INNOVA RESEARCH JOURNAL
ISSN: 2477-9024

UIDE Powered by Arizona State University

ARTÍCULO ORIGINAL
INNOVA Research Journal, ISSN 2477-9024
(Mayo-Agosto 2024). Vol. 9, No.2 pp. 40-57
DOI: https://doi.org/10.33890/innova.v9.n2.2024.2426
URL: http://revistas.uide.edu.ec/index.php/innova/index
Correo: innova@uide.edu.ec

Análisis bibliométrico de la producción científica sobre Lago de Datos

Bibliometric analysis of the scientific production on Data Lake

Galo Mauricio López-Sevilla
Pontificia Universidad Católica del Ecuador, Ambato, Ecuador
glopez@pucesa.edu.ec
https://orcid.org/0000-0003-4699-4875

Ricardo Patricio Medina-Chicaiza
Pontificia Universidad Católica del Ecuador, Ambato, Ecuador
Universidad Técnica de Ambato, Ambato, Ecuador
pmedina@pucesa.edu.ec
ricardopmedina@uta.edu.ec
https://orcid.org/0000-0002-2736-8214

Recepción: 24/10/2023 | Aceptación: 21/02/2024 | Publicación: 10/05/2024

Cómo citar (APA, séptima edición):
López-Sevilla, G. M., y Medina-Chicaiza, R. P. (2024). Análisis bibliométrico de la producción científica sobre Lago de Datos. *INNOVA Research Journal, 9*(2), 40-57. https://doi.org/10.33890/innova.v9.n2.2024.2426

Resumen

El presente trabajo desarrolla un análisis bibliométrico de la producción científica que contextualiza la corriente sobre Lago de Datos. En ese sentido, lago de datos son infraestructuras de almacenamiento y gestión de grandes volúmenes de datos provenientes de diversas fuentes, con la intensión de facilitar su acceso, análisis y compartición. El objetivo de este artículo es mostrar una visión cuantitativa de la producción científica del tema entre los años 2018 y 2022, para así comprender el estado actual de la investigación, identificar tendencias y áreas de investigación emergentes, evaluar el impacto y promover la colaboración entre investigadores. La metodología permitió llevar a cabo una revisión sistemática de la literatura mediante un análisis retrospectivo descriptivo, usándose como fuente de información la base de datos Scopus, misma que reflejó 73 artículos clave. En ese contexto, los resultados resaltan el interés sobre Lago de Datos desde el número de publicaciones de artículos por año hasta el top de principales autores, palabras clave y

revistas respecto a producción científica. De manera que, denotan la importancia y las preferencias en la investigación sobre esta temática relevante para diversos ámbitos o áreas.

Palabras claves: lago de datos; computación en la nube; datos grandes; análisis bibliométrico.

Abstract

This paper develops a bibliometric analysis of the scientific production that contextualizes the Data Lake current. In this sense, data lakes are infrastructures for storage and management of large volumes of data from various sources, with the intention of facilitating their access, analysis and sharing. The objective of this article is to show a quantitative view of the scientific production of the subject between 2018 and 2022, to understand the current state of research, identify trends and emerging research areas, evaluate the impact and promote collaboration among researchers. The methodology allowed carrying out a systematic review of the literature through a descriptive retrospective analysis, using the Scopus database as a source of information, which reflected 73 key articles. In this context, the results highlight the interest on Data Lake from the number of article publications per year to the top of main authors, keywords and journals with respect to scientific production. Thus, they denote the importance and preferences in research on this topic relevant to various fields or areas.

Keywords: data lake; cloud computing; big data; bibliometric analysis.

Introducción

El auge del *Big Data* ha sido un fenómeno disruptivo que ha revolucionado la forma en que las organizaciones gestionan y aprovechan la información. Este cambio radical en la manera en que concebimos y utilizamos los datos ha venido acompañado por una demanda insaciable de espacio de almacenamiento masivo (Rico et al., 2021). Esta necesidad no solo se basa en la capacidad de conservar grandes volúmenes de información, sino también en la habilidad de centralizar datos diversos y facilitar su análisis en tiempo real, impulsando así procesos de negocio más eficientes y orientados hacia resultados (Pasupuleti y Purra, 2017).

En este contexto, las herramientas y técnicas tradicionales de procesamiento y almacenamiento de datos ya no son suficientes para satisfacer las demandas del mundo del *Big Data* (Hernández, Duque, y Moreno, 2017). Se han vuelto obsoletas frente a la magnitud y complejidad de los datos generados en la actualidad, es así como, la cantidad de datos generados en todo el mundo está en constante aumento, así según el informe de Reinsel, Gantz, y Rydning (2020), se espera que para 2025 se generen 175 zettabytes de datos en todo el mundo, lo que supone un aumento exponencial respecto a los datos generados en 2018. La necesidad de una gestión eficiente de datos y su análisis se ha convertido en un aspecto vital en los negocios actuales.

De acuerdo con ello, (Agudelo, 2020; Wieder y Nolte, 2022) destacan que en 2010 James Dixon (CTO de Pentaho – Compañía de Software de Inteligencia de Negocios), acuñó el término Lago de Datos o *Data Lake* como una solución cada vez más popular para aquellas necesidades. Es por ello, que para Shehab, Badawy, y Arafat (2020, p. 92-101), un lago de datos es "un sistema de almacenamiento masivo, destinado a alojar todo tipo de datos, estructurados y no estructurados,

sin un esquema particular definido", siendo así un gran almacén de datos sin ninguna categorización o pauta estricta.

Como resultado, un lago de datos se convierte en un repositorio centralizado que almacena una gran cantidad de datos en su forma original y cruda, sin procesarlos ni transformarlos. Es así como, proviene de diversas fuentes, incluidos sensores, redes sociales, sistemas transaccionales y dispositivos móviles. De esta forma, las empresas podrán acceder y explotar una gran cantidad de información en tiempo real; y a la vez almacenar información estructurada, no estructurada y semiestructurada (Jarke et al., 2013). Entiéndase, por estructurada aquella información que maneja un esquema determinado; no estructurado aquel que no sigue un esquema o modelo de datos, pero si mantiene una estructura interna; y semiestructurada aquella que se organiza mediante etiquetas o *tags* que permiten crear y agrupar un conjunto de datos (Madera y Laurent, 2019).

De manera que, un lago de datos se utiliza para apoyar procesos tradicionales y modernos de extracción, transformación y carga de datos (ETL) (Romero y Melendres, 2023), así también para realizar análisis avanzados, minería de datos, aprendizaje automático y otras aplicaciones de inteligencia artificial (Grossman, 2019; Sakr y Gaber, 2019). Por tanto, existen tecnologías para construir a un lago de datos, incluyéndose *Hadoop, Apache Spark* y *Amazon S3*. En base aquello, entre los beneficios que sobresalen están la escalabilidad, flexibilidad y costos más bajos en el almacenamiento y procesamiento de grandes cantidades de datos (Lorenzo y López, 2022).

