TecnoLógicas
ISSN 0123-7799
ISSN-e 2256-5337
Vol. 20, No. 39
Mayo - agosto de 2017

Big Data: una exploración de investigaciones, tecnologías y casos de aplicación

Big Data: an exploration of research, technologies and application cases

Emilcy J. Hernández-Leal[1],
Néstor D. Duque-Méndez[2] y Julián Moreno-Cadavid[3]

Recibido: 28 de mayo de 2016,
Aceptado: 15 de marzo de 2017

Cómo citar / How to cite
E.J. Hernández-Leal, N.D. Duque-Méndez y J. Moreno-Cadavid, "Big Data: una exploración de investigaciones, tecnologías y casos de aplicación", *TecnoLógicas*, vol. 20, no. 39, mayo - agosto, 2017.

[1] Esp. en Gerencia Estratégica de Proyectos, Estudiante de Maestría en Ingeniería Administrativa, Administradora de sistemas informáticos, Departamento de Ingeniería de la Organización, Facultad de Minas, Universidad Nacional de Colombia, Medellín-Colombia, ejhernandezle@unal.edu.co

[2] PhD. en Ingeniería, MSc. en Ingeniería de Sistemas, Especialista en Sistemas, Ingeniero Mecánico, Facultad de Administración, Departamento de Informática y Computación, Universidad Nacional de Colombia, Manizales-Colombia, ndduqueme@unal.edu.co

[3] PhD. en Ingeniería – Sistemas, MSc. en Ingeniería de Sistemas, Ingeniero de Sistemas e Informática, Departamento de Ciencias de la Computación y de la Decisión, Facultad de Minas, Universidad Nacional de Colombia, Medellín-Colombia, jmoreno1@unal.edu.co

Resumen

Big Data se ha convertido en una tendencia a nivel mundial y aunque aún no cuenta con un concepto científico o académico consensuado, se augura cada día mayor crecimiento del mercado que lo envuelve y de las áreas de investigación asociadas. En este artículo se reporta una exploración de literatura sobre Big Data, que comprende un estado del arte de las técnicas y tecnologías asociadas a Big Data, las cuales abarcan captura, procesamiento, análisis y visualización de datos. Se exploran también las características, fortalezas, debilidades y oportunidades de algunas aplicaciones y modelos que incluyen Big Data, principalmente para el soporte al modelado de datos, análisis y minería de datos. Asimismo, se introducen algunas de las tendencias futuras para el desarrollo de Big Data por medio de la definición de aspectos básicos, alcance e importancia de cada una. La metodología empleada para la exploración incluye la aplicación de dos estrategias, una primera corresponde a un análisis cienciométrico; y la segunda, una categorización de documentos por medio de una herramienta web de apoyo a los procesos de revisión literaria. Como resultados se obtiene una síntesis y conclusiones en torno a la temática y se plantean posibles escenarios para trabajos investigativos en el campo de dominio.

Palabras clave

Big Data, análisis de datos, ciencia de los datos, minería de datos, análisis Big Data.

Abstract

Big Data has become a worldwide trend and although still lacks a scientific or academic consensual concept, every day it portends greater market growth that surrounds and the associated research areas. This paper reports a systematic review of the literature on Big Data considering a state of the art about techniques and technologies associated with Big Data, which include capture, processing, analysis and data visualization. The characteristics, strengths, weaknesses and opportunities for some applications and Big Data models that include support mainly for modeling, analysis, and data mining are explored. Likewise, some of the future trends for the development of Big Data are introduced by basic aspects, scope, and importance of each one. The methodology used for exploration involves the application of two strategies, the first corresponds to a scientometric analysis and the second corresponds to a categorization of documents through a web tool to support the process of literature review. As results, a summary and conclusions about the subject are generated and possible scenarios arise for research work in the field.

Keywords

Big data, data analysis, data science, data mining, big data analysis.

1. INTRODUCCIÓN

El crecimiento en el volumen de datos generados por diferentes sistemas y actividades cotidianas en la sociedad ha forjado la necesidad de modificar, optimizar y generar métodos y modelos de almacenamiento y tratamiento de datos que suplan las falencias que presentan las bases de datos y los sistemas de gestión de datos tradicionales. Respondiendo a esto aparece Big Data, término que incluye diferentes tecnologías asociadas a la administración de grandes volúmenes de datos provenientes de diferentes fuentes y que se generan con rapidez [1].

A pesar de que el término Big Data se asocia principalmente con cantidades de datos exorbitantes, se debe dejar de lado esta percepción, pues Big Data no va dirigido solo a gran tamaño, sino que abarca tanto volumen como variedad de datos y velocidad de acceso y procesamiento. En la actualidad se ha pasado de la transacción a la interacción, con el propósito de obtener el mejor provecho de la información que se genera minuto a minuto [2].

Con el auge del Big Data se ha dado cabida también a un nuevo concepto, Data Science o Ciencia de los Datos, que se usa de forma genérica para hacer referencia a la serie de técnicas necesarias para el tratamiento y manipulación de información masiva desde un enfoque estadístico e informático. Incluyendo también el surgimiento de un nuevo perfil profesional, el "Data Scientist" [3], las personas capacitadas en este perfil deben saber del negocio, de las herramientas computacionales y de análisis e interpretación estadística.

Ahora bien, al revisar Big Data, pensando en la creación de soluciones que incluyan problemas enmarcados en este enfoque, se pueden encontrar cuatro fases donde se agrupan o clasifican las diferentes tecnologías de soporte, estas son: generación, adquisición, almacenamiento y análisis de datos. En [4] se define la primera fase, generación, como un proceso propio de diversas actividades de la sociedad, en estas se genera una cantidad inmensa de datos, que, según su naturaleza, puede estar almacenada y estructurada o puede corresponder a datos sin ninguna estructura, pero con características de gran valor. En la segunda fase, se incluye la colección de todos estos datos generados en la vida diaria, la trasmisión y pre-procesamiento de los mismos es de gran importancia, ya que muchos conjuntos de datos presentan redundancia o datos inútiles y si no se tratan pueden incrementar el espacio de almacenamiento innecesariamente y afectar los resultados de una fase de análisis. La fase de almacenamiento de Big Data ha generado la necesidad de generar estudios y propuestas de nuevas estrategias que permitan afrontar los tipos de datos que no se pueden gestionar con un sistema de gestión de bases de datos relacionales. Surgen entonces, tecnologías de almacenamiento de datos masivos como almacenamiento con conexión directa y el almacenamiento en red, también diferentes motores NoSQL. Finalmente, la fase de análisis debe atender a la necesidad de extraer rápidamente información desde los datos masivos para poder generar valor en las organizaciones y facilitar procesos de toma de decisiones, se requiere de tecnologías que faciliten incluso el análisis en tiempo real.

Siguiendo los lineamientos para la construcción de artículos de revisión [5], este artículo tiene como objetivo presentar una visión general acerca de Big Data incluyendo un análisis cienciométrico de las publicaciones en este campo y haciendo una exploración cuidadosa de una serie de trabajos en el tema, que contemplan aplicaciones, oportunidades, desafíos y retos de Big Data. A su vez, se hace una breve introducción de algunas tecnologías y técnicas adoptadas para la implementación de soluciones a problemas de Big Data. Esta exploración concluye con la presentación de puntos clave y principales aportes encontrados.

Lo que resta del artículo se organiza de la siguiente forma: en la siguiente Sección se presenta la metodología y principales hallazgos de esta exploración. En la Sección 3 se muestran las tecnologías y técnicas para el tratamiento de Big Data. Por su parte, en la Sección 4 se muestran algunas tendencias y retos en el campo, y se finaliza en la Sección 5, con las conclusiones que parten del análisis del contexto presentado en las secciones previas.

2. METODOLOGÍA

El desarrollo de esta exploración se realizó siguiendo dos estrategias. Como primera estrategia, se hizo un acercamiento cienciométrico por medio de la herramienta bibliográfica SCOPUS, un índice bibliográfico que contiene una colección representativa, completa y multidisciplinar a nivel mundial. La segunda estrategia comprende el análisis de algunos trabajos particulares referentes al soporte y estructura conceptual de la temática abordada. Estos fueron seleccionados y clasificados por medio de la herramienta ToS (Tree of Science), desarrollada en la Universidad Nacional de Colombia. A continuación, se detallarán cada una de las estrategias y se mostrarán los resultados obtenidos.

2.1 Primera estrategia de exploración

SCOPUS es una de las más grandes bases de datos de resúmenes y citas de literatura revisadas por pares, contienen artículos de revistas científicas, libros y artículos de congresos, posibilitando tener una visión global de la producción académica e investigativa en campos de la ciencia, tecnología, medicina, artes y humanidades [6]. Además, esta herramienta permite clasificar, refinar y analizar de forma ágil los resultados obtenidos a partir de una ecuación de búsqueda, con ello se puede extraer información relevante de la temática de interés que se esté abordando. Para este acercamiento se utilizó como ecuación de búsqueda "big data" y a continuación se presentan algunos aspectos relevantes que se extrajeron de los resultados en SCOPUS. Se decidió utilizar esta ecuación de búsqueda poco delimitada, porque se pretende presentar un estado general de presencia y tratamiento de la temática.

El total de recursos encontrados fue de 16.902. En la Fig. 1 se presenta el número de documentos publicados por año. Se aprecia que los estudios del tema llevan poco más de un lustro, se puede ver que en el año 2012 es cuando realmente toma fuerza y viene teniendo un crecimiento significativo, pasando de 646 resultados en 2012 a 7508 resultados en 2015. Para el 2016 se presentan los resultados correspondientes a los cuatro primeros meses del año.

Como se aprecia en la Fig. 2, si se revisa según el tipo de recurso, se ve una marcada tendencia hacia los artículos de conferencia, con un total de 9.493 resultados. Los artículos científicos muestran 4.824 resultados, mientras que los capítulos de libro y los libros solo despliegan 388 y 88 resultados respectivamente, lo anterior ratifica la etapa naciente en que se encuentra este campo de estudio, puesto que sus bases teóricas apenas se están consolidando.

Revisando los resultados agrupados por país de publicación, se puede ver una concentración en Estados Unidos y China como se aprecia en la Fig. 3. En los países europeos se encuentra un número también significativo de trabajos, mientras que en Sur América, Oceanía y África, el desarrollo de investigaciones en el campo es aún incipiente.

