Big data, el futuro de las predicciones certeras

Harrison S. Martin Aldana, Julian D. Calderon Rivas & Jose M. Vargas Hidalgo

***Abstract*— Este artículo buscará, principalmente, establecer las razones por las cuales nosotros consideramos que el análisis de datos a través de la utilización de herramientas de Big Data será en el futuro una base sólida para realizar predicciones en muchas áreas de las diferentes temáticas que lo lleguen a requerir, definiremos qué es Big Data y cuál es el proceso que deben llevar los datos desde su recolección hasta el momento en el que se convierten en información relevante para que a través de su análisis se pueda convertir en una base relevante para el establecimiento de patrones de comportamiento en esas áreas.**

Abstract: This article mainly targets to establish the reasons why we consider that the analysis of the data through the Big Data tools is going to be an important base in the future, in order to make predictions in many areas that may need them, we will define what is Big data and the process that should be followed from the collection of the data until the moment that it becomes into relevant information. so by analyzing it, it could transform itself into an important base to establish behavioral patterns on the different areas..

***Palabras Claves*— Análisis predictivo, Big data, tendencia.**

I. INTRODUCCIÓN

Mediante la revisión de bibliografía de diferentes autores, empezaremos por realizar una revisión y posterior análisis de las diferentes definiciones que se pueden encontrar acerca de este concepto, que ha sido materia de investigación más o menos desde la última década de los 90s y que en esta época está buscando consolidarse como una tecnología disruptiva contundente. Sus áreas de aplicación son bastantes, de la mano de tecnologías que también están en un momento de auge tales como el internet de las cosas (IOT por sus siglas en inglés) o las diferentes tecnologías basadas en sensores van a generar una cantidad masiva de datos que van a requerir un almacenamiento, tratamiento y análisis especial. El Big Data se postula como una opción para las necesidades planteadas por estas tecnologías en estos términos y así trabajar de una manera transversal con estos datos para conservarlos y sacar provecho de ellos de la mejor manera a través un análisis predictivo, entre otras.

Desarrollo: [1] Big Data, se trata de aquel conjunto de datos que, por su tamaño ingente, sobrepasa la capacidad de ser gestionado por bases de datos de integración tradicionales. A pesar de que muchos autores consideran esta definición demasiado dispersa, si profundizamos en las características que componen el Big Data, existe un mayor grado de acuerdo en aducir que se fundamenta en el paradigma de la 3 "V" (volumen, variedad y velocidad). El elevado volumen de datos (más de un petabyte) precisa nuevas técnicas de almacenamiento a gran escala y enfoques distintos para recuperar la información; la variedad de las fuentes de datos (texto, audio, vídeo, etc.) hace que las redes relacionales sencillas sean difícilmente aplicables; y, por último, el incesante incremento con que se generan los datos, hace que la velocidad sea un parámetro clave en su manejo. [2] Big Data es un activo de información de gran volumen, alta velocidad y gran variedad que exige formas rentables e innovadoras de procesamiento de información para mejorar el conocimiento y la toma de decisiones. El término Big Data, que en nuestro idioma es un anglicismo, según la RAE el término en español sería Macrodatos, aunque usualmente no se conoce como tal, como podemos observar las definiciones de lo que en realidad es Big Data pueden variar dependiendo el enfoque que le dé el autor, incluso este hecho de definir lo que es Big Data ha sido materia de investigación para ellos. Para realizar una pequeña contextualización vamos a hacer referencia a la Figura 1.