En teoría, lago de datos permite un acceso fácil y rápido a todos los datos del negocio (desde cualquier fuente y en cualquier formato), y también la experimentación, descubrimiento y exploración (Balseca, Colina, y Espinoza, 2021). Si se explota correctamente, un lago de datos se puede obtener una amplia gama de beneficios, entre ellos la facilidad de uso y accesibilidad (Goyal y Malviya, 2023). Según Rawat, Doku, y Garuba (2019), la gestión de la calidad del dato y la privacidad siguen siendo desafíos significativos en un lago de datos, al contrario de los enfoques tradicionales de la gestión de datos, como los *data warehouses* (Kimball y Ross, 2013).

En este sentido, el análisis de datos en un lago de datos ofrece numerosas ventajas en términos de eficiencia y productividad, permite a las organizaciones tomar decisiones más informadas y precisas, así como identificar oportunidades y desafíos en tiempo real, convirtiéndose en una solución cada vez más relevante y necesaria para el análisis de información.

De este modo, se plantea como objetivo mostrar una visión cuantitativa de la producción científica sobre lago de datos publicada en la base de datos Scopus, con la intensión de identificar tendencias y patrones en la investigación de este tema, así como reseñar autores y publicaciones más influyentes en el campo, para con ello proporcionar un contenido general de la evolución histórica de la investigación, lo cual es valioso para comprender el estado actual y el futuro de este campo. Por lo que, es imprescindible señalar que este trabajo sigue un procedimiento para alcanzar los resultados asociados a las preguntas de investigación delineadas durante el planteamiento del objeto de estudio; detallándose en los siguientes apartados.

Metodología

En este estudio bibliométrico descriptivo, se identificó, resumió, sistematizó y analizó la información generada en fuentes de información relevantes para el tema en análisis (Solano et al., 2019). Para ello, se realizó una revisión de los artículos publicados en la base de datos Scopus, reconocida como una de las más utilizadas en el campo de ciencia y tecnología, se siguió el protocolo establecido por la metodología de Revisión Sistemática de Literatura (RSL) para la recopilación y análisis de información secundaria de artículos científicos sobre la temática (Guix, 2018), se definió para ello criterios de inclusión y exclusión establecidos por los autores para obtener la información necesaria para el análisis bibliométrico de lago de datos (Oleo y Said, 2020) y que sea considerada como indicador bibliométrico para analizar ciertas características de la actividad científica del tema (Kitchenham, 2007).

Este trabajo se sustenta en el enfoque cuantitativo que es un método de investigación que se basa en la recopilación y análisis de datos numéricos (Escudero y Cortez, 2018; Hernández, Fernández, y Baptista, 2019), con la intensión de analizar la evolución y contribución de las perspectivas de varios autores sobre la temática, generando datos numéricos para sintetizar la información contenida en un conjunto de documentos. Estos datos incluyen la frecuencia de las palabras clave, la estructura del documento, las relaciones entre los documentos, cantidad de artículos, autores, revistas, idioma, año de publicación, tipo de documento, entre otros.

Procedimiento

El proceso de revisión propuesto en Kitchenham (2007) muy similar a los presentados por Moreno et al., (2018); y Snyder (2019) está estructurado de la siguiente manera:

Paso 1. Planeamiento

- Justificación de la revisión.
- Formulación de las preguntas de investigación.
- Diseño del protocolo de búsqueda.

Paso 2. Ejecución

- Búsqueda y extracción de documentos.
- Síntesis de datos.

Paso 3. Reporte

- Evaluación y exégesis de los resultados.

Paso 1. Planteamiento

a. Justificación de la revisión

Lago de datos es una tecnología emergente que ha ganado una gran popularidad en los últimos años. Sin embargo, la investigación sobre lago de datos aún está en sus primeras etapas, y hay una gran cantidad de información fragmentada disponible. Un análisis bibliométrico de la literatura sobre lago de datos es necesario para sintetizar información y proporcionar una visión general actualizada del estado de la investigación, aportándose beneficios como:

- Resumen de la investigación existente: Se resume la investigación existente sobre lago de datos, incluyéndose las principales tendencias, hallazgos y vacíos.

- Identificación de tendencias y oportunidades: Se identifican tendencias y oportunidades emergentes en el campo de lago de datos. Esto ayuda a los investigadores a mantenerse al día con los últimos avances e identificar áreas prometedoras para la investigación.

- Evaluación de la calidad de la investigación: Se evalúa la calidad de la investigación sobre lago de datos. Esto ayuda a los investigadores a identificar estudios de alta calidad y a evitar estudios sesgados o de baja calidad.

b. Formulación de las preguntas de investigación

Las preguntas de investigación actúan como brújulas que guían la búsqueda y análisis de estudios previos, lo cual permite una evaluación rigurosa y síntesis coherente de la información existente en un campo específico. En ese contexto, a continuación, se definen las preguntas de investigación (Véase Tabla 1):

Tabla 1

Preguntas de investigación sobre lago de datos

Preguntas de investigación
PI. 1 ¿Cuál es el número de publicaciones de artículos por año?
PI. 2 ¿Cuál es el promedio de citas por año?
PI. 3 ¿Cuáles son las fuentes más relevantes?
PI. 4 ¿Cuál es la producción por países?
PI. 5 ¿Cuál es la participación de las instituciones respecto al número de publicaciones?
PI. 6 ¿Cuáles son las referencias más citadas?
PI. 7 ¿Cuáles son las palabras claves más utilizadas?
PI. 8 ¿Cuál es el top de principales autores, palabras clave y revistas respecto a las publicaciones?

Fuente: Elaboración propia

c. Diseño del protocolo de búsqueda

- Criterios de inclusión: determinan los documentos seleccionados para el análisis:

  - Artículos que contengan las palabras clave:

  - *data lake,*
  - lago de datos
- Publicaciones realizadas en el periodo 2018-2022.
- Artículos científicos a texto completo.

- Criterios de exclusión: determinan los documentos que no serán seleccionados para el análisis:
  - Artículos en idioma diferente al inglés o español.
  - Resúmenes, ponencias, actas de conferencia.
  - Artículos que no correspondan a la subárea de conocimiento de Ciencias de la computación

- Fuentes de información: Se selecciona como fuente la base de datos Scopus, exportando la información pertinente a archivos .CSV y .RIS, que contendrán todos los datos útiles para la importación indirecta de referencias a otros gestores.

- Estrategia de búsqueda: Se define la siguiente cadena de búsqueda:

  TITLE-ABS-KEY ("data lake" OR "lagos de datos") AND PUBYEAR > 2017 AND PUBYEAR < 2023 AND (LIMIT-TO (SUBJAREA , "COMP")) AND (LIMIT-TO (DOCTYPE , "ar"))

- Revisión y selección de documentos: Mediante la aplicación de criterios de inclusión y exclusión a través de la cadena de búsqueda, se logró recopilar un conjunto total de 73 documentos que satisfacen los requisitos previamente definidos para la ejecución de un análisis bibliométrico. Estos documentos se han catalogado, organizado y procesados en Rstudio 2023.09.0+463. Estos datos incluyen información relevante como los nombres completos de los autores, el título de los documentos, el año de publicación, las afiliaciones institucionales de los autores, la revista de publicación, el país de origen de la publicación, el número de citas recibidas y otros datos pertinentes para el análisis bibliométrico (Villasís et al., 2020).