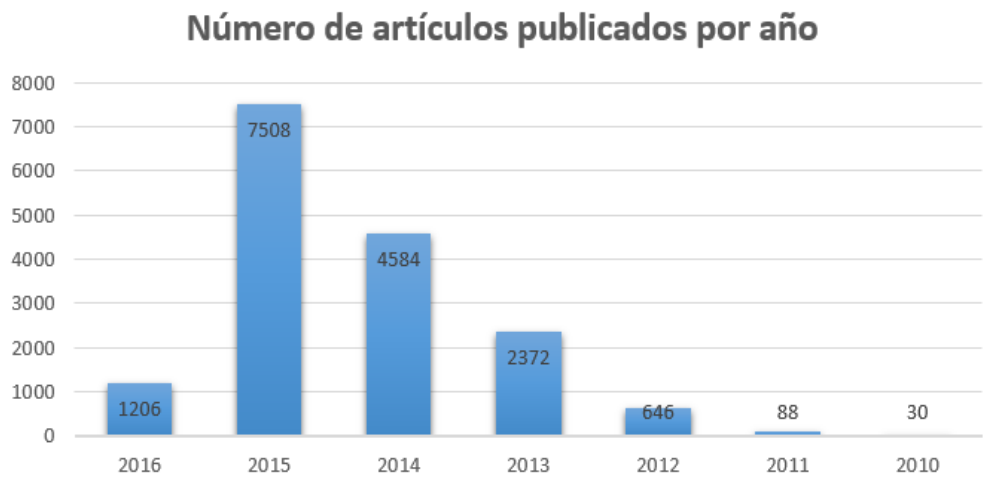

Fig. 1. Número de artículos publicados por año, periodo 2010 – 2016. Fuente: Autores.

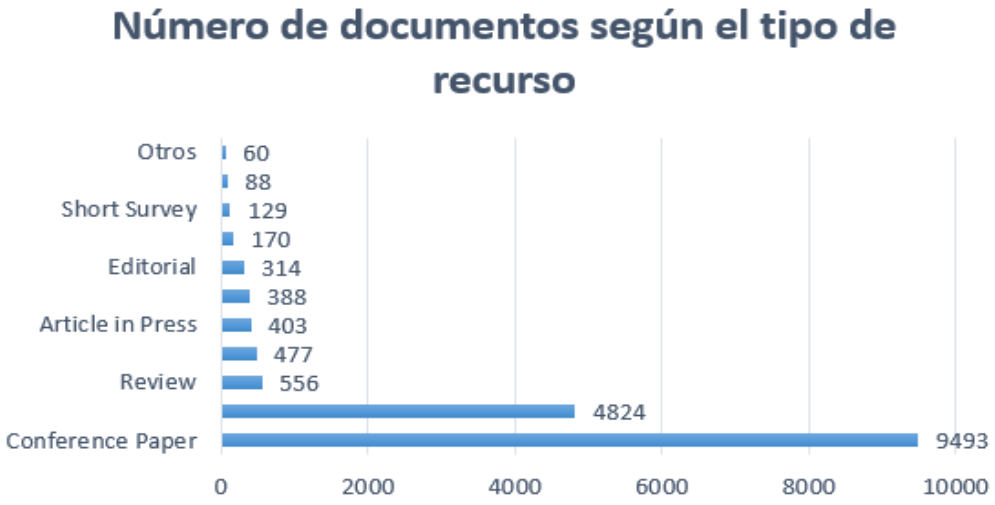

Fig. 2. Número de documentos según el tipo de recurso Fuente: Autores.

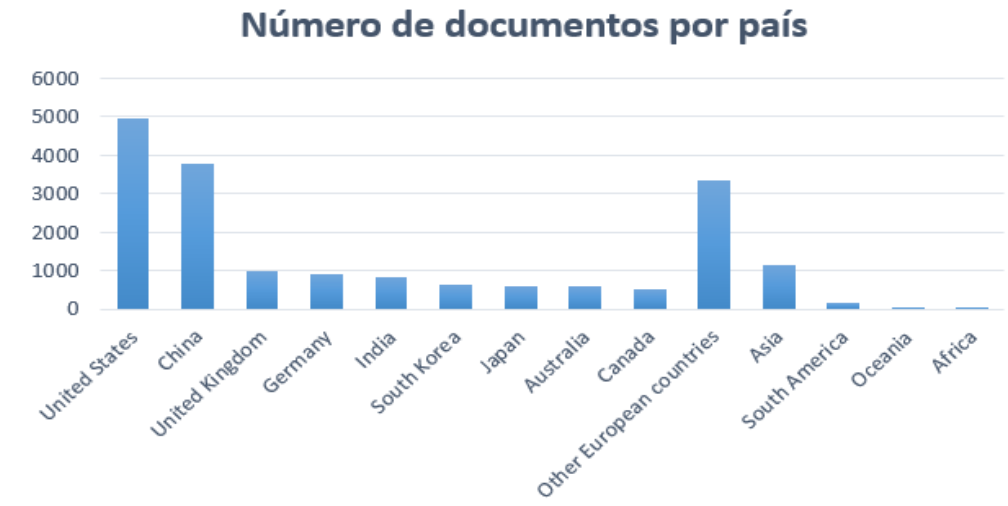

Fig. 3. Número de documentos según el país de publicación. Fuente: Autores.

2.2 Segunda estrategia de exploración

Como segunda estrategia se hizo un análisis detallado de algunos documentos, utilizando para su selección y clasificación una herramienta desarrollada desde el Grupo de Investigación en Ambientes Inteligentes Adaptativos – GAIA – y como parte de una tesis doctoral en la Universidad Nacional de Colombia – Sede Manizales. La herramienta llamada ToS (Tree of Science) [7], funciona en la web y se puede acceder a ella a través del enlace http://tos.manizales.unal.edu.co/.

Los resultados que brinda la herramienta son construidos a partir de la utilización de una serie de algoritmos de redes complejas, los cuales optimizan los resultados de la búsqueda y selección de documentos científicos publicados. Esta herramienta clasifica los documentos en "raíz", "tronco" y "ramas" a partir de la lista de trabajos encontrados. Los documentos raíz hacen referencia a las investigaciones que dan soporte al enfoque o temática abarcada, los documentos tronco son aquellos que dan estructura al tema y los documentos rama son las perspectivas y tendencias.

Para el caso particular, se usaron los siguientes parámetros para la búsqueda:

- Palabras de búsqueda: "Big Data"
- Restricción de años: 2010-2016
- Categoría de Web of Science: computer science information systems
- Tipo de documento: documentos científicos

Se hizo la búsqueda en el índice bibliográfico Web of Science (índice con el cual trabaja la herramienta ToS) y se obtuvo un total de ciento setenta y cuatro (174) artículos para los parámetros de búsqueda. A partir del análisis de este grupo de artículos y de las referencias citadas en los mismos, la herramienta ToS hizo el respectivo refinamiento y retornó diez artículos considerados raíz, en el tronco se clasificaron otros diez y setenta artículos fueron ubicados en las ramas, como se aprecia en la Fig. 4. Cabe aclarar que en este documento no se reportará la totalidad de los artículos arrojados por la herramienta, se ha realizado una selección de los documentos que cubren la temática, permitiendo tener una visión general del estado del arte y de las tendencias y campos de trabajo.

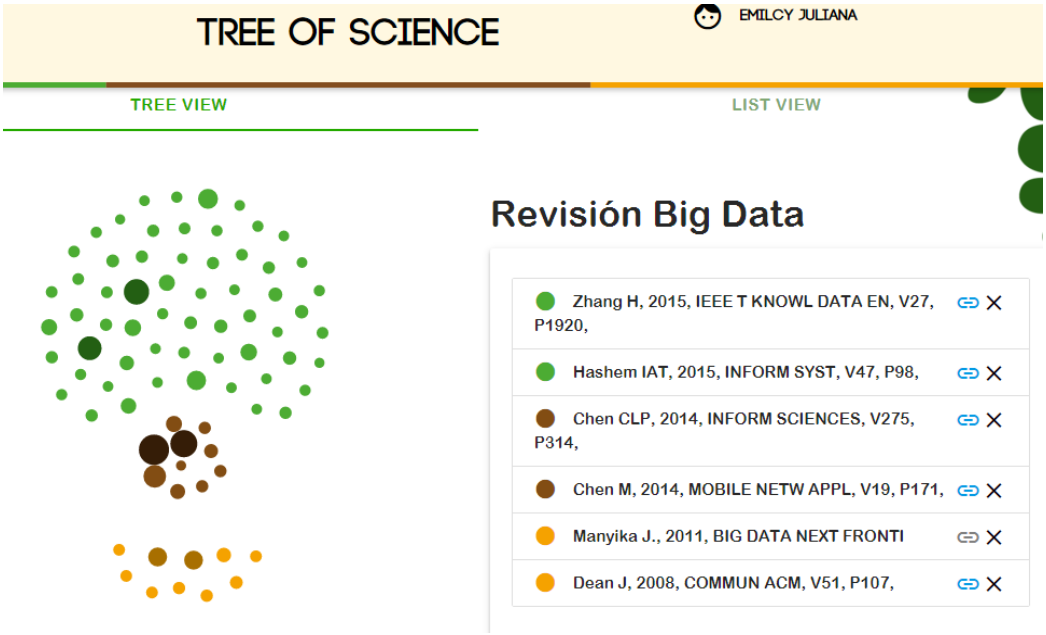

Fig. 4. Estructura del árbol de la ciencia generado por la herramienta ToS.
Fuente: Generado en http://tos.manizales.unal.edu.co/

Seguidamente, se presenta un recuento de cinco de los documentos clasificados como raíz del enfoque, la selección de los documentos presentados en este escrito se hace después del análisis por parte de los autores de la totalidad de los documentos raíz e identificando los más relevantes.

Retomando lo anteriormente mencionado, los documentos raíz corresponden a investigaciones o aportes que dan soporte a la temática. Dos de los artículos corresponden a revisiones acerca de tecnologías que permitieron y dieron pie al nacimiento de la tendencia Big Data: la computación en la nube y el paradigma de programación Map Reduce. Los dos documentos siguientes corresponden a dos libros, el primero concebido por la empresa IBM, en el que se analiza Big Data desde una perspectiva empresarial y una perspectiva tecnológica, cabe señalar que IBM es una de las compañías que provee soluciones Big Data a nivel empresarial. El segundo es una Guía de Hadoop, donde se describen los conceptos asociados a este paradigma, se habla del modelo MapReduce, de otras herramientas Big Data y se presentan algunos casos de estudio donde se ha aplicado Hadoop. Finalmente, se toma un informe producto de los puntos de vista recogidos por el autor en un evento de exploración de Big Data e inferencia de software.