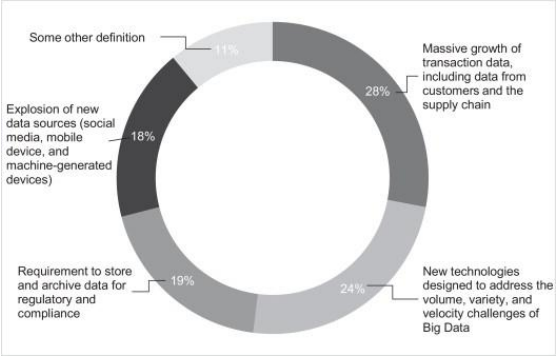


Figura 1. Definición de Big Data basada en una encuesta realizada a 154 ejecutivos a nivel global

Al realizar un análisis de esta figura podemos observar que no hay una definición que tenga, por mucho, más acogida que las otras. Si bien la definición "Crecimiento masivo de datos de transacciones, incluyendo datos de clientes y de la cadena de suministro" es la que tuvo un mayor porcentaje, no fue mucha la diferencia que tuvo con respecto a las otras, lo cual se puede traducir en la variedad de definiciones que se planteó anteriormente. También a través del análisis de todas estas respuestas sabemos que la palabra "tamaño" es lo primero que viene a la cabeza de los encuestados cuando escuchan sobre el término.

Harrison S. Martin Aldana, estudiante de ingeniería de sistemas, harrisons.martina@unilibrebog.edu.co

Julian D. Calderon Rivas, estudiante de ingeniería de sistemas, juliand.calderonr@unilibrebog.edu.co.

Jose M. Vargas Hidalgo, estudiante de ingeniería de sistemas, josem.vargash@unilibrebog.edu.co

Corresponding author: Harrison S. Martin Aldana

La definición que en este contexto encontramos como la más acertada, esto sin el ánimo de afirmar o inferir en ningún momento que las otras no lo sean, es aquella propuesta por TechAmerica Foundation´s Federal Big Data Commision en la que expresa [3] Big data es un término que describe grandes volúmenes de datos de alta velocidad, complejos y variables que requieren técnicas y tecnologías avanzadas para permitir la captura, almacenamiento, distribución, gestión y análisis de la información. También se realiza el planteamiento de las 3V:

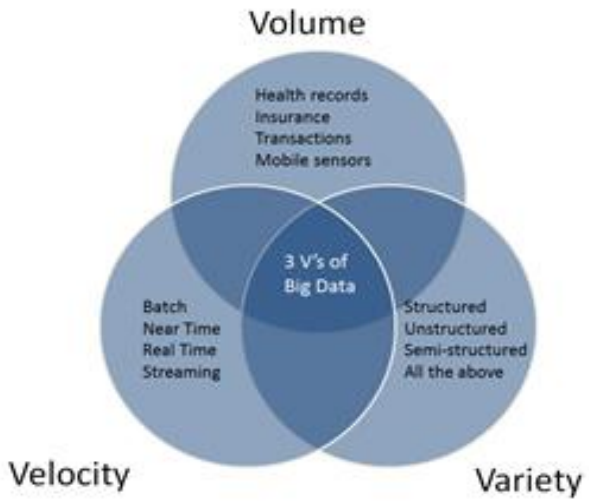


Figura 2. Relación entre las 3 V del Big Data. Tomada del articulo The application of Big Data in medicine: curent implications and future directions 2016 (Christopher Austin).

Volumen: El volumen se refiere a la magnitud que se puede dar a los datos. Los grandes tamaños de datos se reportan en varios terabytes y petabytes. Una encuesta realizada por IBM a mediados de 2012 reveló que poco más de la mitad de los 1144 encuestados consideraron que los conjuntos de datos de más de un terabyte eran considerados Big Data. [4] Un terabyte almacenar tantos datos como cabrían en 1500 CD o 220 DVD, suficiente para almacenar alrededor de 16 millones de fotografías de Facebook. [5] Facebook procesa hasta un millón de fotografías por segundo. Un petabyte es igual a 1024 terabytes. Estimaciones anteriores sugieren que Facebook almacenó 260 mil millones de fotos con un espacio de almacenamiento de más de 20 petabytes.