Paso 2. Ejecución de la búsqueda

a. Búsqueda y extracción de documentos

Para realizar la exploración de los artículos académicos, se efectuó una búsqueda en las bases de datos disponibles hasta el mes de diciembre de 2022, dicha búsqueda se ejecuta el mes de septiembre de 2023. Subsecuentemente, se procedió a la selección de documentos que incluyeran en sus secciones de título, resumen y palabras clave los descriptores "*data lake*" y "lagos de datos", los cuales fueron vinculados mediante el operador booleano "OR".

La extracción de los datos se inicia con la búsqueda a través de las palabras claves. El total de artículos encontrados en la base de datos científica fue 977, que luego de aplicar los criterios de inclusión y exclusión se redujeron a 73 artículos que serán parte del presente estudio.

b. Síntesis de datos

Los metadatos resultantes de la recolección de datos bibliométricos fueron sometidos a un análisis computacional exhaustivo dentro del entorno RStudio que es el entorno de desarrollo para el lenguaje R (Hernández y Lasso, 2021). Este proceso englobó una serie de etapas cruciales que incluyeron la preparación de los datos, la importación utilizando funciones especializadas de lectura de datos, así como la rigurosa limpieza y preparación de estos, orientada a la depuración de valores nulos, la normalización de identificaciones de autores, la mitigación de errores tipográficos, entre otros procedimientos (Gül y Ayik, 2023).

Dentro de la plataforma RStudio, se emplearon paquetes y herramientas de visualización altamente sofisticadas con el fin de generar gráficos descriptivos y representaciones gráficas de los datos, destinados a facilitar la interpretación de los resultados derivados del análisis bibliométrico. Estas visualizaciones revisten una relevancia crucial en la exploración y exposición de las tendencias y patrones inherentes a los datos bibliométricos, proporcionando una perspicaz comprensión del panorama investigativo subyacente. Además, los datos de las publicaciones elegidas fueron sometidos a un proceso de análisis adicional y la creación de representaciones gráficas descriptivas mediante la aplicación científica *ScienceScape* (Perilla, Orjuela, y Parra, 2020).

Resultados y Discusión

Paso 3. Reporte

En esta sección, se procederá a abordar las interrogantes de investigación previamente delineadas durante el planteamiento de la investigación. Seguidamente, se llevará a cabo una minuciosa evaluación y exégesis de los resultados asociados a cada una de las mencionadas preguntas de investigación.

PI. 1 ¿Cuál es el número de publicaciones de artículos por año?

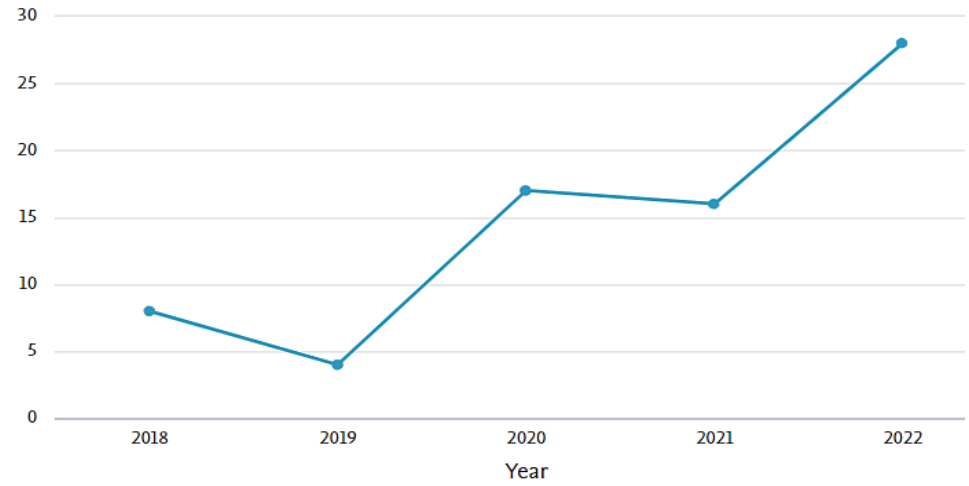
Tras analizar detenidamente un total de 73 artículos seleccionados previamente para el estudio de la producción científica en el ámbito de lago de datos, se concluye que el año 2022 ha sido testigo del mayor número de contribuciones con un total de 28 artículos (38.4%). El año 2021 sigue en importancia con 16 artículos (21.9%), seguido por el año 2020 con 17 artículos (23.3%) y el año 2019 con 4 artículos (5.5%). Por último, se destaca un modesto número de 8 artículos (11.0%) que emergieron en el año 2018.

La tendencia evidenciada en este análisis muestra un crecimiento notable en el número de artículos publicados a lo largo de los años, con un marcado aumento en el año 2022. Esto indica un interés en expansión en esta área de estudio, posiblemente impulsado por avances tecnológicos y la creciente relevancia de lago de datos en diversas disciplinas. Asimismo, la estabilidad en la producción de artículos en años anteriores y la presencia de contribuciones desde el año 2018 sugieren que esta temática se ha mantenido como un área de investigación continua y sostenible en el tiempo.

Para una representación visual completa de esta evolución temporal en términos de publicaciones, se presenta la Figura 1.

Figura 1

Artículos publicados por año



Fuente: Elaboración propia.

PI. 2 ¿Cuál es el promedio de citas por año?

La Tabla 2 exhibe el promedio de citas anuales, evidenciando que los ocho trabajos emitidos en 2018 ostentan una media de 5.31 citaciones por año durante su período de vigencia, destacándose como los de mayor valor citativo anual. En el año 2019, se publicaron cuatro trabajos sobre el mismo tema, con un promedio de 4.70 citaciones anuales. Por su parte, los trabajos de 2020 presentan una media de 4.26 citaciones por año, mientras que en 2021 esta cifra desciende a 3.69, y en 2022, se reduce aún más a 2.16 citaciones anuales.

En conclusión, los datos revelan que los trabajos publicados en 2018 exhiben el mayor impacto en términos de citaciones anuales, mientras que las publicaciones posteriores muestran una tendencia a la disminución en el número de citaciones por año, lo que sugiere un interés continuo en los trabajos de 2018 en comparación con las obras más recientes.

Tabla 2

Promedio de citas por año

Año	MeanTCperArt	Número de artículos	MediaTCperArt	CitableYears
2018	31.88	8.00	5.31	6
2019	23.50	4.00	4.70	5
2020	17.06	17.00	4.26	4
2021	11.06	16.00	3.69	3

Año	MeanTCperArt	Número de artículos	MediaTCperArt	CitableYears
2022	4.32	28.00	2.16	2

Fuente: Elaboración propia

PI. 3 ¿Cuáles son las fuentes más relevantes?