Map Reduce [8] es un modelo de programación asociado a las implementaciones que requieren procesamiento y generación de grandes bases de datos. Los cómputos se hacen en términos de una función de mapeo y otra de reducción y el cálculo se hace de forma paralelizada. Los autores muestran Map Reduce como un modelo que facilita el trabajo con sistemas paralelos y distribuidos, ya que oculta detalles de paralelización, tolerancia a fallos, optimización y balance de carga. Es necesario optimizar los recursos de red cuando se trabaja con Map Reduce, por ello es bueno leer los datos desde discos locales y reducir la cantidad de datos enviados a través de la red. También la ejecución redundante disminuye el impacto de las máquinas lentas, pérdida de datos y fallos de máquina.

Otra de las bases de Big Data es la computación en la nube o cloud computing logrando que los desarrolladores ya no requieran de grandes inversiones en hardware, la elasticidad de recursos sin necesidad de pagar por servicios Premium de gran escala es un hito fundamental en la historia de las tecnologías de la información. Cloud computing se convirtió en un tema popular y objeto de artículos, workshops, conferencias y revistas. Se auguró el crecimiento de cloud computing independientemente de si los servicios adquiridos son a bajo o alto nivel de abstracción. Se afirma que el almacenamiento, cómputo y las redes deben concentrarse en la escalabilidad horizontal de los recursos virtualizados en lugar del rendimiento de un solo nodo. Se planteó la necesidad de que las aplicaciones de software tuviesen una rápida escalabilidad y que los sistemas de hardware fuesen diseñados a escala de contenedor [9].

Big Data surge como una nueva era en la exploración y utilización de datos. Desde la perspectiva empresarial Big Data no representa solo grandes volúmenes de datos, se deben considerar los patrones extraídos a partir de los datos y que pueden generar procesos de innovación. Desde la perspectiva tecnológica se presenta Hadoop como la principal herramienta desarrollada para el tratamiento de Big Data, incluyendo el manejo de sistemas de archivos distribuidos y el paradigma de programación Map Reduce. En la primera parte, correspondiente a la perspectiva empresarial, se presenta una comparación entre las soluciones Big Data y las soluciones tradicionales de Datawarehouse. Sin querer buscar una ganadora, se expone la ventaja de usar Datawarehouse cuando se trata de analizar datos estructurados que vienen de varios sistemas y de mediciones relativamente estables. Respecto a las plataformas basadas en Hadoop, funcionan bien con datos semiestructurados y desestructura-

dos, así como también cuando se requiere de procesos de descubrimiento de datos [10].

Partiendo de la necesidad de almacenamiento y análisis de los datos se desarrolla el ecosistema Hadoop, los sistemas de archivos distribuidos, el desarrollo de aplicaciones con MapReduce, el lenguaje de consultas Hive y otras herramientas como HBase, ZooKeeper y Sqoop. En [11] se presenta una guía completa, tanto de forma conceptual como con ejemplos de aplicación de Hadoop y de varias herramientas asociadas a este. Uno de los casos estudiados es el de Hadoop y Hive para Facebook. Facebook inicialmente usaba data warehousing sobre una instancia Oracle, sin embargo, con su crecimiento se tuvo que pensar en nuevas alternativas, Hadoop fue atractiva porque ya se usaba en Yahoo para procesamientos internos y usaba el modelo MapReduce popularizado por Google.

El crecimiento de los datos, como la explosión de las redes móviles, la computación en la nube y las nuevas tecnologías son descritas en [12]. Esto ha dado un aumento al incomprensible mundo de la información, que se suele describir como Big Data. Este informe captura los puntos de vista recogidos durante un evento de exploración de temas de Big Data e inferencia de software. Las compañías que han sido pioneras en el uso de analíticas profundas sobre grandes bases de datos han sido las que operan sobre internet, como son los motores de búsqueda, los sitios de redes sociales y los sitios de comercio en línea. Sin embargo, el desarrollo de nuevos tipos de sensores remotos como telescopios, videocámaras, monitores de tráfico, máquinas de resonancia magnética, sensores químicos y biológicos y sensores de monitoreo ambiental, se han generado nuevos flujos de datos digitales. Así mismo, las personas a través de sus teléfonos celulares, computadores personales, sitios web y otro tipo de dispositivos digitales generan grandes flujos de datos personales. Lo anterior deja ver que Big Data presenta oportunidades incalculables para la formulación de investigación científica, acelera la innovación y puede ayudar a mejorar ámbitos que van desde la salud hasta el Gobierno. También se abren nuevas oportunidades de negocio porque surgen mecanismos que permiten entender las dinámicas de negocio en tiempo real, como el comportamiento de los consumidores, las actividades de vida nocturna, los mercados, entre otros. Cabe anotar que Big Data presenta también retos y peligros, ya que las tecnologías de datos son cada vez más penetrantes, intrusivas y difíciles de entender.

A manera de resumen de los principales documentos considerados raíz, en la Tabla 1 se presenta una síntesis de estos.

Los documentos ubicados en el tronco, son aquellos que dan estructura a la temática o campo de estudio, hacen referencia a estudios de revisión frente a los avances, desafíos y perspectivas de Big Data y tecnologías asociadas, estos son presentados a continuación. En este caso también se optó por presentar los cinco documentos que después de la revisión por parte de los autores son considerados los más relevantes.

Tabla 1. Síntesis de los principales documentos raíz. Fuente: Autores.

Autores - Año	Título	Tipo de documento	Dimensión, campo o herramienta analizado	Síntesis del documento
Dean, J. Ghemawat, S. 2008	MapReduce: simplified data processing on large clusters	Artículo en revista científica	MapReduce	Map Reduce es uno de los enfoques que se muestra como base sólida de las soluciones Big Data, ya que desde el paradigma de distribución de procesamiento se pueden afrontar problemas de tratamiento de grandes volúmenes de datos que las herramientas tradicionales no soportan.
Armbrust, M. et al. 2010	A view of cloud computing	Artículo en revista científica	Cloud Computing	Cloud computing, o computación en la nube es una tendencia que logró virtualizar procesos que requerían de grandes inversiones en hardware, las cuales no siempre podían ser afrontadas por las organizaciones. Con ello se ha permitido también, que el crecimiento de los datos y su procesamiento se pueda escalar.
Zikopoulos, P. et al. 2011	Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data	Libro	Perspectivas Big Data: Empresarial, tecnológica	Tomando dos perspectivas, la empresarial y la tecnológica se analiza Big Data y se concluye que se debe realizar un análisis previo a la implementación de soluciones desde este enfoque, puesto que según el tipo de datos que se manejen, las soluciones tradicionales pueden seguir siendo más eficientes respecto a las que implementan Hadoop o procesamiento distribuido, en otras ocasiones lo indicado es tener una solución mixta.
White, T. 2011	Hadoop: The Definitive Guide	Libro	Ecosistema Hadoop	Se presenta todo el ecosistema Hadoop, tecnologías de almacenamiento, procesamiento y análisis de datos adaptadas a diferentes tipos de datos. Desde esta guía se presentan casos de éxito de soluciones Big Data desde lo conceptual y aplicadas a casos de estudio.
Bollier, D. 2010	The Promise and Peril of Big Data	Informe	Analíticas Big Data	El mundo presenta una tendencia cada vez más marcada hacia la generación de datos. Ya sea desde la interacción de las personas con las nuevas tecnologías, las mediciones de variables del entorno o los flujos de datos personales. Todas estas fuentes de datos se convierten en posibilidades de investigación científica e innovación empresarial.

En [13] se presenta Big Data, sus aplicaciones, las oportunidades y desafíos de estas tecnologías, así como también técnicas de última generación que se han adoptado para hacer frente a los problemas de Big Data. Se discuten algunas metodologías utilizadas para tratar cantidades considerables de datos como es la computación granular, la computación en la nube, la computación bio-inspirada y la computación cuántica. Destacan el papel que han jugado los datos como promotores de diferentes campos científicos, como la astronomía, la meteorología, la bioinformática y la biología computacional. Dichos campos basan gran parte de su descubrimiento científico en el análisis de grandes volúmenes de datos. Otro de los aportes significativos, es la descripción de los principios para el diseño de sistemas Big Data. Estos son: (1) Buenas arquitecturas y frameworks son necesarios y de alta prioridad. (2) Soportar una variedad de métodos analíticos. (3) No hay un tamaño definido para

todo. (4) Conducir el análisis de los datos. (5) El procesamiento debe ser distribuido. (6) El almacenamiento de los datos debe ser distribuido. (7) Es necesaria una coordinación entre las unidades de procesamiento y de datos.

En [4] se revisan algunas de las tecnologías relacionadas a Big Data como computación en la nube, Internet de las cosas, centros de datos y Hadoop. También se mencionan las fases de la cadena de valor de Big Data y finalmente se examinan algunos casos de aplicación como gestión empresarial, internet de las cosas, redes sociales, aplicaciones médicas, inteligencia colectiva y redes eléctricas inteligentes. En cuanto a las fases de Big Data, se definen cuatro principales: generación, adquisición, almacenamiento y análisis de datos. Los autores afirman que, sin tratar de predecir el futuro, el panorama de Big Data se concentrará en: datos con escalas y diversidad cada vez mayores y estructuras mucho más complejas, la necesidad de rendimiento de los recursos de datos, Big Data promoverá la fusión transversal de la ciencia, tendrá grandes retos de visualización de datos y una orientación a los datos cada vez más marcada. A su vez, se presentan los desafíos de Big Data, estos requieren de un esfuerzo investigativo y son agrupados en las siguientes categorías: investigación teórica, desarrollo tecnológico, implicaciones prácticas y seguridad de datos.