La variedad se refiere a que tan heterogéneos desde el punto de vista estructural son un conjunto de datos. Los avances tecnológicos permiten a las empresas utilizar diversos tipos de datos estructurados, semiestructurados y no estructurados. Los datos estructurados, que constituyen solo el 5% de todos los datos existentes se refieren a los datos tabulares que se encuentran en las hojas de cálculo o bases de datos relacionales. Texto, imágenes, audio y video son ejemplos de datos no estructurados, que a veces carecen de la organización estructural requerida por las máquinas para el análisis. Abarcando desde los datos totalmente estructurados hasta aquellos datos no estructurados en lo absoluto, el formato de datos semiestructurados no cumple con estándares estrictos. Extensible Markup Language (XML), un lenguaje textual para intercambiar datos en la Web, es un ejemplo típico de datos semiestructurados. Los documentos XML contienen etiquetas de datos definidas por el usuario que los hacen legibles por máquina.

Velocidad: Se refiere a la razón (de cambio) a la que se generan los datos y la velocidad a la que se deben analizar y actuar sobre ellos. La proliferación de dispositivos digitales, como teléfonos inteligentes y sensores, ha llevado a una tasa de creación de datos sin precedentes y está impulsando una creciente necesidad de análisis en tiempo real y planificación basada en la evidencia. Incluso los minoristas están generando datos de alta frecuencia. Wal-Mart, por ejemplo, procesa más de un millón de transacciones por hora. [6] Los datos que provienen de dispositivos móviles y fluyen a través de aplicaciones móviles producen torrentes de información que se pueden usar para generar ofertas personalizadas en tiempo real para los clientes diarios. Estos datos proporcionan información sólida sobre los clientes, como la ubicación geoespacial, la demografía y los patrones de compra anteriores, que se pueden analizar en tiempo real para crear un valor agregado para el cliente.

Hay que tener en cuenta que para realizar el proceso del análisis de los datos se debe tener claro el objetivo que se desea. [7] se pueden distinguir diferentes tipos de análisis". Estos son:

Análisis descriptivo: Según Angel M. Rayo [8] es cuando se quiere saber qué hacer para que suceda algo .Esto quiere decir que es el paso de resumir y simplificar los datos para que sean más manejables y visualizar el escenario en el que se encuentra.

Análisis Predictivo: Como se mencionó anteriormente a groso modo, el análisis de datos en tiempo real puede llegar a ser un factor crucial en muchos aspectos, como ejemplo tenemos información tal como ubicación del cliente, sus preferencias al momento de comprar, el tipo de compras que hace, las épocas del año en las que aumenta y disminuye el volumen de compras, entre otra gran cantidad de información que puede ser generada por una aplicación de compras online y que mediante su utilización puede ayudar al vendedor a realizar ofertas especializadas a cada cliente, este sería un claro ejemplo de un análisis de tendencias, en este caso aplicado a las ventas.

Análisis Prescriptivo: Es muy importante para cualquier compañia conocer de qué forma deben actuar ante una calamidad y estipular cuál va a ser el plan de acción identificando las posibles coincidencias de estas.

Un ejemplo claro de la aplicación de este análisis es una aplicación de transito como Waze la cual ayuda a elegir una mejor ruta teniendo en cuenta la distancia, velocidad en la que puedo viajar y condiciones del tráfico.

Para el proceso del análisis de los datos es de vital importancia conocer cuáles son los tipos que podemos encontrar en este proceso, [9] existen 3 tipos de datos:

Datos Estructurados Son aquellos que tradicionalmente se han usado en el tratamiento de datos y sus características principales son que se pueden almacenar en tablas y tienen clara una definición de longitud y formato"

Figura 3 Datos estructurados y no estructurados. Tomada del artículo Diferencias entre datos estructurados y no estructurados 2017.

Datos no estructurados: Estos al contrario de los estructurados son aquellos que su tratamiento se realiza de la misma forma en la que fueron recolectados o recogidos. Estos no tienen un formato que permita que se puedan almacenar en forma tradicional, tenemos como ejemplos los correos electrónicos, presentaciones de power point, archivos PDF o los procesadores de texto.

Datos semiestructurados: Esta es la combinación de las dos anteriores, ya que siguen una estructura en específico, pero esta no es lo suficiente regular como para gestionarla como con los estructurados; dentro de estos datos podemos encontrar los HTML.