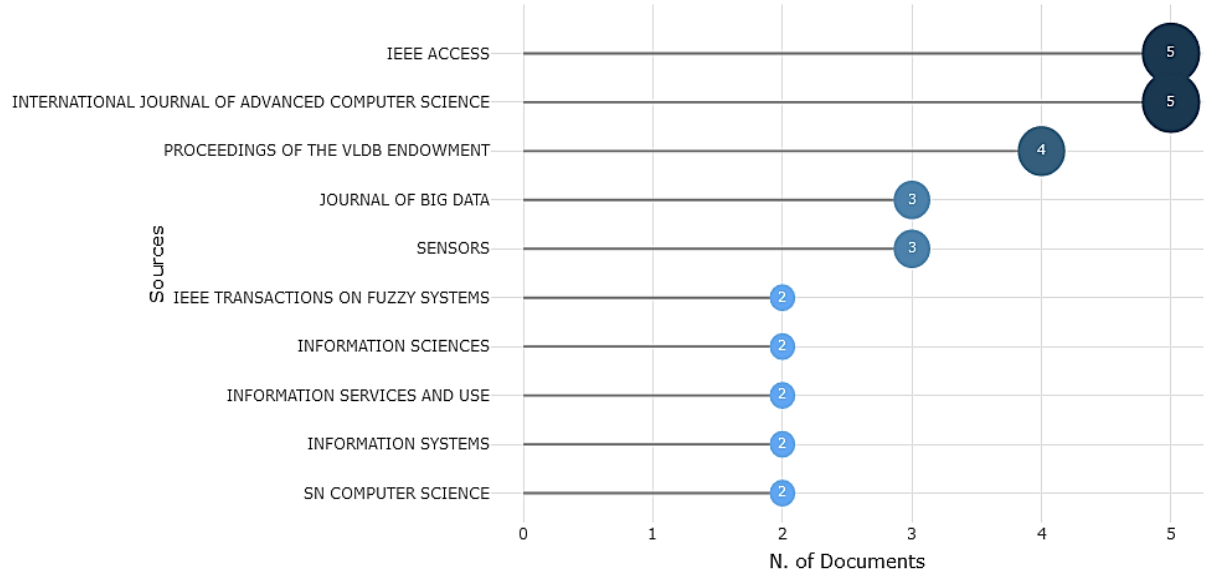
Las investigaciones sobre lago de datos se focalizan en 10 revistas de alcance global, que concentran el 41% de las contribuciones científicas relacionadas al tema. En este grupo las de *IEEE Access* y la *International Journal of Advanced Computer Science and Applications* es suman 5 publicaciones cada una, la fuente *Proceedings of the VLDB Endowment* ha publicado 4 trabajos y las revistas *Big Data* y *Sensores* 3 artículos cada una. Con dos trabajos publicados aparecen otras fuentes que se detallan en la Figura 2.

Es relevante destacar que el Instituto de Ingeniería Eléctrica y Electrónica (IEEE) ostenta una posición de liderazgo en el ámbito de investigación examinado, esta distinción refleja la excepcional calidad y el rigor científico inherente a sus publicaciones. Esta afirmación se sustenta en el notorio *Research Impact Score* de 25.2 atribuido a *IEEE Access* según el portal Research.com en 2023, indicando una significativa influencia en la comunidad científica (Research.com, 2023).

En contraste, la Revista International *Journal of Advanced Computer Science and Applications* exhibe un *Research Impact Score* de 1.3 (Research.com, 2023), indicándose un nivel más modesto de influencia en el ámbito investigativo. Por otro lado, la revista *Proceedings of the VLDB Endowment* sobresale con un destacado *Research Impact* de 13.20 (Research.com, 2023), lo que respalda su relevancia en el campo de estudio considerado.

Figura 2

Fuentes relevantes



Fuente: Elaboración propia

PI. 4 ¿Cuál es la producción por países?

La producción de investigación en el campo de lago de datos se distribuye de manera heterogénea en diferentes países, como se desprende de un análisis de los 73 artículos investigados en el período comprendido entre 2018 y 2022.

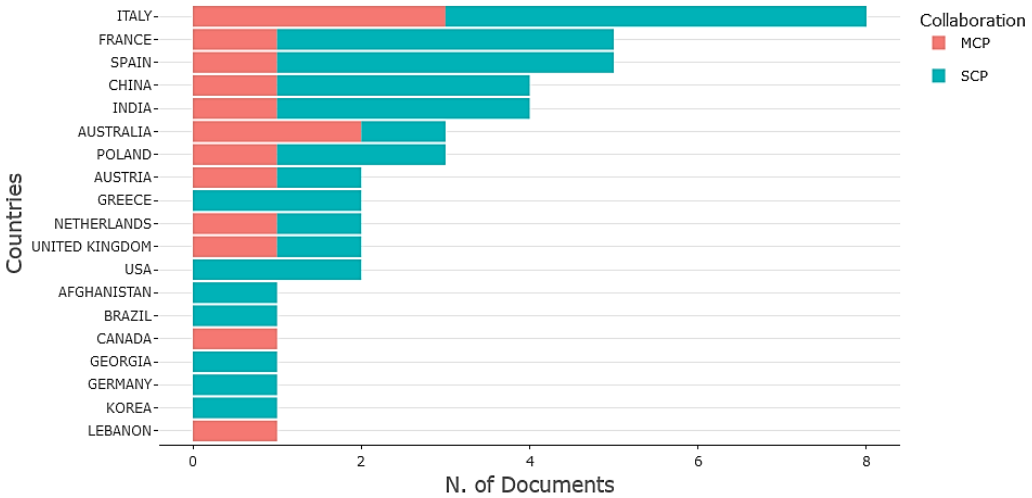
En términos cuantitativos, como lo muestra la Figura 3, Italia destaca como el líder en investigaciones sobre lago de datos, contribuyendo con el 8% del total de artículos publicados en el mencionado período, de los cuales se puede también observar que la mayoría son publicados de manera independiente en cada país (PC: Colaboración simple), mientras que de manera equitativa una minoría es producida con colaboración múltiple entre autores de varios países (MCP: colaboración múltiple). Italia ha demostrado una significativa presencia en la producción científica de la disciplina.

Francia y España, emergen como destacados actores en el ámbito de la investigación, contribuyendo con un 5% de los artículos publicados cada uno. Esto resalta la importancia de su participación en la generación de conocimiento en esta área, superando a naciones americanas en términos de producción científica.