Desde una perspectiva de la administración de los datos, en [14] se presenta una discusión acerca de la diversidad de Big Data, las necesidades de integración y limpieza, las consultas e indexación y finalmente la minería y análisis sobre Big Data. El inicio de Big Data va directamente relacionado con el crecimiento de los datos generados por la sociedad. Estos datos suelen caracterizarse por su heterogeneidad y por la variedad de fuentes desde las cuales provienen, sin embargo, se pueden clasificar estas fuentes de acuerdo con donde son generadas. Los autores proponen las siguientes: contenidos generados por usuarios, estos vienen de aplicaciones que cuentan con usuarios masivos, por ejemplo tweets o blogs; datos transaccionales, son generados por sistemas masivos que procesan transacciones y operaciones como por ejemplo lectores de radio frecuencia, transacciones empresariales, entre otras; datos científicos, estos son producidos por aplicaciones o experimentos de datos-intensivos, por ejemplo datos del genoma o datos de asistencia sanitaria; datos web, provienen de los procesos que soportan aplicaciones web como búsquedas y minería, también de los billones de páginas web que existen; Grafos de datos, corresponden a un enorme número de nodos de información y las relaciones entre estos nodos. Adicionalmente, se habla de la reducción de Big Data, entendida como la reducción de las cantidades exorbitantes a los segmentos significativos, se presentan técnicas como machine learning y el procesamiento paralelo masivo para este fin.

Es importante también, tener en cuenta cómo en el área de la industria y los negocios se ha presentado una explosión en el número de datos, causada principalmente por el rápido desarrollo del internet, nuevos conceptos como el internet de las cosas y la computación en la nube. Big data se ha constituido como un "tópico caliente" que atrae la atención no solo de la industria, sino también de la academia y del Gobierno. Los autores presentan desde diferentes perspectivas el significado y las oportunidades que nos brinda el ecosistema Big Data y dan una serie de condiciones necesarias para que un proyecto de Big Data sea exitoso. En primer lugar, se deben tener claros los requerimientos independientemente de si son técnicos, sociales o económicos. En segundo lugar, para trabajar de forma eficiente con Big Data se requiere explorar y encontrar la estructura central o el kernel de los datos a ser procesados, ya que al tener esto se puede caracterizar el comportamiento y las propiedades subyacentes a Big Data. En tercer

lugar, se debe adoptar un modelo de administración top-down, se puede considerar también un modelo bottom-up, sin embargo, solo serviría cuando se trata de problemas específicos, y luego tratar de unirlos para formar una solución completa es complejo. Por último, los autores exponen la necesidad de abordar desde los proyectos Big Data soluciones integradas, no con esfuerzos aislados [15].

Los retos que se desprenden del consumo y creación de información a través de la red incluyen necesidades de captura, manejo y procesamiento de grandes volúmenes de datos. En [16] los autores proponen un teorema llamado "HACE" (Heterogeneous, Autonomous, Complex y Evolving), con el cual buscan describir las características de la revolución de Big Data. El teorema plantea la existencia de un gran volumen de datos heterogéneos y provenientes de fuentes autónomas con control distribuido y descentralizado, y que trata de explorar relaciones complejas y cambiantes entre los datos. Los autores plantean que hay un gran desafío para descubrir conocimiento útil desde Big Data. La heterogeneidad se refiere a los diferentes tipos de representaciones para los mismos individuos, y la diversidad de características se refiere a la variedad a la hora de representar cada observación particular. Las fuentes de datos autónomas con control distribuido y descentralizado son, según los autores, la principal característica de las aplicaciones de Big Data. Al ser autónomas, cada fuente de datos tiene la capacidad de generar y recopilar información sin la participación de un ente de control centralizado. Se plantea, además, que un marco de trabajo para el procesamiento de Big Data presenta ciertos desafíos de investigación, los cuales se pueden reunir en una estructura de tres niveles. La parte central, la "plataforma de minería de Big Data" (nivel I), que se enfoca en el acceso a los datos de bajo nivel y computación. Los desafíos en el intercambio de información y la privacidad, los dominios de aplicación de Big Data y el conocimiento forman el nivel II, que se concentra en la semántica de alto nivel, las aplicaciones de dominio de conocimiento y los problemas de privacidad del usuario. Ya en el nivel III se presentan los desafíos en los actuales algoritmos de minería.

Cabe resaltar, que los documentos considerados tronco, para este caso, corresponden a revisiones del estado del arte en Big Data. En la Tabla 2 se presenta una síntesis de los mismos.

En la exploración se encontró que el termino Big Data ha tenido gran acogida en la comunidad, representado esto en el surgimiento de tecnologías, técnicas y enfoques.

Sin embargo, se presenta aún una marcada tendencia hacia los aportes de tipo conceptual, son pocos los resultados y hallazgos que permitan realmente vislumbrar de forma tangible sus beneficios frente a otras tendencias o tecnologías tradicionales. Los trabajos se concentran, en su gran mayoría, en asociar Big Data a grandes volúmenes de datos o a la distribución de procesamiento. En el primer caso, no es claro cuál es la cantidad de datos que permite esta calificación; y para el segundo, no hay coincidencia en determinar para qué tipo de datos el procesamiento distribuido consigue mejores resultados. La volatilidad y variabilidad aún no reciben la atención necesaria. Con lo anterior, se ratifica que existen numerosos vacíos conceptuales y tecnológicos en los cuales se pueden plantear trabajos investigativos y prácticos.

Tabla 2. Síntesis de los documentos tronco. Fuente: Autores

Autores - Año	Título	Tipo de documento	Referencias revisadas	Síntesis del documento
Chen, P. Zhang, C. 2014	Data-intensive applications, challenges, techniques and technologies: A survey on Big Data	Artículo en revista científica	207	Presentan Big Data como el inicio de una era de innovación, competitividad, productividad y revolución científica. El principal aporte del documento se encuentra en el detalle que realizan de diferentes herramientas y técnicas potenciales para resolver los problemas de Big Data desde cada una de sus fases
Chen, M. Mao, S. Liu, Y. 2014	Big Data: A Survey	Artículo en revista científica	156	La revisión se concentra en las cuatro fases de valor de Big Data: generación, adquisición, almacenamiento y análisis de datos. Introduciendo en cada fase una exploración general, técnicas y últimos avances. También se presentan a aplicaciones de Big Data en campos como el empresarial, salud y medicina, internet de las cosas (IoT) y redes sociales
Chen, J. et al 2013	Big data challenge: a data management perspective	Artículo en revista científica	36	Se hace una revisión corta enfocada a los cuatro pasos, que según los autores y según una perspectiva de administración de datos, se deben considerar en Big Data, estos pasos son: integración, reducción, consulta e indexación y análisis y minería. Se clasifican las fuentes de datos en: contenidos generados por usuarios, datos transaccionales, datos científicos, datos web y grafos de datos
Jin, X. et al 2015	Significance and Challenges of Big Data Research	Artículo en revista científica	21	Se hace una breve revisión de las oportunidades e importancia de Big data, pero se enfatiza en cómo hacer un proyecto de Big Data exitoso. Para ello, se da una serie de recomendaciones, como tener claridad en los requerimientos, encontrar el centro de los datos a procesar, caracterizar el comportamiento y propiedades del problema, ya que cada dominio de datos es específico
Wu, X. et al 2014	Data Mining with Big Data	Artículo en revista científica	57	Consideran Big Data como una tendencia emergente y la minería de datos sobre Big Data como una necesidad en todos los campos de la ciencia y la ingeniería. Los autores consideran que las tecnologías de Big Data pueden permitir la detección de información más relevante y precisa para entender la sociedad en tiempo real

3. TRATAMIENTO DE BIG DATA

Como se ha venido comentando, el tratamiento de Big Data ha exigido el desarrollo de soluciones computacionales que permitan afrontar las necesidades y retos que traen consigo los grandes volúmenes de datos, su variedad de fuentes y la velocidad con que se generan.

A continuación, se da una breve descripción de algunas tecnologías y técnicas de Big Data, los artículos referenciados en esta sección comprenden algunos de los documentos "ramas" encontrados en la exploración con la herramienta ToS, otros hacen parte de la búsqueda inicial en Scopus y otros son fuentes adicionales consultadas por los autores para ampliar el tema y cubrir el objetivo de brindar una visión del estado del arte referente a la temática abordada.

3.1 Tecnologías Big Data

Como tecnologías de Big Data se clasifican aquellas que dan soporte a la captura, transformación, procesamiento y análisis de los datos, ya sean estructurados, semiestructurados o no estructurados. Seguidamente, en la Fig. 5, se muestran las

tecnologías de Big Data que se revisarán en este documento. Se decide presentar estas tecnologías ya que son software de libre uso y que permite la generación de soluciones de Big Data de acuerdo con las necesidades particulares de un dominio de datos u organización. Cabe aclarar que existen un mayor número de tecnologías que soportan Big Data, tanto libres como propietarias, pero para efectos de este documento se ha acotado de acuerdo con lo anteriormente expuesto y tomando las tecnologías que dieron las bases iniciales al ecosistema Big Data.

Hadoop

Hadoop es una librería de Apache definida como un framework que permite hacer procesamiento de datos distribuido sobre volúmenes de datos de considerable tamaño sobre clúster. Está diseñado pensando en brindar poder de escalamiento desde un par de servidores hasta cientos de máquinas o nodos, las cuales manejan almacenamiento y procesamiento local [17].

Hadoop cuenta con dos componentes principales, el HDFS, sistema de archivos distribuidos que permite distribuir los ficheros en distintas máquinas y MapReduce, framework que permite al desarrollador aislarse de la programación paralela, permite ejecutar programas escritos en lenguajes de programación conocidos (p.e Java) en el clúster de Haddop. El HDFS cuenta con tres pilares básicos. Namenode, se ocupa del control de acceso y tiene la información sobre la distribución de datos en el resto de nodos. Datanodes, son los encargados de ejecutar el cómputo, es decir, las funciones Map y Reduce, sobre los datos almacenados de manera local en cada uno de dichos nodos. Jobtracker, este nodo se encarga de las tareas y ejerce el control sobre la ejecución del proceso de MapReduce. Además, el HDFS cuenta con las siguientes características fundamentales:

- Tolerancia a fallos
- Acceso a datos en streaming
- Facilidad para el trabajo
- Modelo sencillo de coherencia
- Portabilidad de convivencia

Varios trabajos donde se ha tomado Hadoop como base y se ha potencializado algunas de sus características o se ha fusionado con otra herramienta o tecnología. Ejemplos de esto se pueden encontrar en [18]–[20].