Como se ha indicado el Big Data aporta grandes perspectivas las cuales abren paso a nuevas oportunidades y modelos de negocio para que las empresas se desenvuelvan y apunten a nuevos nichos de mercado en la actualidad. Para ello es muy importante tener en cuenta las siguientes tres acciones [10]:

Integrar:

El Big Data incorpora datos de varias fuentes por lo cual se requieren nuevas estrategias y tecnologías que permitan analizar conjuntos de datos de uno o más terabytes.

Gestionar:

Para poder gestionar estos datos se requiere un lugar en donde guardar dichos volúmenes, estos pueden ser en la nube o en servidores físicos, todo puede depender de la necesidad de las persona que desea analizar los datos, en la actualidad los servicios de almacenamiento de datos se están volviendo más populares ya que integran seguridad y continuidad de los datos, a diferencia de los servidores que pueden estar expuestos a desastres naturales y a fallos.

Analizar:

Este es el paso más importante del Big Data ya que de aquí es donde parten las soluciones, e ideas que cada empresa necesite, sin este paso toda la recolección de datos quedaría en eso. Es muy importante que se continúe explorando en nuevos descubrimientos con el análisis de los datos ya que en la actualidad en bien más importante de una compañía es la información.

Basándonos en estas tres acciones podemos comenzar con el proceso de descubrimiento de la información, existe un proceso llamado el KDD (Knowledge Discovery in Databases, en inglés) el cual básicamente es el proceso que se genera de analizar la información que se tiene o los datos con el fin de poder llegar a una conclusión o decisión. Una definición más acertada es la siguiente: *"El descubrimiento de conocimiento en bases de datos es un campo de la inteligencia artificial de rápido crecimiento, que combina técnicas del aprendizaje de máquina, reconocimiento de patrones, estadística, bases de datos, y visualización para automáticamente extraer conocimiento (o información), de un nivel bajo de datos (bases de datos*)" [11].

El KKD lo que busca combinar es el proceso de descubrimiento y análisis de la información.

Debemos dejar en claro que el KDD no es un producto de software desarrollado si no que es un proceso que va a estar compuesto de varias etapas para el análisis de la información.

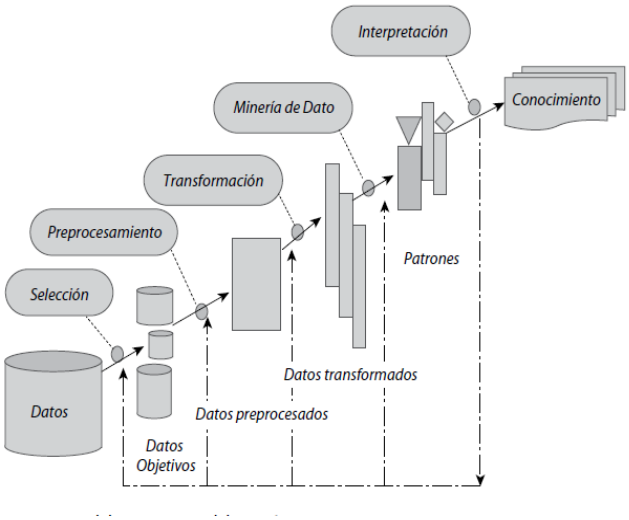


Figura 4.El proceso de descubrimiento de conocimiento en bases de datos. En Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional Timarán-Pereira, S. R., Hernández-Arteaga, I., Caicedo-Zambrano, S. J., Hidalgo-Troya, A. y Alvarado-Pérez, J. C.Pg (65) (2016) .

Como se puede evidenciar en la figura 4 el proceso del KDD consta de las siguientes etapas:

Etapa de selección: En esta etapa en en donde se escogen cuáles van a ser los datos a analizar dependiendo del objeto del estudio, este paso es importante ya que es la raíz de todo el proceso, se deben seleccionar de manera minuciosa para que en el momento de que se estén analizando los datos no ocurra que faltaron datos o sobraron.