China e India comparten el tercer lugar en la lista, cada una con una contribución del 4% en la producción de artículos. Este nivel de participación indica su relevancia en el panorama investigativo internacional en este campo. Además de los países mencionados anteriormente, hay una serie de naciones, como Australia, Polonia, Austria, Grecia, Países Bajos, Reino Unido, entre otros, que también contribuyen a la producción científica, aunque con un porcentaje menor en comparación con los líderes mencionados. Esta diversidad de contribuciones resalta la naturaleza global de la investigación en lago de datos y muestra la influencia de diversos actores a nivel internacional en la generación de conocimiento en esta área.

Figura 3

Producción por países



Fuente: Elaboración propia

PI. 5 ¿Cuál es la participación de las instituciones respecto al número de publicaciones?

En el contexto del análisis de la producción científica en el ámbito de lago de datos, se ha constatado que, de un total de 10 instituciones académicas que se muestran en la tabla 3 y que han contribuido con investigaciones en esta área de estudio, la entidad *Fondazione Policlinico Universitario A. Gemelli IRCCS* de Roma, Italia, sobresale con un total de 27 publicaciones atribuidas a su autoría. Del mismo modo, se han identificado diecisiete trabajos científicos publicados por la prestigiosa *Université Clermont Auvergne*, ubicada en Francia.

En una perspectiva análoga, se ha detectado que la *Silesian University of Technology* ha realizado una contribución significativa al campo, con la publicación de diez artículos científicos de relevancia. Asimismo, *La Trobe University* ha aportado al cuerpo de conocimiento con la publicación de ocho manuscritos científicos de alta calidad. Además, el *Consorzio Nazionale Interuniversitario per le Telecomunicazioni y DISI – University of Bologna* han contribuido de manera destacada con seis trabajos científicos cada una en este dominio de investigación.

Finalmente, las cuatro instituciones restantes también han contribuido de manera significativa, cada una con cinco publicaciones. Aunque su producción es menor en comparación con las principales instituciones, no se puede subestimar su contribución a la literatura científica en este campo. Estas instituciones son*: School of Engineering of the Polytechnic of Porto (ISEP), Shanghai Jiao Tong University, King's College London* y *University of New South Wales.*

En general, estos resultados sugieren que existe un interés y una actividad crecientes en la investigación de lago de datos, con una concentración significativa de producción en algunas instituciones líderes. Esto puede ser un indicativo de las áreas de especialización de estas instituciones y podría señalar oportunidades de colaboración en el futuro para avanzar en el conocimiento en este emocionante campo de estudio.

Tabla 3

Instituciones de afiliación de autores

Affiliation	Articles
FONDAZIONE POLICLINICO UNIVERSITARIO A. GEMELLI IRCCS	27
UNIVERSITÉ CLERMONT AUVERGNE	17
SILESIAN UNIVERSITY OF TECHNOLOGY	10
LA TROBE UNIVERSITY	8
CONSORZIO NAZIONALE INTERUNIVERSITARIO PER LE TELECOMUNICAZIONI	6
DISI – UNIVERSITY OF BOLOGNA	6
SCHOOL OF ENGINEERING OF THE POLYTECHNIC OF PORTO (ISEP)	5
SHANGHAI JIAO TONG UNIVERSITY	5
UNIVERSITY COLLEGE LONDON	5
UNIVERSITY OF NEW SOUTH WALES	5

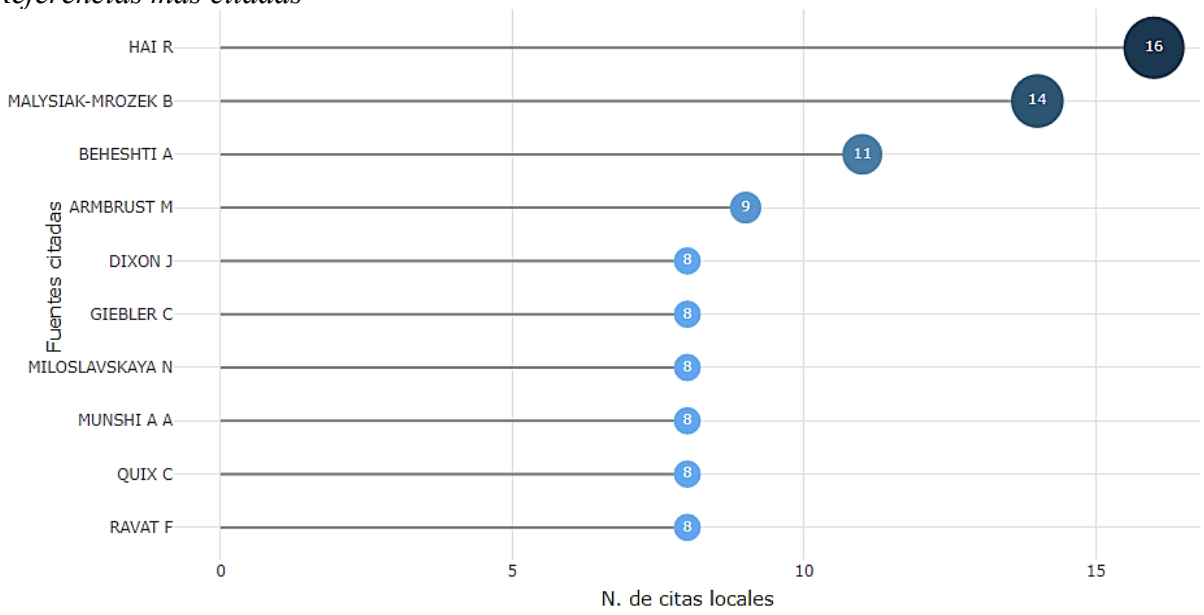
Fuente: Elaboración propia

PI. 6 ¿Cuáles son las referencias más citadas?

La representación gráfica de la Figura 4 muestra las diez fuentes primordiales que han sido objeto de múltiples menciones en los manuscritos relativos al tema de análisis. Rihan Hai, afiliado a la Universidad de Aachen en Alemania, ostenta la máxima preeminencia en este contexto, con un total de dieciséis menciones en sus obras científicas. De igual manera, Bożena Małysiak-Mrozek, procedente de la Universidad de Tecnología de Silesia en Polonia, se distingue con un total de catorce citaciones en sus publicaciones. Amin Beheshti, cuya filiación recae en la Universidad de Nueva Gales del Sur en Sídney, Australia, se evidencia con once menciones a sus trabajos en los manuscritos objeto de análisis. Mientras tanto, los trabajos de Michael Armbrust han sido objeto de cita en nueve ocasiones, y se advierte la presencia de un conjunto de autores cuyas investigaciones han sido referenciadas en ocho ocasiones, conformando un sólido grupo de relevancia en la esfera científica. Estos datos reflejan la influencia y pertinencia de los trabajos de dichos autores en el contexto de la investigación sobre el tema de lago de datos.