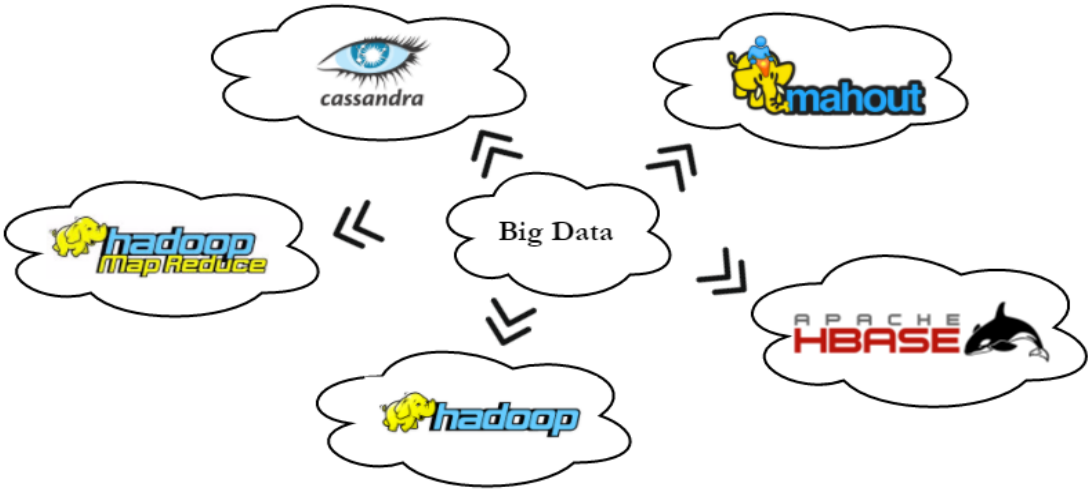

Fig. 5. Tecnologías Big Data. Fuente: Autores.

MapReduce

MapReduce es un modelo de programación que se ha asociado también a la implementación de estrategias de procesamiento de grandes conjuntos de datos que puede ser aplicado a una gran variedad de tareas del mundo real [8]. Este modelo de programación fue utilizado inicialmente por Google para resolver el problema de ranking de páginas ("Page Rank"). El modelo se basa en los siguientes conceptos: iteraciones sobre los datos de entrada, construcción de los pares clave-valor a partir de cada pieza de entrada, agrupación de los valores intermedios de acuerdo con las claves, iteración sobre los grupos resultantes y reducción de cada grupo [21]. En la Fig. 6 se presenta el esquema de un proceso MapReduce y seguidamente, se hace una descripción de cada una de fases que involucra.

Mapeo: se aplica en paralelo para cada uno de los ítems en la entrada de datos. Por medio de la tarea de mapeo (Map) a cada llamada se asignará una lista de pares clave-valor (key-value). Por cada clave generada se crea un grupo, el framework agrupa todos los pares con la misma clave extraídos de todas las listas tratadas.

Reducción: se aplica en paralelo para el grupo asociado a una clave. El resultado es la producción de una colección de valores para cada dominio.

Distribución y ordenamiento: tiene dos misiones, por una parte, se encarga de ordenar por clave todos los resultados emitidos por los mapper y por otra parte recoge todos los valores intermedios pertenecientes a una clave para combinarlos en una lista asociada a ella.

Las características de MapReduce se resumen a continuación:

- Distribución y paralelización automáticas
- Tolerancia a fallos y a redundancias
- Transparencia
- Escalabilidad horizontal
- Localización de los datos
- Herramientas de monitorización

Este paradigma ha sido implementado en numerosas aplicaciones, algunos ejemplos se pueden encontrar en los siguientes documentos, todos se caracterizan por el uso de MapReduce como base de su implementación. En [22] se presenta una herramienta para el análisis de producción mediante simulaciones a gran escala, en [23] se introduce una estrategia para la extracción de patrones significativos a partir de textos de fecha y hora, por su parte en [24] se muestra la implementación paralela de redes neuronales multicapa sobre cloud computing clusters, en [25] se evalúa MapReduce para la realización de minería de texto en información biomédica y en [26] se reporta la utilización del paradigma para la construcción de un sistema de recomendación de artículos considerado como un problema dentro del alcance de las soluciones de Big Data. Como se puede apreciar, son variados los campos de dominio y problemáticas que pueden ser abordadas mediante la adopción de MapReduce para la simplificación de complejos.

HBase

Es una base de datos Hadoop, distribuida y escalable. HBase ha sido desarrollada por Apache y se recomienda su uso cuando se necesita acceso a lectura y escritura de datos en tiempo real sobre Big Data. El objetivo de HBase es el almacenamiento de tablas de gran tamaño, con billones de filas por millones de columnas [27]. Esta base de datos no relacional fue modelada después de Bigtable de Google [28], es open source, distribuida y versionada. HBase provee capacidades similares a Bigtable sobre Hadoop y HDFS. Algunas de sus principales características son:

- Escalabilidad modular y linear
- Estricta consistencia de lectura y escritura
- Facilidad de uso de la API de Java para el acceso de clientes
- Bloqueo de la caché para consultas en tiempo real
- Soporte de para exportar métricas a través del subsistema de métricas de Hadoop

Cassandra

La base de datos Cassandra, propiedad de Apache, brinda escalabilidad y alta disponibilidad sin comprometer el rendimiento. Se considera una plataforma ideal para tratar problemas de datos críticos, puesto que cuenta con escalabilidad lineal y la tolerancia a fallos en el hardware o en la infraestructura en la nube [29]. Cassandra ofrece un modelo de datos que cuenta con comodidad para la indexación de columnas, soporte a la desnormalización y materialización a las vistas y un poderoso almacenamiento en caché integrado. Es un sistema de almacenamiento distribuido con un modelo de datos que soporta un control dinámico sobre el diseño y el formato de los datos [30]. Algunos de los principales atributos de Cassandra son:

- Tolerancia a fallos, por medio de la replicación automática de los datos en múltiples nodos
- Descentralización, uso de muchos nodos idénticos, sin cuellos de botella en la res
- Durable, diseñada para evitar la pérdida de datos
- Elasticidad, capacidad de añadir nuevas máquinas para aumentar el rendimiento de lectura y escritura

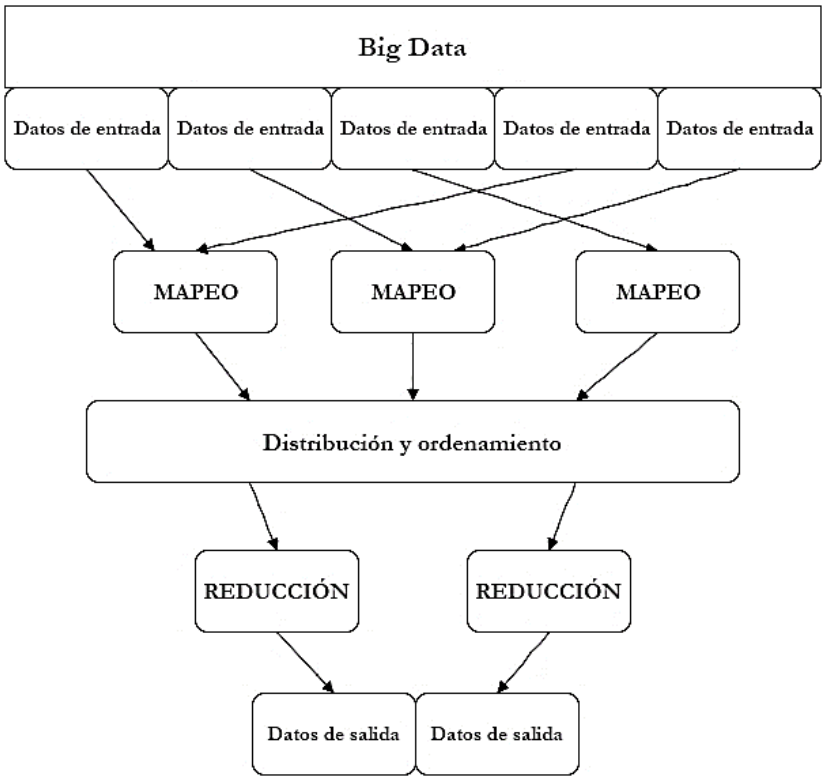

Fig. 6. Esquema general de los procesos MapReduce. Fuente: Autores

Mahout

Mahout es un proyecto de Apache que tiene como objetivo ofrecer un ambiente para la creación rápida de aplicaciones de aprendizaje máquina escalables y eficientes [31]. Mahout ofrece una suite de algoritmos para clustering, categorización, filtrado colaborativo, clasificación y programación evolutiva. Algunas de sus principales aplicaciones prácticas se enmarcan en la realización de clúster de documentos, recomendaciones y organización de contenidos [32]. El machine learning o aprendizaje máquina es el trasfondo principal de Mahout y corresponde a un subcampo de la inteligencia artificial que se centra en el mejoramiento de procesamientos computacionales a partir del análisis de experiencias previas. Mahout desde su aparición ha seguido siendo un proyecto en desarrollo, crecimiento y expansión. Grant Ingersoll en [33] presenta una descripción de algunos de los más recientes algoritmos implementados en Mahout, resumiéndolos en la Tabla 3, la cual se presenta a continuación.

Técnicas Big Data

En cuanto a técnicas de Big Data, se dará una breve introducción, cabe aclarar que existen diferentes clasificaciones y que muchas de estas técnicas se aplican tanto en soluciones Big Data como en otros enfoques. En [34] se presenta una clasificación de las técnicas de Big data en técnicas estadísticas, métodos de optimización, minería de datos, técnicas de machine learning (aprendizaje máquina), técnicas de clasificación y Clustering y técnicas de análisis y regresión. Para efectos de este documento se describen, sin entrar en detalle, la minería de datos, el aprendizaje máquina, el reconocimiento de patrones, los algoritmos genéticos y las reglas de asociación**.**

Tabla 3. Algunos algoritmos en Mahout. Fuente: Tomado y adaptado de [33].