Etapa de pre-procesamiento (limpieza): En esta etapa se analiza la calidad de los datos, se usan operaciones para la reducción de los datos "ruidosos" y se implementan estrategias para el uso de los datos desconocidos, los datos ruidosos [12] son valores que están significativamente fuera del rango de valores esperados; se deben principalmente a errores humanos,a cambios en el sistema, a información no disponible a tiempo y a fuentes.

Por su parte los datos que son desconocidos son aquellos que cuentan con un valor que no fue capturado al momento de la obtención de la información. Luego de clasificar e identificar estos datos se comienza con el proceso de limpieza que consiste en utilizar métodos estadísticos como la media o la moda para reemplazarlos

Etapa de transformación (reducción): Es el tratamiento preliminar de los datos que se obtuvieron en el proceso de selección y que posteriormente se procesaron. En esta etapa se buscan características relevantes para representar los datos dependiendo de los objetivos que se deseen. Se utilizan métodos de reducción de dimensiones o de transformación para disminuir el número efectivo de variables bajo consideración o para encontrar representaciones invariantes de los datos Dichos metodos de reduccion pueden ayudar a simplificar una tabla de una base de datos de forma horizontal o verticalmente. En donde la primera se basa en eliminar los elementos idénticos que van a quedar como el resultado de la sustitución de atributos. Por su parte la reducción vertical realiza la eliminación de atributos que o sirven o no van de la mano con los objetivos del problema.

Etapa de minería de datos: [13]El objetivo de la etapa minería de datos es la búsqueda y descubrimiento de patrones insospechados y de interés, aplicando tareas de descubrimiento como clasificación, clustering, patrones secuenciales, asociaciones, entre otras.

Estas técnicas de minerías de datos van a crear modelos que son predictivos o descriptivos en donde los predictivos van a permitir estimar valores futuros o que antes se desconocían, un ejemplo claro es predecir el comportamiento de clientes dependiendo de su estado civil, edad, género y profesión. Por su parte los modelos descriptivos tienden a estipular o identificar patrones que explican o minimizan los datos; lo cual sirve demasiado para para explorar todas las propiedades de los datos que están siendo analizados a diferencia de los predictivos.

[14]La escogencia de un algoritmo de minería de datos incluye la selección de los métodos por aplicar en la búsqueda de patrones en los datos, así como la decisión sobre los modelos y los parámetros más apropiados, dependiendo del tipo de datos (categóricos, numéricos) por utilizar.

Hay que enfatizar que dentro de la minería de datos existen tareas que pueden ser complejas de desarrollar por un algoritmo, dentro des estas se encuentran:

1. ***Clasificación:*** Este proceso permite obtener resultados a partir de un aprendizaje supervisado. dicho proceso se realiza en 2 pasos el primero con la construcción de un modelo en el cual cada elemento de un conjunto de elementos de la base de datos tiene una clase que se conoce como etiqueta que va a ser determinada por uno de los atributos de la base de datos. [15] Este conjunto sirve para la construcción del modelo llamado conjunto de entrenamiento el cual escoge randomicamente el número de elementos de la base de datos. En el segundo paso se utiliza el modelo para clasificar la información.
2. ***Segmentación:*** El proceso de agrupar objetos físicos o abstractos en clases de objetos similares se llama segmentación o clustering o clasificación no supervisada [16]. Básicamente lo que hace es agrupar un conjunto de datos, esto va a permitir identificar subpoblaciones homogéneas de datos, por ejemplo, si se aplicara en una base de datos de clientes servirá para identificar qué clientes tienen las mismas

tendencias o comportamiento al momento de realizar compras.

3. ***Asociación:*** Esta tarea va a descubrir patrones en forma de reglas, que va a demostrar que hechos ocurren frecuentemente en un conjunto de datos determinados.

Etapa de interpretación (evaluación de los datos): En esta etapa se comienza a interpretar cada uno de los patrones que descubrimos en las etapas anteriores, y puede que se retorne a estas para realizar posibles iteraciones. En esta etapa se puede incluir el análisis de los patrones extraídos, y se pueden quitar o destruir los patrones que tienden a ser completamente irrelevantes o redundantes, adicionalmente se traducen los patrones que son importantes o útiles para el proyecto.

