentes contextos con el fin de incrementar y mejorar la comunicación entre el personal del equipo de trabajo (Markopoulos et al., 2024). Incorporar un enfoque multiperspectivo logra que se integren la comprensión del negocio, la integración de datos, la estadística, las hipótesis, el modelado, la visualización y el razonamiento analítico a lo largo del ciclo de vida del desarrollo de proyectos de *big data* (Ahmad et al., 2022).

Plantear la metodología como una guía en la identificación de diferentes requisitos es mencionado por Kavakli et al. (2020) y para el presente análisis es identificado como una característica común en el resto de los otros marcos de trabajo.

Otra de las características comunes a las metodologías es que las mismas sean iterativas para fomentar la retroalimentación que ayude a los profesionales de la tecnología de la información (TI) en la definición y validación de arquitecturas *big data* (Tardío et al., 2020). La iteración sobre las fases definidas permite la detección y corrección de errores, para así obtener resultados que colaboren en el proceso de toma de decisiones (Caffetti et al., 2023).

La búsqueda constante para trabajar con datos de calidad se menciona por García-Gil et al. (2021) donde se indica que la calidad de los datos se consigue mediante la aplicación de varias técnicas de preprocesamiento de datos para permitir diferentes enfoques del conjunto de datos.

3. CONCLUSIONES

En el presente trabajo se han revisado las diferentes metodologías aplicadas en proyectos de *big data* con el fin de dar respuesta a las preguntas de investigación planteadas inicialmente.

En la actualidad, las industrias deben dar una pronta e innovadora respuesta a las demandas del mercado, las cuales consideran que la masividad de los datos complejiza la resolución de problemas y el proceso de resolución genera altos costos.

Se analizan trabajos que reconocen a la gestión de los datos como parte importante dentro de un proyecto de *big data*, cuya imprecisa definición de los requerimientos puede llevar al fracaso. Incluso, la baja calidad de los datos puede resultar en problemas futuros, como inconsistencias, mientras que la imprecisión de resultados obtenidos podría conducir a una errónea toma de decisiones. Asimismo, la base de este tipo de proyectos es el análisis de la demanda comercial, los elementos y las necesidades del negocio, como así los procesos de etapas iterativas que permiten el modelo inicial.

En cuanto a limitaciones técnicas, se pueden mencionar la falta de experiencia para guiar el desarrollo de arquitecturas de *big data* y la necesidad de contar con medidas de seguridad que impidan ataques cibernéticos. Se reconoce la necesidad de contar con perspectivas multidisciplinarias que promuevan el trabajo colaborativo y creativo.

Por lo tanto, se requieren investigaciones futuras que permitan indagar y continuar desarrollando enfoques colaborativos entre distintos roles, dada la variedad de las temáticas identificadas como probables causas raíz que llevan al fracaso de los proyectos.

La presente revisión permite identificar puntos de interés para dar continuidad a futuras líneas de investigación de estudios que puedan aportar nuevos conocimientos que sean de relevancia en el campo de la gestión de proyectos de *big data*. Por ello, se enfatiza en la necesidad de abordar la definición, el análisis y la especificación de los requisitos funcionales y no funcionales, los cuales se consideran desde diferentes perspectivas con el objetivo de lograr una mejor armonización entre las metas del negocio y las funcionalidades del sistema a desarrollar como propuesta de solución.

REFERENCIAS

Abdul Hamid, K., Abu Bakar, M., Jalar, A., & Hakim Badarisman, A. (2021). Incorporation of big data in methodology of identifying corrosion factors in the semiconductor

package. En *2021 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)* (pp. 1-4). https://doi.org/10.1109/ICECCE52056.2021.9514240

Ahmad, Z., Yaacob, S., Ibrahim, R., & Farahwani, W. (2022). The review for visual analytics methodology. En *2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications* (pp. 1-10). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/HORA55278.2022.9800100

Bahit, E. (2012). *Scrum & Extreme Programming (para programadores)*. Creative Commons.

Caffetti, Y. A., Eckert, K., Ruidías, H. J., & Vera Laceiras, M. S. (2023). Data cleansing en entornos big data: mapeo sistemático de la literatura. En S. Rodríguez, M. Giménez y M. A. Molina (Comps.), *XXVIII Congreso Argentino de Ciencias de la Computación – CACIC 2022* (pp. 75-79). Editorial de la Universidad Nacional de La Rioja. https://repositoriosdigitales.mincyt.gob.ar/vufind/Record/SEDICI_1d437c59c0d397280848f3cfd422df97

Dai, H.-N., Wang, H., Xu, G., Wan, J., & Imran, M. (2019). Big data analytics for manufacturing internet of things: opportunities, challenges and enabling technologies. *Enterprise Information Systems*, *14*(9-10), 1279-1303. https://doi.org/10.48550/arXiv.1909.00413