Figura 4

Referencias más citadas



Fuente: Elaboración propia

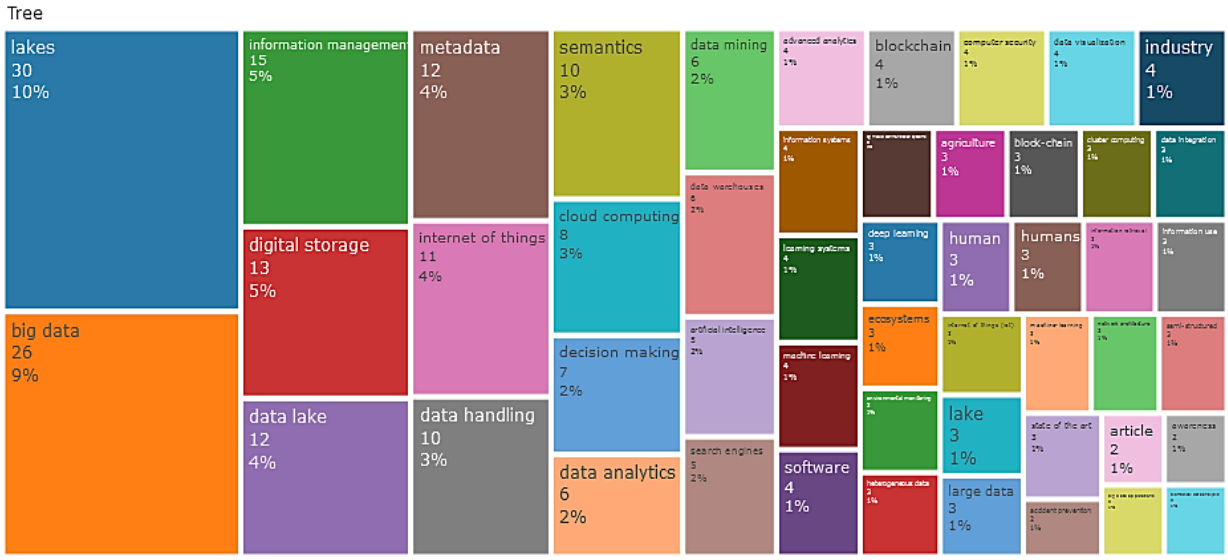
PI. 7 ¿Cuáles son las palabras claves más utilizadas?

La Figura 5 exhibe un diagrama de árbol que desglosa las palabras clave extraídas de los documentos sometidos a consulta. Destaca que el término *lakes* prevalece de manera significativa en un total de 30 documentos, mientras que *big data* ocupa la segunda posición, presentándose en 26 de los artículos analizados. A cierta distancia, observamos la presencia de las expresiones *information management, digital storage, data lake* y *metadata.*

Este análisis revela la distribución de estas palabras clave en un corpus de 73 documentos, destacando la prominencia de *lakes* y *big data*. Sin embargo, es relevante señalar que otras expresiones técnicas, como *internet of things, data handling, semantics* y *cloud computing*, también juegan un papel significativo en el discurso científico, a pesar de su menor frecuencia de aparición en los documentos. Este hallazgo subraya la importancia de un conjunto diverso y amplio de conceptos dentro del contexto de la investigación en cuestión, tal como se muestra en la Figura 6.

Figura 5

Palabras clave en diagrama de árbol



Fuente: Elaboración propia

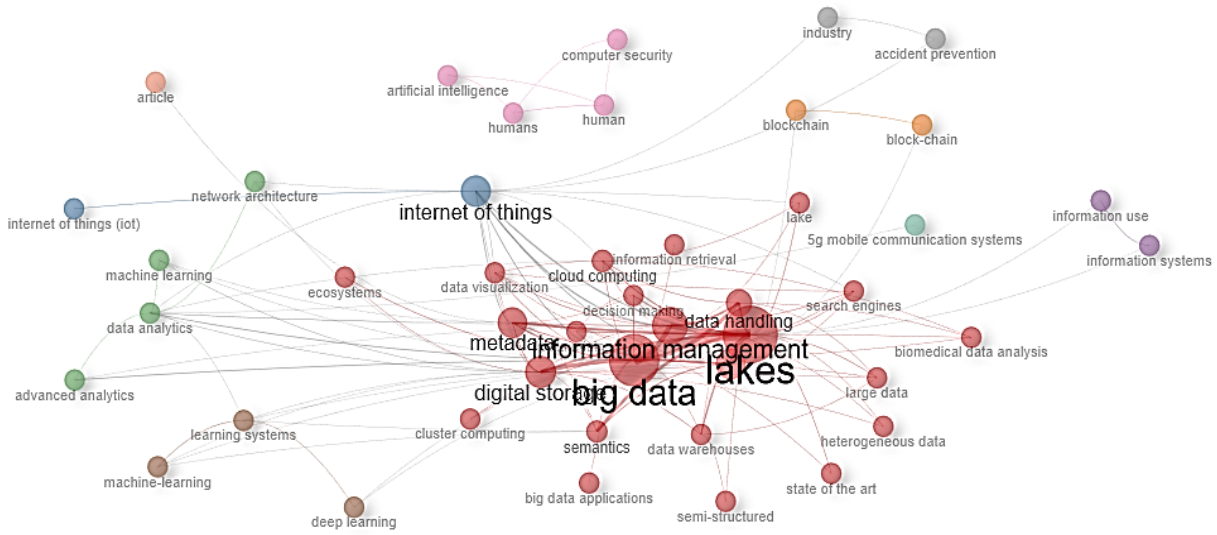
La representación gráfica de co-ocurrencia, ilustrada en la Figura 6, despliega de manera integral las interconexiones y asociaciones entre las palabras clave coexistentes en los documentos bajo escrutinio (López, Amado, y Miotto, 2022). Esto proporciona una percepción significativa de las tendencias y pautas emergentes en el compendio de información analizado. La red de co-ocurrencia resalta las relaciones intrínsecas entre términos compartidos en un mismo documento, y su centralidad indica la importancia de las palabras clave que mantienen múltiples vínculos con otras en el ámbito de estudio. Entre estas palabras clave centrales se incluyen *lakes, big data, information management, digital storage, metadata* y otras de naturaleza afín.

Observamos, adicionalmente, la presencia de grupos discernibles de nodos o clústeres, que denotan subtemas específicos en el ámbito investigado. Estos subtemas abarcan áreas como analítica de datos, el dominio de datos voluminosos, inteligencia artificial, transacciones, aprendizaje automático, sistemas de información y otros.

En síntesis, esta visualización de la red de co-ocurrencia ofrece una valiosa comprensión de relaciones semánticas y estructura de información subyacente en los documentos examinados, lo que contribuye al enriquecimiento del conocimiento en el campo de estudio.

Figura 6

Red de co-ocurrencia en palabra clave



Fuente: Elaboración propia

PI. 8 ¿Cuál es el top de principales autores, palabras clave y revistas respecto a las publicaciones?

La Figura 7 presenta un análisis de los flujos de datos interconectados entre autores, palabras clave y revistas, en el contexto de la producción científica relacionada con la temática de lago de datos.