Algoritmo	Descripción breve	Aplicaciones
Regresión logística, resuelta por gradiente estocástico descendiente (SGD)	Clasificador brillante, rápido, simple y secuencial, capaz de aprendizaje on-line en entornos exigentes	Recomendación de publicidad, clasificación de textos
Modelos ocultos de Markov (HMM)	Implementaciones secuenciales y paralelas del algoritmo clásico de clasificación diseñado para modelar procesos del mundo real cuando el proceso de generación subyacente es desconocido	Etiquetado de texto, reconocimiento del discurso
Descomposición de valor singular (SVD)	Diseñado para reducir el ruido en matrices grandes, haciendo con esto que sean más pequeñas y que sea más fácil trabajar con ellas	Clasificación para realizar selección de recursos automáticamente
Almacenamiento en clúster Dirichlet	Enfoque de almacenamiento en clúster basado en modelo, que determina la propiedad con base en si los datos se ajustan al modelo subyacente	Almacenamiento en clúster para datos con sobreposición o jerarquía
Almacenamiento en clúster espectral	Es una familia de enfoques similares que usa un enfoque basado en gráficas para determinar la membresía a clúster	Almacenamiento en clúster para conjuntos de datos grandes y no vistos
Almacenamiento en clúster Minhash	Utiliza una estrategia de hash para agrupar elementos similares, produciendo así clústeres	Clúster
Numerosas mejoras de recomendador	Co-ocurrencia distribuida, SVD, mínimos cuadrados alternantes	Recomendaciones en sitios de citas, e-commerce, recomendaciones de películas o de libros
Colocaciones	Implementación de colocación reducida por correlacionamiento	Encontrando frases estadísticamente interesantes en texto

Minería de Datos

La minería de datos (data mining - DM) se puede definir como el proceso de extracción de conocimiento a partir de cúmulos de datos. Se suele utilizar el término minería de datos como sinónimo de descubrimiento de conocimiento, pero realmente no son sinónimos, la minería de datos es solo un paso en el proceso de descubrimiento de conocimiento [35]. La minería de datos nace de la necesidad de conocer información útil a partir de los bases de datos o Datawarehouse, con el crecimiento de los datos disponibles, la inteligencia de negocios tuvo que dar paso a la aplicación de la minería de datos en soluciones empresariales y comerciales, puesto que de esta manera se permite el descubrimiento automático o semiautomático de información relevante a partir de estos cúmulos de datos. En las ciencias y la ingeniería existe un amplio rango de problemas y dominios de aplicación para la minería de datos [36]. Se encuentran soluciones a partir de minería de datos para problemas de los campos de mercadeo, comercio, salud, predicción, transporte, meteorología, entre otros.

Machine learning

Aprendizaje máquina es un área de investigación bastante reconocida en las ciencias de la computación, principalmente comprende el descubrimiento de modelos, patrones y regularidades en los datos [37]. El aprendizaje máquina puede ser visto desde dos enfoques, los simbólicos y los estadísticos. Los primeros trabajan aprendizaje inductivo de descripciones simbólicas, mientras que los segundos se centran en los métodos de reconocimiento de patrones o en la estadística. En los últimos años, el uso del aprendizaje máquina se ha extendido con rapidez [38], se ven aplicaciones en dominios como detección de fraudes, sistemas de recomendación [39], detección de spam [40], predicciones financieras [41], comercio y mercadeo [42], [43], entre otros. Los algoritmos de aprendizaje máquina se clasifican en supervisados y no supervisados.

Reconocimiento de patrones

El reconocimiento de patrones (Pattern Recognition) es una técnica que se aplica principalmente en procesos de ingeniería, computación y matemáticas que tiene como objetivo extraer información, a partir de un cúmulo de datos, que brinde la posibilidad de establecer propiedades o relaciones entre estos datos. En el procesamiento de patrones generalmente se usan algoritmos de optimización, puesto que su intención es hallar una mejor solución respecto a un criterio definido, teniendo en cuenta que un proceso de optimización es una situación que requiere elegir desde un conjunto de alternativas, la que lleve al fin requerido con el costo mínimo [44].

Algoritmos genéticos

Los algoritmos genéticos (genetic algorithms - GA) son una técnica aplicada en la ingeniería computacional, pero que parte de la concepción biológica de la genética. Estos algoritmos comprenden un enfoque que busca dar solución a diversos problemas matemáticos intangibles que no han podido tener solución desde otros enfoques matemáticos tradicionales [45]. Los algoritmos genéticos utilizan también operaciones genéticas como la mutación, recombinación y cruce. En [46] se definen los algoritmos genéticos como métodos de búsqueda estocásticos diseñados para explorar problemas complejos, con el fin de encontrar una solución óptima, generalmente usando información propia del problema como guía de la búsqueda. Los algoritmos genéticos se enmarcan dentro de los Algoritmos Evolutivos (Evolutionay Algorithms – EA) siendo uno de los componentes más importantes, junto con la programación genética y las estrategias evolutivas. En [47] se presentan como componen-

tes esenciales de los algoritmos genéticos los siguientes:

- Estrategia de codificación que determina la forma en que se representará la solución en forma de cromosomas
- Población de cromosomas o individuos
- Mecanismo para la evaluación de cada cromosoma
- Procedimiento de selección/reproducción
- Operadores genéticos: cruce, mutación
- Probabilidades para los operadores genéticos
- Un criterio de finalización

Aprendizaje de reglas de asociación

El aprendizaje de reglas de asociación (Association rule learning), es un método para encontrar las relaciones entre variables en grandes bases de datos, su objetivo es identificar reglas usando algunas medidas de relación de intereses, por ejemplo, en el caso de las redes sociales, se trataría de revisar las personas que posiblemente le interesarían seguir a otras dependiendo de sus amistades o seguidores. En el caso de tiendas de productos, podría ser la revisión de los productos que se compran juntos con frecuencia para sugerirlos a un cliente que adquiera uno de los productos relacionados. Algunas aplicaciones del aprendizaje de reglas de asociación se encuentran en [48]–[50].

4. TENDENCIAS Y RETOS EN BIG DATA

En esta sección se presenta la exploración de algunos trabajos que presentan una visión general de las tendencias y enfoques en el desarrollo de investigaciones en el campo de Big Data.

En [51] se muestra una revisión del estado del arte en cuanto a sistemas de almacenamiento para grandes volúmenes de datos, incluyendo un comparativo entre los Sistemas de Administración de Bases de Datos (DBMS) tradicionales y los nuevos enfoques NoSQL (Not Only SQL). En el trabajo se considera la necesidad de que estos sistemas sigan garantizando características como: escalabilidad, fiabilidad, durabilidad, tiempos de respuesta, interfaces de consulta, esquemas de particionamiento y estructura o carencia de esta. Se describen los modelos de almacenamiento NoSQL: depósitos llave-valor, basado en documentos, tabular y orientados a grafos. Los autores afirman que los sistemas NoSQL se adecuan a casos en los que se necesita atender a muchos usuarios sin perder rendimiento, como puede pasar en el caso de las redes sociales. Por su parte, recomiendan los sistemas de bases de datos relacionales cuando se trata de garantizar integridad referencial, se requiere el uso de conexiones entre servidores y clientes, consultas arbitrarias, estandarización, herramientas de análisis y pruebas de rendimiento.

En [52] presentan la revisión de varios aspectos relacionados con Big Data, tales como contenido, alcance, métodos, ventajas, desafíos, ejemplos y privacidad de los datos. La revisión realizada por los autores muestra que incluso con las herramientas y técnicas disponibles en la actualidad y la literatura al respecto, existen muchos puntos a ser considerados, desarrollados, mejorados y analizados. Es claro que la cantidad de datos ha ido en aumento, lo cual exige que también las técnicas de análisis y tratamiento de datos se hagan más competitivas, el reto no es solo para recoger y gestionar el gran volumen y diferentes tipos de datos, sino también para extraer valor significativo de estos. Se presentan como las principales barreras para la implementación de analíticas de Big Data: la carencia de expertos en el tema de Big Data, el costo, el manejo de la privacidad en la manipulación de los datos, la dificultad en el diseño de sistemas de análisis, la falta de software que soporte grandes bases de datos permitiendo análisis con tiempos de procesamiento rápido, los problemas de escalabilidad, la incapacidad de hacer

que Big Data sea utilizable por usuarios finales, la falta de rapidez en la carga de datos con los sistemas de gestión de bases de datos actuales y la ausencia de un modelo de negocio convincente y rentable en torno al tema.

En [4] los autores analizan algunas tecnologías relacionadas con Big Data como computación en la nube, internet de las cosas, centros de datos y Hadoop. También se enfocan en la discusión de los desafíos técnicos y adelantos en cada una de las fases de Big Data: generación, adquisición, almacenamiento y análisis de datos. El análisis de Big Data tiene que afrontar muchos desafíos, se requieren considerables esfuerzos investigativos, los cuales se pueden agrupar en los problemas abiertos presentados en la Fig. 7.

En [53] se hace énfasis en la utilización de técnicas de Inteligencia Artificial (IA) para facilitar la captura y estructuración de grandes volúmenes de datos y también cómo se han implementado para el análisis de estos. Se presentan algunas preocupaciones respecto a la integración de IA con Big Data, que no se resuelven solo con pensar en la distribución y paralelización, sino que requieren otros análisis. Las técnicas de IA para el tratamiento de Big Data permiten la delegación de tareas complejas de reconocimiento de patrones, aprendizaje y otras tareas basadas en enfoques computacionales, la IA contribuye a la velocidad en la manipulación de los datos, facilitando la toma de decisiones rápidas. Por ejemplo, muchas operaciones de la bolsa son hechas por sistemas basados en IA en lugar de personas, la velocidad de las operaciones puede aumentar y una transacción puede conducir a otras. Existen varios problemas emergentes asociados a la IA y Big Data, en primer lugar, la naturaleza de algunos de los algoritmos de machine-learning son difícilmente usados en ambientes como MapReduce, por lo cual se requiere de su adaptación. En segundo lugar, Big Data trae consigo datos "sucios", con errores potenciales, incompletos o de diferente precisión, la IA puede ser usada para identificar y limpiar estos datos sucios. En tercer lugar, la visualización de los datos, con la IA se puede lograr incluir la captura de capacidades de visualización de conocimiento para facilitar el análisis de datos, un enfoque es crear aplicaciones inteligentes de visualización para determinados tipos de datos. En cuarto lugar, ya que las tecnologías de almacenamiento evolucionan, es cada vez más factible proporcionar a los usuarios, casi en tiempo real, análisis de bases de datos más grandes, lo que acelera las capacidades de toma de decisiones.