Dastgerdi, A., & Gandomani, T. (2021). On the appropriate methodologies for data science projects. En *2021 International Conference on Information Technology* (pp. 667-673). Institute of Electrical and Electronic Engineers. https://doi.org/10.1109/ICIT52682.2021.9491712

Davenport, T. (2006). *Competing on Analytics*. Harvard Business Review https://hbr.org/2006/01/competing-on-analytics

Funde, S., & Swain, G. (2022). Big data privacy and security using abundant data recovery techniques and data obliviousness methodologies. *IEEE Access*, *10*, 105458-205484. https://doi.org/10.1109/ACCESS.2022.3211304

García-Gil, D., García, S., Xiong, N., & Herrera, F. (2021). Smart data driven decision trees ensemble methodology for imbalanced big data. *Cognitive Computation*, *16*, 1572-1588. https://doi.org/10.48550/arXiv.2001.05759

Jin, W., Yang, J., & Fang, Y. (2020). Application methodology of big data for emergency management. En *2020 IEEE 11th International Conference on Software Engineering and Service Science (ICSESS)* (pp. 326-330). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/ICSESS49938.2020.9237653

Kavakli, E., Sakellariou, R., Eleftheriou, I., & Mascolo, J. (2020). Towards a multi-perspective methodology for big data requirements. En *2020 IEEE International*

Conference on Big Data (pp. 5719-5720). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/BigData50022.2020.9378406

Khan, N., Alsaqer, M., Shah, H., Badsha, G., Abbasi, A., & Salehian, S. (2018). The 10 Vs, Issues and Challenges of Big Data. En *ICBDE '18* (pp. 52-56). Association for Computing Machinery. https://doi.org/10.1145/3206157.3206166

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing Systematic Literature Reviews in Software Engineering*. Keele University; Durham University Joint Report.

Krasteva, I., & Ilieva, S. (2021). Adopting agile software development methodologies in big data projects – a systematic literature review of experience reports. En *2020 IEEE International Conference on Big Data* (pp. 2028-2033). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/BigData50022.2020.9378118

Manzano, F., & Avalos, D. (2023). Análisis de calidad de los datos en las estadísticas públicas y privadas, ante la implementación del Big Data. *Ciencias Administrativas*, *11*(22), 1-11. https://doi.org/10.24215/23143738e119

Markopoulos, D., Tsolakidis, A., Karanikolas, N., Marinagi, A., & Skourlas, C. (2024). Applying soft system methodology for a clearer understanding of the future intensive care units. En *PCI '23: Proceedings of the 27th Pan-Hellenic Conference on Progress in Computing and Informatics* (pp. 163-170). Association for Computing Machinery. https://doi.org/10.1145/3635059.3635084

Ontiveros, E. (Dir.), Sabater, V. (Coord.)., Vizcaíno, D., Romero, M., & Llorente, A. (2018). *Economía de los datos. Riqueza 4.0*. Fundación Telefónica, Ariel España.

Project Management Institute. (2017). *A guide to the project management knowledge. PMBOK Guide* (6.ª ed).

Reggio, G., & Astesiano, E. (2020). Big-Data/Analytics projects failure: A literature review. En *2020 46th Euromicro Conference on Software Engineering and Advanced Applications* (pp. 246-255). https://doi.org/10.1109/SEAA51224.2020.00050

Saltz, J., & Hotz, N. (2020). Identifying the most common frameworks data science teams use to structure and coordinate their projects. En *2020 IEEE International Conference on Big Data* (pp. 2038-2042). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/BigData50022.2020.9377813

Shu, W., Sun, W., & Li, Y. (2020). The development trend of design methodology under the influence of artificial intelligence and big data. En *ICDLT '20: Proceedings of the 2020 4th International Conference on Deep Learning Technologies* (pp. 104-108). Association for Computing Machinery. https://doi.org/10.1145/3417188.3417214

Song, X., Zhang, H., Akerkar, R., Huang, H., Guo, S., Zhong, L., Ji, Y., Opdahl, A. L., Purohit, H., Skupin, A., Pottathil, A., & Culotta, A. (2020). Big data and emergency management: concepts, methodologies, and applications. *IEEE Transactions on Big Data*, *8*(2), 397-419. https://doi.org/10.1109/TBDATA.2020.2972871

Tardío, R., Maté, A., & Trujillo, J. (2020). An iterative methodology for defining big data analytics architectures. *IEEE Access*, *8*, 210597-210616. https://doi.org/10.1109/ACCESS.2020.3039455

Zúñiga, F., Mora Poveda, D., & Llerena Llerena, W. (2023). El Big Data y su implicación en el marketing. *Revista de Comunicación de la SEECI*, *56*, 302-321. https://doi.org/10.15198/seeci.2023.56.e832