Se observa que 30 autores han empleado de manera coincidente las palabras clave *data lake* y *data lakes* en sus investigaciones. Además, 24 de estos autores han abordado el concepto de *big data* en sus trabajos, mientras que 23 artículos han relacionado su producción científica con las palabras clave asociadas a la gestión de datos, tales como *data governance, data warehouse, data integration, data management* y *data science.*

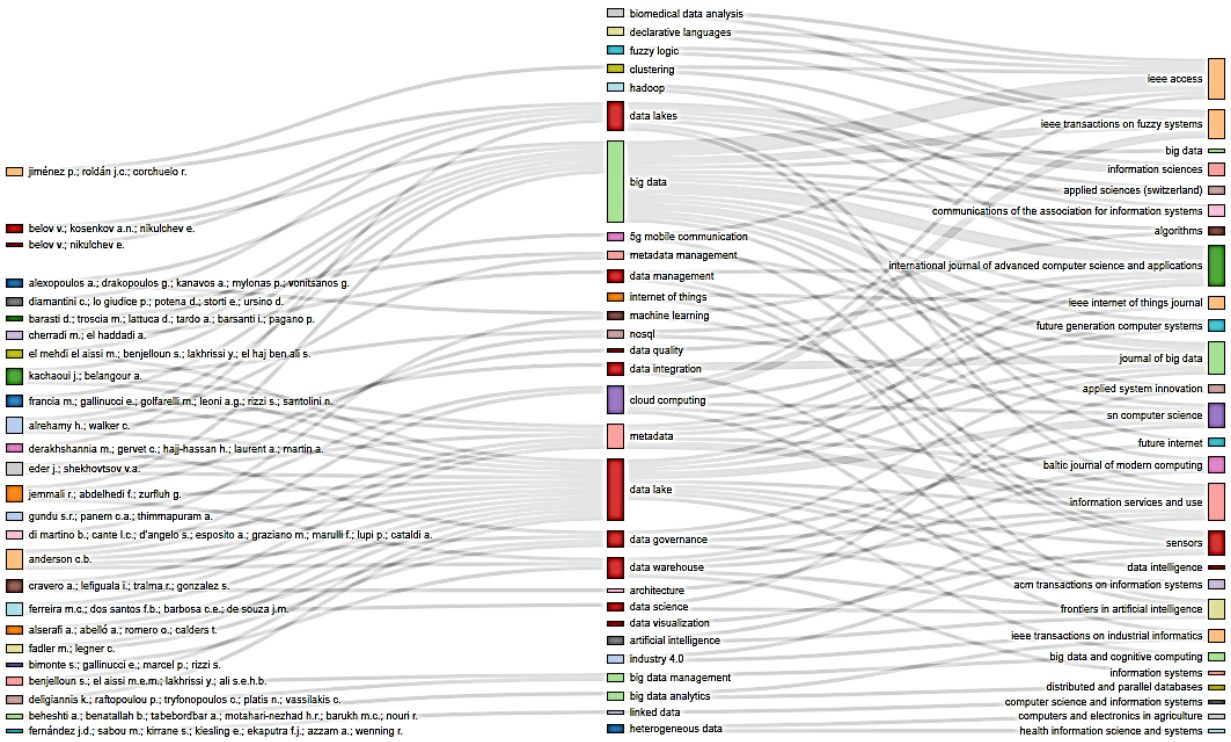
Los flujos de estas palabras clave se distribuyen de manera heterogénea a lo largo de varias revistas especializadas, entre las cuales se destacan *IEEE Access, International Journal of Advanced Computer Science and Applications, Proceedings of the VLDB Endowment, Journal of Big Data, Sensors* y otras publicaciones relevantes en el ámbito científico.

Este análisis revela una interesante convergencia de autores y palabras clave en la literatura científica relacionada con lago de datos, lo que indica un enfoque creciente en aspectos clave de

la gestión y utilización de grandes conjuntos de datos. La dispersión de estos flujos hacia diferentes revistas sugiere un interés generalizado en esta temática dentro de la comunidad científica, lo que puede tener implicaciones significativas para el avance de la investigación en este campo en particular.

Figura 7

Relación entre autores, palabras clave y revistas



Fuente: ScienceScape.

Conclusiones

Se observa una tendencia de crecimiento constante en la producción de artículos relacionados con lago de datos a lo largo de los años, alcanzando su punto máximo en 2022. Esto sugiere un creciente interés en esta área, posiblemente impulsado por avances tecnológicos y su creciente relevancia interdisciplinaria. La estabilidad en la producción de artículos en años anteriores indica que esta temática ha mantenido su sostenibilidad como área de investigación.

Las publicaciones de 2018 destacan por su mayor impacto en términos de citaciones anuales, mientras que las publicaciones más recientes muestran una tendencia a la disminución en el número de citaciones por año. Este patrón podría indicar un interés continuo en trabajos más antiguos en comparación con los más recientes.

La investigación sobre lago de datos se concentra en revistas especializadas, siendo *IEEE Access* la líder tanto en cantidad de publicaciones como en un destacado *Research Impact Score*, lo que refleja la calidad y el rigor científico inherentes a las publicaciones de IEEE. Además, la revista *Proceedings of the VLDB Endowment* también sobresale en este ámbito, demostrando la diversidad de fuentes que respaldan la globalidad de la investigación en lago de datos.

Italia lidera en investigaciones sobre lago de datos, seguida de cerca por Francia y España, con China e India contribuyendo de manera significativa. Esto subraya la relevancia de estos países en el panorama investigativo internacional, y la diversidad de contribuciones de otras naciones enfatiza la naturaleza global de la investigación en esta área.

Algunas instituciones académicas se destacan en la producción científica en el campo de lago de datos, como la *Fondazione Policlinico Universitario A. Gemelli IRCCS* en Roma, Italia, y la *Université Clermont Auvergne* en Francia. Estas instituciones líderes podrían ofrecer oportunidades de colaboración en el futuro y reflejan áreas de especialización en la investigación.

Finalmente, el diagrama de árbol destaca que el término *lakes* prevalece en un total de 30 documentos, seguido de una red co-ocurrencia que demuestra interconexiones y asociaciones entre palabras clave coexistentes en los documentos bajo escrutinio, tales como: *big data, information management, digital storage, metadata,* otras. De ahí que, tanto empresas como instituciones están aumentando la tecnología en sus procesos y necesitan mayor volumen para almacenar los datos, resaltándose así el denominado lago de datos.

Referencias Bibliográficas

Agudelo Patiño, J. C. (2020). *Data lakes: aplicaciones, herramientas y arquitecturas. Colombia: Universidad Tecnológica de Pereira* [Trabajo de Grado, Universidad Tecnológica de Pereira, Colombia]. https://bit.ly/49CWmUB

Balseca-Chávez, F., Colina-Vargas, A. M., y Espinoza-Mina, M. A. (2021). Identificación de amenazas informáticas aplicando arquitecturas de Big Data. *INNOVA Research Journal, 6*(3), 141-167. https://revistas.uide.edu.ec/index.php/innova/article/view/1860/1953

Escudero, C., y Cortez, L. (2018). *Técnicas y métodos cualitativos para la investigación científica.* Editorial Utmach. https://bit.ly/3I5ZGvd

Goyal, P., & Malviya, R. (2023). Challenges and opportunities of big data analytics in healthcare. *Health Care Science, 2*(6), 1-11.