También [13] se consolida el conocimiento descubierto para incorporarlo en otro sistema para posteriores acciones o, simplemente, para documentarlo y reportarlo a las partes interesadas; también para verificar y resolver conflictos potenciales con el conocimiento previamente descubierto.

Además de las fases descritas anteriormente generalmente se incluye otra fase previa para la determinación de las necesidades ya sea de la organización o del proyecto que se desea analizar para llegar al planteamiento del problema y sus objetivos. También se incluye otra etapa al finalizar la de interpretación que es donde los resultados que se obtuvieron comienzan a integrarse o usarse para la toma de decisiones o realizar acciones.

Tecnologías y herramientas de análisis de big data.

[15]Los tipos de datos no estructurados y semiestructurados generalmente no encajan bien en los almacenes de datos tradicionales que se basan en bases de datos relacionales orientadas a conjuntos de datos estructurados. Además, es posible que los almacenes de datos no puedan manejar las demandas de procesamiento planteadas por los conjuntos de big data que deben actualizarse con frecuencia, o incluso de manera continua, como en el caso de los datos en tiempo real sobre el comercio de acciones, las actividades en línea de los visitantes del sitio web. o el rendimiento de las aplicaciones móviles.

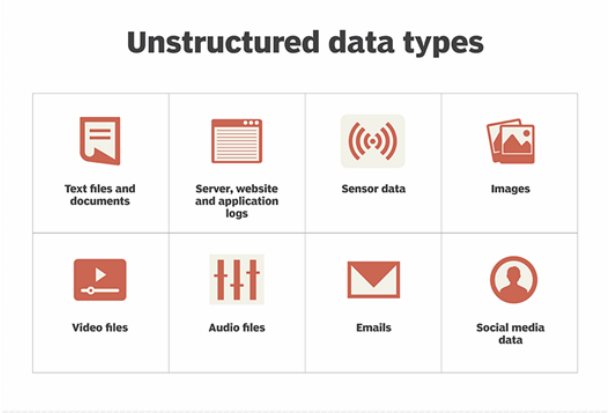
Como resultado, muchas de las organizaciones que recopilan, procesan y analizan big data recurren a las bases de datos NoSQL , así como a Hadoop y sus herramientas complementarias, que incluyen:

- ***YARN:*** una tecnología de administración de clústeres y una de las características clave de la segunda generación de Hadoop.
- ***MapReduce***: un marco de software que permite a los desarrolladores escribir programas que procesan grandes cantidades de datos no estructurados en paralelo en un grupo de procesadores distribuidos o computadoras independientes.
- ***Spark***: un marco de procesamiento paralelo y de código abierto que permite a los usuarios ejecutar aplicaciones de análisis de datos a gran escala en sistemas agrupados.
- ***HBase***: un almacén de datos clave / valor orientado a la columna creado para ejecutarse sobre el Sistema de archivos distribuidos de Hadoop (HDFS).
- ***Hive***: un sistema de almacenamiento de datos de código abierto para consultar y analizar grandes conjuntos de datos almacenados en archivos Hadoop.
- ***Kafka***: un sistema de mensajería distribuida de publicación / suscripción diseñado para reemplazar a los intermediarios de mensajes tradicionales.
- ***Pig***: una tecnología de código abierto que ofrece un mecanismo de alto nivel para la programación paralela de trabajos de MapReduce ejecutados en clusters de Hadoop.

¿Cómo funciona el análisis de datos grandes?

[17]En algunos casos, los clusters de Hadoop y los sistemas NoSQL se usan principalmente como plataformas de aterrizaje y áreas de preparación para los datos antes de que se carguen en un almacén de datos o en una base de datos analítica para el análisis, generalmente en una forma resumida que es más propicia para las estructuras relacionales.