En [54] presentan una descripción consolidada del concepto de Big Data, partiendo de las definiciones dadas por profesionales y académicos del campo, como se ve en la Fig. 8. Sin embargo, el artículo se concentra en revisar los métodos de análisis usados para Big Data. Se destaca que Big Data no tiene un verdadero sentido si solo se trata de un gran cúmulo de datos, su valor potencial se desbloquea solo cuando estos datos son aprovechados para impulsar la toma de decisiones. Para ello es necesario mover y dar significado a los datos, esto se puede hacer por medio de dos subprocesos principales: la gestión y análisis de datos. La gestión de datos implica procesos y tecnologías de apoyo para adquirir, almacenar, preparar y recuperar los datos para su análisis. El análisis, por su parte, se refiere a las técnicas utilizadas para adquirir inteligencia a partir de Big Data.

Investigación teórica	Desarrollo tecnológico	Implicaciones prácticas	Seguridad de datos
• ***Problemas fundamentales:*** Big Data no está formal ni estructuralmente definido y los modelos existentes no se verifican en sentido estricto. • ***Estandarización:*** se requiere un sistema de evaluación de la calidad de los datos, un estándar de pre-procesamiento, simplificación y detección. • ***Evolución de los modos de computación:*** la transferencia de datos se ha convertido en un cuello de botella, esto exige el desarrollo de nuevos algoritmos de computación intensiva para afrontar los datos intensivos.	• ***Formato de conversión:*** la heterogeneidad de los datos es una característica de Big Data por ello al contar con un formato de conversión más eficiente se podrá extraer más valor • ***Transferencia:*** este aspecto en Big Data suele ser muy costoso pero inevitable e involucra la generación, adquisición, trasmisión, almacenamiento y otras transformaciones de los datos • ***Rendimiento en tiempo real:*** definir un ciclo de vida y computar la tasa de depreciación de los datos y construir un modelo de computación en tiempo real influirán en los resultados de análisis de Big Data. • ***Procesamiento:*** se involucran problemas como re-utilización de los datos, re-organización y el fenómeno del "data exhaust" que trae consigo muchos datos erróneos en la adquisición.	• ***Administración:*** se requieren muchos esfuerzos para la generación de nuevos modelos de almacenamiento, integración de datos con múltiples estructuras y gestión de datos distribuidos • ***Búsqueda, minería y análisis:*** es necesario contar con algoritmos para búsqueda distribuida, sistemas de recomendación masiva, minería de datos en tiempo real, minería de imágenes y de texto, entre otros • ***Integración y procedencia:*** es un desafío ya que se tienen múltiples patrones de datos y un gran número de datos redundantes, así como también los datos proceden de varios datasets • ***Aplicaciones:*** el estudio de Big Data está en una etapa inicial, por lo cual la necesidad de aplicaciones en diferentes ciencias y campos es inminente.	• ***Privacidad de datos:*** esto incluye la protección de los datos personales durante la adquisición y durante el almacenamiento, trasmisión y uso; se requiere de mayor claridad y reglamentación en este aspecto • ***Calidad de datos:*** la baja calidad de los datos se ve reflejada en una pobre usabilidad de los mismos. La calidad de los datos se refleja en la precisión, integridad, redundancia y consistencia • ***Mecanismos de seguridad:*** se deben desarrollar métodos de encriptación capaces de abordar la diversidad y gran escala de Big Data • ***Seguridad de la información en las aplicaciones de Big Data:*** se presentan oportunidades para el desarrollo de nuevos mecanismos de seguridad informática, en sistemas de detección de intrusos, entre otros.

Fig. 7. Problemas abiertos en Big Data. Fuente: Elaborado a partir de [4].

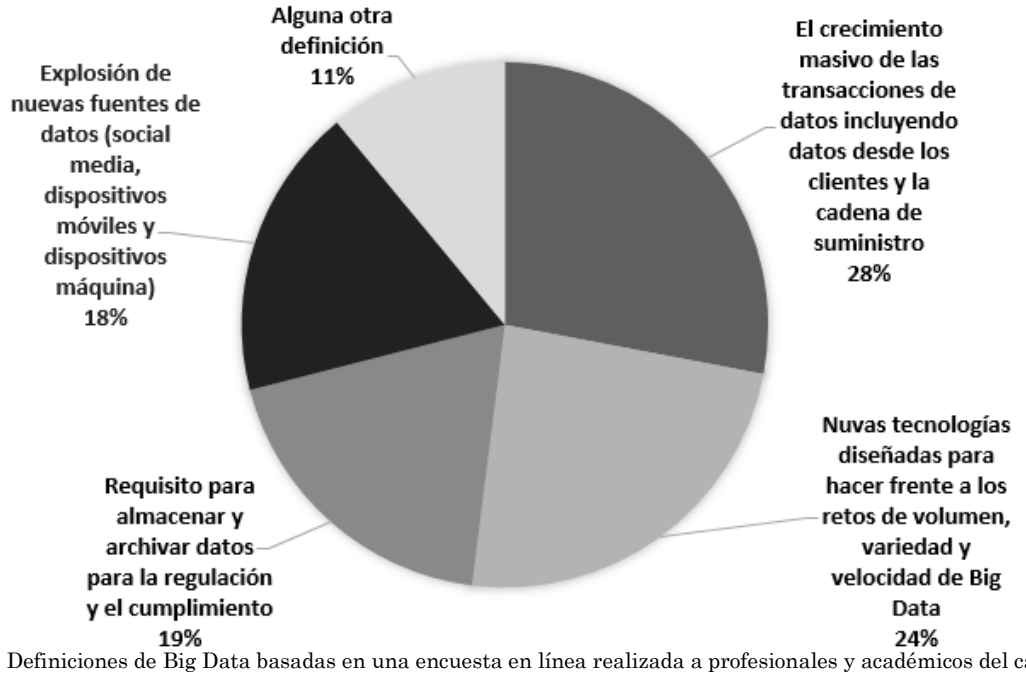

Fig. 8. Definiciones de Big Data basadas en una encuesta en línea realizada a profesionales y académicos del campo. Fuente: Adaptado de [54].

Los métodos de análisis de Big Data a los que hacen referencia los autores se enfocan en los tipos de datos tratados, por lo que se describen analíticas de texto, analíticas de audio, analíticas de social media y analíticas predictivas. Estas últimas, las predictivas, se basan principalmente en los métodos estadísticos, sin embargo, hay algunos factores que requieren el desarrollo de nuevos métodos estadísticos para Big Data. En primer lugar, los métodos estadísticos convencionales se concentran en una pequeña muestra de la población y los resultados se generalizan a toda la población, pero para el caso de Big Data, las muestras son enormes y representan la mayoría o la totalidad de la población. En segundo lugar, en términos de eficiencia de cómputo muchos métodos convencionales para muestras pequeñas no se logran escalar hasta Big Data. El tercer factor corresponde a algunos de los rasgos distintivos de Big Data: la heterogeneidad, la acumulación de ruido, las falsas correlaciones y la endogeneidad incidental.

Los trabajos revisados en esta sección muestran el estado actual del enfoque Big Data y de las tendencias que giran en torno a este, incluyendo el planteamiento de varios campos de investigación que se encuentran abiertos, principalmente relacionados con la optimización de los sistemas de almacenamiento para grandes volúmenes de datos, los cuales todavía presentan falencias en cuanto al tratamiento de distintos tipos de datos a la vez, la optimización de consultas complejas y operaciones sobre los datos.

También se ve la diversidad de planteamientos que presentan los autores en cuanto al concepto de Big Data y las características que este debe atender. Es claro que el tema ha tomado un carácter de moda mundial y que se ha dejado de asociar solo a la característica de gran tamaño. Se ven también posibilidades de explorar la aplicación de Big Data a nuevos dominios de datos, ya que actualmente se han concentrado en social media, medicina, bioinformática y seguridad, principalmente.

5. CONCLUSIONES

Teniendo en cuenta los objetivos planteados para la realización de esta explora-

ción, el principal aporte logrado con el artículo es la caracterización en un solo documento de trabajos, enfoques y herramientas recientes relacionadas con el término en boga Big Data, que puede servir como referente para trabajos posteriores y para la consulta de investigadores que deseen adelantar trabajos en el marco de los campos de estudio abiertos que se dejan ver tras la exploración presentada.

Este artículo se trazó como objetivo mostrar algunos trabajos desarrollados entorno a la temática y describir tecnologías y técnicas de Big Data, notándose que siguen siendo materia de investigación y discusión, generando la posibilidad de proponer alternativas y modelos basados en la táctica de divide y vencerás.

Las tecnologías asociadas al enfoque de Big Data ya han comenzado a tomar madurez y se vislumbran grandes oportunidades y retos en su utilización, optimización y adaptación a diferentes dominios de datos. Sin embargo, ya se encuentran resultados que muestran sus beneficios en aspectos como la reducción de tiempos, optimización de recursos y mayor flexibilidad. Existe una estrecha relación entre diferentes métodos y tecnologías para la construcción de soluciones que integren las capacidades de cada uno de estos y las potencien en nuevas propuestas.

Big Data no trata solo de grandes volúmenes de datos, sino que incluye otras dimensiones significativas en el tratamiento de datos, como son la variedad, velocidad y veracidad. No obstante, una implementación de Big Data requiere altos costos en expertos, mayor tiempo de adaptación tecnológica, dificultad para implementar nuevos análisis y percepción limitada. Big Data no busca sustituir a los sistemas tradicionales, sino construir una nueva tendencia donde se construyan arquitecturas de sistemas que permitan manejar todas las peticiones. Y ya ha logrado incentivar en la comunidad académica y comercial el desarrollo de tecnologías de apoyo que toman los paradigmas base y los emplean en la construcción de soluciones particularizadas a problemas de entornos de investigación y producción reales.

6. AGRADECIMIENTOS

Este trabajo está en el marco del proyecto titulado: "Consolidación de las líneas de investigación del Grupo de Investigación en Ambientes Inteligentes Adaptativos GAIA" con código 32059, en el marco de la convocatoria interna de investigación de la Facultad de Administración 2015, para la formulación y ejecución de proyectos de consolidación y/o fortalecimiento de los grupos de investigación de la Universidad Nacional de Colombia, sede Manizales.

7. REFERENCIAS

[1] K.C. Li, H. Jiang, L. T. Yang, and A. Cuzzocrea, *Big Data: Algorithms, Analytics, and Applications*, Chapman &. CRC Press, 2015.