Grossman, R. (2019). Data lakes, clouds, and commons: a review of platforms for analyzing and sharing genomic data. *Trends in Genetics, 35*(3), 223-234. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6474403/

Guix, J. (2018). El análisis de contenidos: ¿qué nos están diciendo?. *Calidad Asistencial, 23*(1), 26-30. https://www.elsevier.es/es-revista-revista-calidad-asistencial-256-articulo-el-analisis-contenidos-que-nos-S1134282X08704640

Gül, M., & Ayik, Z. (2023). Enrichment studies in gifted education: a bibliometric analysis with RStudio. *Participatory Educational Research, 10*(3), 266-284. https://dergipark.org.tr/en/pub/per/issue/76200/1257077

Hernández, E., Duque, N., y Moreno, J. (2017). Big data: una exploración de investigaciones, tecnologías y casos de aplicación. *Tecnológicas, 20*(39), 1-24. https://revistas.itm.edu.co/index.php/tecnologicas/article/view/685/671

Hernández, M., y Lasso, E. (2021). *Revisión bibliográfica de las investigaciones realizadas sobre páramos en las últimas cinco décadas* [Trabajo de Grado, Universidad de los Andes, Colombia].

Hernández, R., Fernández, C., y Baptista, M. (2019). *Metodología de la investigación.* México: McGraw-Hill.

Jarke, M., Lenzerini, M., Vassiliou, Y., & Vassiliadis, P. (2013). Fundamentals of Data Warehouses. *Sigmod Record, 32*(2), 55-56.

Kimball, R., & Ross, M. (2013). *The data warehouse toolkit.* Toronto: John Wiley & Sons.

Kitchenham. (2007). *Guidelines for performing Systematic Literature Reviews in.* Inglaterra: University of Durham. https://bit.ly/3UQP3Uu

López, V., Amado, A., y Miotto, U. (2022). Un enfoque bibliométrico a los procedimientos gráficos como método de investigación. *Expresión Gráfica Arquitectónica, 27*(45), 218-231. https://polipapers.upv.es/index.php/EGA/article/view/16451

Lorenzo, P., y López, G. (2022). *Análisis, diseño e implementación de una arquitectura de servicios cloud para un lago de datos en el ambito turístico* [Trabajo de Grado, Universidad de La Coruña, España]. https://ruc.udc.es/dspace/handle/2183/32125

Madera, C., & Laurent, A. (2019). The next information architecture evolution: the data Lake wave. *Management of Digital EcoSystems, 6*(9), 1-8. https://hal-lirmm.ccsd.cnrs.fr/lirmm-01399005/document

Moreno, B., Muñoz, M., Cuellar, J., Domancic, S., y Villanueva, J. (2018). Revisiones Sistemáticas: definición y nociones básicas. *Clínica de Periodoncia, Implantología y Rehabilitación Oral, 11*(3), 184-186.

Oleo, C., y Said, E. (2020). La producción científica en el estudio de experiencia de usuario en la educación: caso Web of Science y Scopus. *Perspectiva, 32*(7), 1-7. https://humanas.blog.scielo.org/es/2020/05/21/la-produccion-cientifica-en-el-estudio-de-experiencia-de-usuario-en-la-educacion-caso-web-of-science-y-scopus/

Pasupuleti, P., & Purra, B. (2017). *Data lake development with big data.* Reino Unido: Packt Publishing. https://www.packtpub.com/product/data-lake-development-with-big-data/9781785888083

Perilla, R., Orjuela, W., y Parra, C. (2020). *Análisis de futuro: algunos métodos alternativos a la caja de herramientas de la prospectiva francesa.* Colombia: Universidad del Tolima. https://bit.ly/3wphmzj

Rawat, D., Doku, R., & Garuba, M. (2019). Cybersecurity in big data era: from securing big data to data-driven security. *Transactions on Services Computing, 20*(1), 1-18. https://ieeexplore.ieee.org/ielaam/4629386/9642441/8673585-aam.pdf

Reinsel, D., Gantz, J., & Rydning, J. (2020). *The digitization of the world from edge to core.* Estados Unidos: Seagate. https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

Research.com. (2023). *Best Journals - Computer Science - IEEE Access.* https://research.com/journal/ieee-access

Research.com. (2023). *Best Journals - Computer Science - International Journal of Advanced Computer Science and Applications.*

Research.com. (2023). *Best Journals - Computer Science - Proceedings of the VLDB Endowment.* https://research.com/journal/proceedings-of-the-vldb-endowment-1

Rico, D., Maestre, G., Medina, Y., y Areniz, Y. (2021). Universidad inteligente: factores claves para la adopción de internet de las cosas y big data. *Ibérica de Sistemas y Tecnologías de Información, 4*(1), 63-79.

Romero, A., y Melendres, J. (2023). Uso de data warehouse para la toma de decisiones empresariales: una revisión literaria. *Científica de Sistemas e Informática, 3*(2), 1-12. https://bit.ly/3uxrF3T

Sakr, S., & Gaber, M. (2019). *Large scale and big data: processing and management.* Estados Unidos: Auerbach Publications. https://www.routledge.com/Large-Scale-and-Big-Data-Processing-and-Management/Sakr-Gaber/p/book/9781138033948

Shehab, N., Badawy, M., & Arafat, H. (2020). Big Data Analytics Concepts, Technologies Challenges, and Opportunities. *Advanced Intelligent Systems and Informatics, 11*(14), 92-101. https://www.researchgate.net/publication/336219391_Big_Data_Analytics_Concepts_Technologies_Challenges_and_Opportunities

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research, 10*(4), 333-339. https://www.sciencedirect.com/science/article/pii/S0148296319304564

Solano, E., Castellanos, S., López, M., y Hernández, J. (2019). La bibliometría: una herramienta eficaz para evaluar la actividad científica postgraduada. *MediSur, 7*(4), 59-62. http://scielo.sld.cu/pdf/ms/v7n4/v7n4a745.pdf

Villasís, M., Rendón, M., García, H., Miranda, M., y Escamilla, A. (2020). La revisión sistemática y el metaanálisis como herramientas de apoyo para la clínica y la investigación. *Alergia México, 67*(1), 62-72. https://www.scielo.org.mx/pdf/ram/v67n1/2448-9190-ram-67-01-62.pdf

Wieder, P., & Nolte, H. (2022). Toward data lakes as central building blocks for data management and analysis. *Frontiers in Big Data, 16*(2), 1-18. https://www.researchgate.net/publication/362793690_Toward_data_lakes_as_central_building_blocks_for_data_management_and_analysis