Sin embargo, con mayor frecuencia, los usuarios de análisis de datos grandes están adoptando el concepto de un lago de datos Hadoop que sirve como el repositorio principal para las corrientes entrantes de datos en bruto. En tales arquitecturas, los datos pueden analizarse directamente en un clúster de Hadoop o ejecutarse a través de un motor de procesamiento como Spark. Al igual que en el almacenamiento de datos, la gestión de datos de sonido es un primer paso crucial en el proceso de análisis de big data. Los datos que se almacenan en el Sistema de archivos distribuidos de Hadoop deben organizarse, configurarse y particionarse correctamente para obtener un buen rendimiento de los trabajos de integración de extracción, transformación y carga (ETL) y consultas analíticas.



Una vez que los datos están listos, se pueden analizar con el software comúnmente utilizado para los procesos analíticos avanzados. Eso incluye herramientas para la minería de datos , que se filtran a través de conjuntos de datos en busca de patrones y relaciones; análisis predictivo , que construye modelos para pronosticar el comportamiento del cliente y

otros desarrollos futuros; Aprendizaje automático , que utiliza algoritmos para analizar grandes conjuntos de datos; y el aprendizaje profundo , una rama más avanzada del aprendizaje automático.

La minería de texto y el software de análisis estadístico también pueden desempeñar un papel en el proceso de análisis de big data, al igual que el software de BI y las herramientas de visualización de datos. Tanto para ETL como para aplicaciones analíticas, las consultas se pueden escribir en MapReduce , con lenguajes de programación como R , Python , Scala y SQL , los lenguajes estándar para bases de datos relacionales que son compatibles con las tecnologías de SQL-on-Hadoop.

Usos y desafíos del análisis de big data

Las aplicaciones de análisis de big data a menudo incluyen datos tanto de sistemas internos como de fuentes externas, como datos meteorológicos o datos demográficos sobre consumidores compilados por proveedores de servicios de información de terceros. Además, las aplicaciones de análisis de transmisión por secuencias se están volviendo comunes en los entornos de big data a medida que los usuarios buscan realizar análisis en tiempo real de los datos introducidos en los sistemas Hadoop a través de los motores de procesamiento de flujos, como Spark, Flink y Storm .

Los primeros sistemas de big data se implementaron principalmente en las instalaciones, particularmente en grandes organizaciones que recolectaron, organizaron y analizaron grandes cantidades de datos. Pero los proveedores de plataformas en la nube, como Amazon Web Services (AWS) y Microsoft, han facilitado la configuración y administración de los clusters de Hadoop en la nube, al igual que los proveedores de Hadoop como Cloudera y Hortonworks, que admiten su distribución del marco de big data. en las nubes AWS y Microsoft Azure . Los usuarios ahora pueden girar clusters en la nube, ejecutarlos durante el tiempo que lo necesiten y luego desconectarlos con precios basados en el uso que no requieren licencias de software continuas.

Las posibles dificultades de las iniciativas de análisis de big data incluyen la falta de habilidades de análisis interno y el alto costo de contratar científicos de datos con experiencia e ingenieros de datos para llenar los vacíos.

Recientemente, la proliferación y el avance de la inteligencia artificial y las tecnologías de aprendizaje automático han permitido a los proveedores producir software para el análisis de grandes datos que es más fácil de usar, en particular para la creciente población de científicos de datos ciudadanos. Algunos de los proveedores líderes en este campo incluyen Alteryx, IBM, Microsoft y Knime.

La cantidad de datos que suele estar involucrada, y su variedad, pueden causar problemas de gestión de datos en áreas que incluyen la calidad , la coherencia y la gobernabilidad de los datos. Además, los silos de datos pueden resultar del uso de diferentes plataformas y almacenes de datos en una arquitectura de big data. Además, la integración de Hadoop, Spark y otras herramientas de big data en una arquitectura cohesiva que satisfaga las necesidades de análisis de big data de una organización es una propuesta desafiante para muchos equipos de análisis y TI.