[2] H. Mohanty, P. Bhuyan, and D. Chenthati, *Big Data: A Primer*, vol. 11. Springer, 2015.

[3] W. M. P. van der Aalst, "Data Scientist: The Engineer of the Future," in *Enterprise Interoperability VI*, no. 7, K. Mertins, F. Bénaben, R. Poler, and J.-P. Bourrières, Eds. Springer International Publishing, 2014, pp. 13–26.

[4] M. Chen, S. Mao, and Y. Liu, "Big Data: A Survey," *Mob. Networks Appl.*, vol. 19, no. 2, pp. 171–209, Apr. 2014.

[5] L. A. Montenegro Mora, "¿Cómo elaborar un artículo de revisión?," San Juan de Pasto, Nariño, Colombia, 2013.

[6] Elsevier, "Scopus The largest database of peer-reviewed literature," *Scopus Elsevier*. 2016. [Online]. Available: https://www.elsevier.com/solutions/scopus.

[7] S. Robledo Giraldo, G. Osorio Zuluaga, and C. López Espinosa, "Networking en pequeña empresa: una revisión bibliográfica utilizando la teoría de grafos," *Rev. Vínculos*, vol. 11, no. 2, pp. 6–16, 2014.

[8] J. Dean and S. Ghemawat, "MapReduce," *Commun. ACM*, vol. 51, no. 1, p. 107, Jan. 2008.

[9] M. Armbrust, I. Stoica, M. Zaharia, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, and A.

Rabkin, "A view of cloud computing," *Commun. ACM*, vol. 53, no. 4, p. 50, Apr. 2010.

[10] P. Zikopoulos and C. Eaton, *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*, 1st ed. McGraw-Hill Osborne Media, 2011.

[11] T. White, *Hadoop: The Definitive Guide*, 2nd ed. United States of America: O'Reilly Media, Inc, 2010.

[12] D. Bollier, "The Promise and Peril of Big Data," Washington, DC, 2010.

[13] C. L. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," *Inf. Sci. (Ny).*, vol. 275, pp. 314–347, 2014.

[14] J. Chen, Y. Chen, X. Du, C. Li, J. Lu, S. Zhao, and X. Zhou, "Big data challenge: a data management perspective," *Front. Comput. Sci.*, vol. 7, no. 2, pp. 157–164, Apr. 2013.

[15] X. Jin, B. W. Wah, X. Cheng, and Y. Wang, "Significance and Challenges of Big Data Research," *Big Data Res.*, vol. 2, no. 2, pp. 59–64, Jun. 2015.

[16] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding, "Data mining with big data," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 97–107, Jan. 2014.

[17] T. A. S. Foundation, "Welcome to Apache™ Hadoop®!," *hadoop*. 2016. [Online]. Available: http://hadoop.apache.org/

[18] M. Klein, R. Sharma, C. H. Bohrer, C. M. Avelis, and E. Roberts, "Biospark: scalable analysis of large numerical datasets from biological simulations and experiments using Hadoop and Spark," *Bioinformatics*, vol. 33, no. 2, pp. 303–305, Jan. 2017.

[19] A. Aji, F. Wang, H. Vo, R. Lee, Q. Liu, X. Zhang, and J. Saltz, "Hadoop GIS: a high performance spatial data warehousing system over mapreduce," *Proc. VLDB Endow.*, vol. 6, no. 11, pp. 1009–1020, 2013.

[20] A. M. Aly, H. Elmeleegy, Y. Qi, and W. Aref, "Kangaroo," in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining - WSDM '16*, 2016, pp. 397–406.

[21] R. Lämmel, "Google's MapReduce programming model — Revisited," *Sci. Comput. Program.*, vol. 70, no. 1, pp. 1–30, Jan. 2008.

[22] K. Lee, K. Jung, J. Park, and D. Kwon, "ARLS: A MapReduce-based output analysis tool for large-scale simulations," *Adv. Eng. Softw.*, vol. 95, pp. 28–37, May 2016.

[23] J.-D. Wang, "Extracting significant pattern histories from timestamped texts using MapReduce," *J. Supercomput.*, vol. 72, no. 8, pp. 3236–3260, Aug. 2016.

[24] H. Zhang and N. Xiao, "Parallel implementation of multilayered neural networks based on Map-Reduce on cloud computing clusters," *Soft Comput.*, vol. 20, no. 4, pp. 1471–1483, Apr. 2016.

[25] Y. Ji, Y. Tian, F. Shen, and J. Tran, "Experimental Evaluations of MapReduce in Biomedical Text Mining," in *Information Technology: New Generations*, Springer, 2016, pp. 665–675.

[26] S. Singh and N. Ahuja, "Article recommendation system based on keyword using map-reduce," in *2015 Third International Conference on Image Information Processing (ICIIP)*, 2015, pp. 548–550.

[27] T. A. S. Foundation, "Apache HBase," *Apache HBase*. 2016. [Online]. Available: http://hbase.apache.org/

[28] G. C. Deka, "A Survey of Cloud Database Systems," *IT Prof.*, vol. 16, no. 2, pp. 50–57, Mar. 2014.

[29] T. A. S. Foundation, "The Apache Cassandra Project," *Apache Cassandra*. 2015.

[30] E. Dede, B. Sendir, P. Kuzlu, J. Hartog, and M. Govindaraju, "An Evaluation of Cassandra for Hadoop," in *2013 IEEE Sixth International Conference on Cloud Computing*, 2013, vol. 2013, pp. 494–501.

[31] T. A. S. Foundation, "Apache Mahout: Scalable machine learning and data mining," *Apache Mahout*. 2016.

[32] G. Ingersoll, "Introducing Apache Mahout," *IBM developerWorks*. 2009. [Online]. Available: http://www.ibm.com/developerworks/java/library/j-mahout/

[33] G. Ingersoll, "Apache Mahout: Aprendizaje escalable con máquina para todos," *IBM developerWorks*. 2012. [Online]. Available: http://www.ibm.com/developerworks/ssa/library/j-mahout-scaling/

[34] S. M. D. MUJEEB and L. K. NAIDU, "A Relative Study on Big Data Applications and Techniques," *Int. J. Eng. Innov. Technol.*, vol. 4, no. 10, pp. 133–138, 2015.

[35] J. Han, J. Pei, and M. Kamber, "Data mining: concepts and techniques," 3rd ed., E. Inc., Ed. Morgan Kaufmann Publishers, 2011, p. 703.

[36] R. L. Grossman, C. Kamath, P. Kegelmeyer, V. Kumar, and R. Namburu, *Data Mining for Scientific and Engineering Applications*, vol. 2. Boston, MA: Springer US, 2013.

[37] R. S. Michalski, J. G. Carbonell, and T. M. Mitchell, *Machine learning: An artificial intelligence approach*. Springer Science & Business Media, 2013.

[38] P. Domingos, "A few useful things to know about machine learning," *Commun. ACM*, vol. 55, no. 10, p. 78, Oct. 2012.

[39] I. Portugal, P. Alencar, and D. Cowan, "The Use of Machine Learning Algorithms in Recommender Systems: A Systematic Review," *arXiv*, vol. 4, pp. 1–16, Nov. 2015.

[40] M. Crawford, T. M. Khoshgoftaar, J. D. Prusa, A. N. Richter, and H. Al Najada, "Survey of review spam detection using machine learning techniques," *J. Big Data*, vol. 2, no. 1, p. 23, Dec. 2015.

[41] Wei-Yang Lin, Ya-Han Hu, and Chih-Fong Tsai, "Machine Learning in Financial Crisis Prediction: A Survey," *IEEE Trans. Syst. Man, Cybern. Part C (Applications Rev.*, vol. 42, no. 4, pp. 421–436, Jul. 2012.

[42] R. Dash and P. K. Dash, "A hybrid stock trading framework integrating technical analysis with machine learning techniques," *J. Financ. Data Sci.*, vol. 2, no. 1, pp. 42–57, Mar. 2016.

[43] J. Patel, S. Shah, P. Thakkar, and K. Kotecha, "Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques," *Expert Syst. Appl.*, vol. 42, no. 1, pp. 259–268, Jan. 2015.

[44] E. Cuevas, D. Zaldívar, and M. Perez-Cisneros, *Applications of Evolutionary Computation in Image Processing and Pattern Recognition*, 1st ed., vol. 100. Cham: Springer International Publishing, 2016.

[45] K.-F. Man, K. S. TANG, and S. Kwong, *Genetic Algorithms: Concepts and Designs*. Springer Science & Business Media, 2012.

[46] G. Luque and E. Alba, *Parallel Genetic Algorithms: Theory and Real World Applications*, vol. 367. Springer, 2011.

[47] U. Maulik, S. Bandyopadhyay, and A. Mukhopadhyay, *Multiobjective Genetic Algorithms for Clustering: Applications in Data Mining and Bioinformatics*. Springer Science & Business Media, 2011.

[48] A. E. Doub, M. L. Small, A. Levin, K. LeVangie, and T. R. Brick, "Identifying users of traditional and Internet-based resources for meal ideas: An association rule learning approach," *Appetite*, vol. 103, pp. 128–136, Aug. 2016.

[49] H. Sundell, R. Konig, and U. Johansson, "Pragmatic Approach to Association Rule Learning in Real-World Scenarios," in *2015 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2015, pp. 356–361.

[50] R. Sarno, R. D. Dewandono, T. Ahmad, M. F. Naufal, and F. Sinaga, "Hybrid association rule learning and process mining for fraud detection," *IAENG Int. J. Comput. Sci.*, vol. 42, no. 2, pp. 1–14, 2015.

[51] S. Jaramillo Valbuena and J. M. Londoño, "Sistemas para almacenar grandes volúmenes de datos," *Rev. Gerenc. Tecnológica Informática*, vol. 13, no. 37, pp. 17–28, 2015.

[52] S. Sagiroglu and D. Sinanc, "Big data: A review," in *2013 International Conference on Collaboration Technologies and Systems (CTS)*, 2013, pp. 42–47.

[53] D. E. O'Leary, "Artificial Intelligence and Big Data," *IEEE Intell. Syst.*, vol. 28, no. 2, pp. 96–99, Mar. 2013.

[54] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *Int. J. Inf. Manage.*, vol. 35, no. 2, pp. 137–144, Apr. 2015.