REFERENCIAS:

[1] José Martinez. ““Big Data”; aplicación y utilidad para el sistema sanitario” [En línea.].2015. Disponible en: https://www.researchgate.net/publication/274259789_Big_Data_application_and_use_for_the_health_system

[2] Gideon Gartner“What is Big data” [En línea] Disponible en: https://www.gartner.com/it-glossary/big-data/

[3] TechAmerica Foundation´s Federal Big Data Commision”” [En línea] Disponible en: https://bigdatawg.nist.gov/_uploadfiles/M0068_v1_3903747095.pdf

[4] Schroeck, Shockley, Smart, Romero-Morales y Tufano [En linea], 2012 Disponible en : https://www.bdvc.nl/images/Rapporten/GBE03519USEN.PDF

[5] Beaver, Kumar, Li, Sobel y Vajgel [En linea] Disponible en: http://ijcsit.com/docs/Volume%207/vol7issue2/ijcsit2016070270.pdf

[6] Kenneth Cukier [En línea] 2010 Disponible en: http://www.elboomeran.com/obra/1760/big-data-la-revolucion-de-los-datos-masivos/

[7]Computer traning by Netmind “Análisis de datos en Big Data: tipos y fases de analisis, 2016” .[En línea.] Disponible en: https://www.bit.es/knowledge-center/analisis-de-datos-en-big-data/ . Accedido[03-oct-2018]

[8]Computer traning by Netmind “Análisis de datos en Big Data: tipos y fases de analisis, 2016” .[En línea.] Disponible en: https://www.bit.es/knowledge-center/analisis-de-datos-en-big-data/. Accedido[03-oct-2018]

[9]Computer traning by Netmind “Tipos de datos en Big Data: clasificación por categoría y origen, 2016” .[En línea.] Disponible en: https://www.bit.es/knowledge-center/tipos-de-datos-en-big-data/.Accedido[03-oct-2018]

[10]Oracle, “¿Qué es Big Data”, [En linea]. Disponible en:https://www.oracle.com/co/big-data/guide/what-is-big-data.html. Accedido[03-oct-2018]

[11]Salvador Garcia, Sergio Ramirez, Julian Luengo y Francisco Herrera. “*Big Data Procesamiento y calidad de datos*”, 2016 Disponible en

:https://sci2s.ugr.es/sites/default/files/ficherosPublicaciones/2133_Nv237-Digital-sramirez.pdf. Acedido[03-oct-2018]

[13]Margaret Rouse. “Big Data analytics” 2018 Disponible en https://searchbusinessanalytics.techtarget.com/definition/big-data-analytics . Accedido [03-oct-2018]

[14] Chen, Han y Yu. [En línea] 1996. Disponible en: http://sedici.unlp.edu.ar/handle/10915/21206

[12]Timarán-Pereira, S. R., Hernández-Arteaga, I., Caicedo-Zambrano, S. J., Hidalgo-Troya, A. y Alvarado-

Pérez, J. C. (2016). El proceso de descubrimiento de conocimiento en bases de datos. En Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional (pp. 63-86). Bogotá: Ediciones Universidad Cooperativa de Colombia.

[13]Quinlan, J. (1986). Induction of Decision Trees. Machine Learning Journal, 1(1), 81-106.

FAYY97 Usama Fayyad y Evangelos Simoudis. (1997). Data Mining and Knowledge Discovery in Databases.

[15]Timarán-Pereira, S. R., Hernández-Arteaga, I., Caicedo-Zambrano, S. J., Hidalgo-Troya, A. y Alvarado-

Pérez, J. C. (2016). El proceso de descubrimiento de conocimiento en bases de datos. En Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional (pp. 63-86). Bogotá: Ediciones Universidad Cooperativa de Colombia.

[16]Timarán-Pereira, S. R., Hernández-Arteaga, I., Caicedo-Zambrano, S. J., Hidalgo-Troya, A. y Alvarado-

Pérez, J. C. (2016). El proceso de descubrimiento de conocimiento en bases de datos. En Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional (pp. 63-86). Bogotá: Ediciones Universidad Cooperativa de Colombia.

[17] Jiawei Han “Data Mining, concepts and techniques” [En linea] 2001 http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf