Polo del Conocimiento

Pol. Con. (Edición núm. 80) Vol. 8, No 4
Abril 2023, pp. 656-675
ISSN: 2550 - 682X
DOI: 10.23857/pc.v8i3

Mecanismos para el procesamiento de big data. Limpieza, transformación y análisis de Datos

Mechanisms for the processing of big data. Data cleaning, transformation and analysis

Negociação para o processamento de big data . Limpeza, transformação e análise de dados

Ricardo Orlando Malla- Valdiviezo [I]
ricardo.malla@utm.edu.ec
https://orcid.org/0000-0003-0841-7495

Oscar Alexander López -Gorozabel [II]
oscar.lopez@utm.edu.ec
https://orcid.org/0000-0002-0640-9953

Jorge Armando Arévalo- Indio [III]
jorge.arevalo@utm.edu.ec
https://orcid.org/0009-0002-7257-3337

Cesar Humberto Tóala- Briones [IV]
cesar.toala@utm.edu.ec
https://orcid.org/0009-0008-8975-6651

Correspondencia: ricardo.malla@utm.edu.ec

Ciencias de la Computación
Artículo de Investigación

*** Recibido:** 23 de febrero de 2023 ***Aceptado:** 14 de marzo de 2023 *** Publicado:** 01 de abril de 2023

I. Magister en Informática Empresarial UNIANDES, Ing. En Sistemas Informáticos Universidad Técnica de Manabí, Docente Universidad Técnica de Manabí, Ex - Coordinador Zonal de TIC MSP – Zona 4, Ex - Coordinador de Metas Institucionales y Asesor Educativo MINEDUC Zona 4, Sub Secretario de atención intergeneracional MIES.

II. Ingeniero de Sistemas Informáticos. Licenciado en Trabajo Social. Máster en Ingeniería de Software y Sistemas Informáticos por la Universidad Internacional de la Rioja. Docente de la Universidad Técnica de Manabí. Portoviejo, Ecuador.

III. Ingeniero En Sistemas Informáticos, Magister En Educación Informática, Analista Distrital de Soporte Técnico de la Dirección Distrital 13D09 Paján Salud, Docente de la Universidad técnica de Manabí. Portoviejo, Ecuador

IV. Ingeniero en Sistemas Informáticos, Analista de planificación institucional en la Universidad Técnica de Manabí, Magister en informática empresarial por la UNIANDES y un Máster Universitario en Evaluación de la Calidad y Procesos de Certificación en Educación Superior por la Universidad Internacional de la Rioja.

Resumen

Actualmente, la masificación de información en Internet, ha provocado el desarrollo de nuevas herramientas de análisis, por lo que se ha vuelto indispensable la adquisición de mecanismos sistemáticos que permitan administrar eficientemente la Big Data, todo esto con el fin de garantizar que las organizaciones y empresas puedan tomar decisiones relacionadas con un previo y efectivo análisis de datos. La Big Data es conocida como una combinación de datos estructurados, semiestructurados y no estructurados, los cuales son recopilados por las organizaciones, para luego ser procesados y presentados de manera pública o privada, posteriormente estos datos pueden usarse en proyectos de aprendizaje automático, modelado predictivo y otras aplicaciones de análisis avanzado. Se ha evidenciado que, en la actualidad la mayor parte de los datos no están estructurados, lo que dificulta la optimización de las tareas de procesamiento de datos y dado que el proceso de generación de datos no tiene fin, los procesos de recopilación y administración de información se han convertido en actividades más complejas. La presente investigación es de carácter analítico-sintético, debido a que se descomponen las partes a estudiar, realizándose un análisis y síntesis sobre la definición de la Big Data, tipos y mecanismos utilizados en el procesamiento de datos, además se presentan diversas herramientas utilizadas para administrar información voluminosa, asi mismo se identifican los algoritmos más eficientes para el procesamiento de información acorde a las necesidades de las organizaciones. Luego de integrar las partes estudiadas, se genera como resultado, una guía actualizada sobre los algoritmos y aplicaciones a utilizar en cada fase del procesamiento de datos, con el objetivo de que se facilite la toma de decisiones en las organizaciones.

Palabras Clave: Big Data; Mecanismos; Análisis de Datos; Herramientas de análisis;Algoritmos de procesamiento.

Abstract

Currently, the massification of information on the Internet has led to the development of new analysis tools, so it has become essential to acquire systematic mechanisms to efficiently manage Big Data, all this in order to ensure that organizations and companies can make decisions related to a previous and effective data analysis. Big Data is known as a combination of structured, semi-structured and unstructured data, which are collected by organizations, and then processed and presented publicly or privately, then these data can be used in machine learning projects, predictive

modeling and other advanced analytics applications. It has been evidenced that, at present most of the data are not structured, which makes it difficult to optimize data processing tasks and since the data generation process is never ending, the information collection and management processes have become more complex activities. The present research is of an analytical-synthetic nature, due to the fact that the parts to be studied are decomposed, performing an analysis and synthesis on the definition of Big Data, types and mechanisms used in data processing, in addition to presenting various tools used to manage voluminous information, as well as identifying the most efficient algorithms for processing information according to the needs of organizations. After integrating the parts studied, the result is an updated guide on the algorithms and applications to be used in each phase of data processing, with the objective of facilitating decision making in organizations.
Keywords: Big Data; Mechanisms; Data Analysis; Analysis tools; Processing algorithms.

Resumo

Actualmente, a massificação da informação na Internet levou ao desenvolvimento de novas ferramentas de análise, pelo que se tornou essencial adquirir mecanismos sistemáticos para gerir eficazmente os Grandes Dados, tudo isto para garantir que as organizações e empresas possam tomar decisões relacionadas com uma análise prévia e eficaz dos dados. Big Data é conhecido como uma combinação de dados estruturados, semi-estruturados e não estruturados, que são recolhidos por organizações e depois processados e apresentados pública ou privadamente, e podem ser utilizados em projectos de aprendizagem de máquinas, modelação preditiva e outras aplicações analíticas avançadas. Tornou-se evidente que, actualmente, a maioria dos dados não está estruturada, o que dificulta a optimização das tarefas de processamento de dados e, uma vez que o processo de geração de dados nunca termina, os processos de recolha e gestão da informação tornaram-se mais complexos. Esta investigação é de natureza analítico-sintética, devido ao facto de as partes a estudar serem discriminadas, realizando uma análise e síntese sobre a definição de Grandes Dados, tipos e mecanismos utilizados no processamento de dados, para além de apresentar várias ferramentas utilizadas para gerir informação volumosa, bem como identificar os algoritmos mais eficientes para o processamento da informação de acordo com as necessidades das organizações. Após a integração das partes estudadas, é gerado como resultado, um guia

actualizado sobre os algoritmos e aplicações a utilizar em cada fase do processamento de dados, com o objectivo de facilitar a tomada de decisões nas organizações.

Palavras-chave: Grandes Dados, Mecanismos, Análise de Dados, Ferramentas de Análise, Algoritmos de Processamento, Algoritmos de Processamento.

Introducción

Después del surgimiento del Internet, se han venido dado avances muy significativos en el ámbito de las telecomunicaciones, educación, negocios y entretenimiento, todas estas actividades han generado que exista gran cantidad de información, con diferentes estructuras, cuyo volumen de datos, complejidad y velocidad de crecimiento poseen características imposibles de procesar a través de modelos y herramientas tradicionales, lo que dificulta su procesamiento. Por otra parte, Zaheer & Zaynah (2019), afirman que la generación de información masiva en Internet o más conocida como Big Data suele ser incontrolable pero muy necesarias para las empresas en la actualidad, por lo cual se requiere de herramientas tecnológicas eficientes e inclusive de algoritmos que permitan mejorar el estudio de los datos (p. 2).

El análisis de Big Data en la actualidad toma cada vez más importancia en los mercados actuales, debido a su increíble utilidad, sobre todo en el entorno empresarial. Gracias a la difusión de los datos, el internet y las nuevas tecnologías, las empresas están recolectando constantemente un mayor volumen de información en tiempo real, tales como: datos de operación, clientes, proveedores y de todos los frentes de operaciones de las mismas. Cuando los datos llegan a integrarse, existen posibilidades de analizar y crear soluciones que permitan mejorar los procesos de toma de decisiones, es decir, se mejora la competitividad de las empresas. El estudio de los datos puede abaratar costos a las empresas, crear nuevos productos o servicios, entre otros.

Según Hernández, Duque & Moreno (2017), manifiesta que:

> Para la realización del manejo de los datos es indispensable contar con dos componentes de suma importancia como lo son el hardware y el software; del lado del hardware se cuenta con tecnologías de alto nivel como arquitecturas de MPP, que agiliza el procesamiento de los datos y del lado del software aparecen las tecnologías que ayudan en el correcto manejo de los datos no estructurados o semiestructurados; para estas necesidades se acude a las tecnologías como Spark o Hadoop, que son especialmente diseñadas para el manejo de información estructurada, no estructurada o semiestructurada (p. 18).

En la presente investigación se pretende identificar las diversas herramientas utilizadas para

recopilación, transformación y análisis de información voluminosa, asi mismo se estudian las características de los algoritmos más eficientes para el procesamiento de información con el objetivo de elaborar una guía metódica para el procesamiento de datos en las organizaciones.

Metodología

En el presente trabajo se utilizará el método analítico-sintético, debido a que se descomponen todas las partes que conforman el objeto de estudio, luego se procede a estudiar cada una de las partes de manera individual y más adelante se integran dichas partes con el fin de estudiarlas de manera holística e integral. Las partes a estudiar en esta investigación son las herramientas y algoritmos utilizados en las diversas fases del procesamiento de datos; dicha información es recopilada a través de la base de datos Google Académico y de los distintos estudios realizados en revistas científicas indexadas como ScieLo, Redalyc, Latindex, haciendo uso de palabras claves como: Big Data, Herramientas de análisis, Algoritmos para tratamiento de datos, entre otros.

Introducción al procesamiento de datos

Para Liarte (2019), el procesamiento de datos es el proceso de recopilar, limpiar, transformar y analizar información valiosa desde diversas fuentes, en la actualidad las empresas suelen utilizar técnicas de aprendizaje automático y estadísticas para recopilar información. El procesamiento de datos surge como una disciplina técnica a partir del apogeo del término Big Data que está relacionado a la información a gran escala, el procesamiento de datos posee una estructura ordenada en su aplicación, según (García et al, 2016) existen fases indispensables en el procesamiento de los datos, estas son:

- Limpieza de datos.
- Transformación de datos.
- Análisis de datos.

Limpieza de datos

Carranza (2022), la define como un proceso de corrección o eliminación de datos formateados de manera incorrecta, pudiendo estar duplicados o incompletos dentro de un gran conjunto de datos, el proceso de limpieza suele ser ejecutado desde software especializado, estos pueden ser:

- Apache Spark: software de código abierto que permite la limpieza y transformación de datos a gran escala de manera eficiente, incluye módulos para la administración de datos, como: Spark SQL y DataFrames.
- Talend: software de código abierto que ofrece una amplia gama de funciones de limpieza de datos, pudiéndolos integrar desde diferentes fuentes (csv, bases de datos, entre otros).
- Databricks: Plataforma basada en la nube, dedicada a la limpieza y transformación de datos a gran escala, incorpora módulos de análisis como: Spark DataFrames.
- Trifacta: Herramienta Online, donde los usuarios pueden explorar y transformar en tiempo real, grandes volúmenes de datos, en comparación a las anteriores Trifacta posee una interfaz visual Drag and Drop (arrastrar y soltar) que facilita las actividades a los usuarios sin conocimientos técnicos.
- Informática Power Center: Herramienta de integración para datos empresariales, con funciones de limpieza y normalización de datos a gran escala.

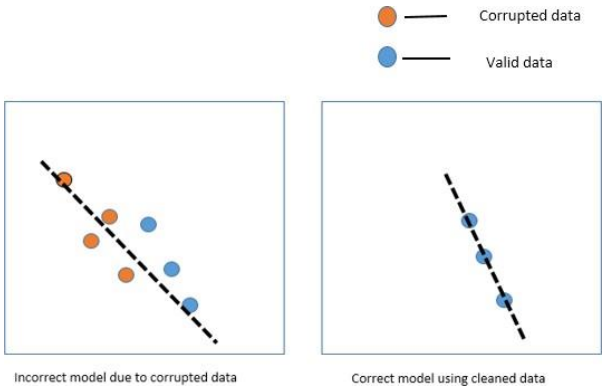


Gráfico 1: Representación de limpieza de datos.

Realizado por: Rahman (2019).

Transformación de datos

Luego de la limpieza de datos o Data Cleasing se procede a transformar los datos, por ende (Aguilar et al, 2022), define los siguientes pasos, para una transformación correcta:

- Compresión de datos, se transforman los datos a un formato en que se puedan gestionar de manera más sencilla y eficiente.
- Cifrado de datos, se traducen los datos a otro código para poder protegerlos.

Según Arias (2016), el proceso de transformación de datos se:

Torna mucho más complejo cuando se tiene una exuberante cantidad de datos no estructurados, por lo que

el proceso Extraer, Transformar y Cargar (en adelante ETL) suele convertirse en un cuello de botella, costoso de administrar y con un bajo desempeño. Por lo general, el proceso de transformación de datos se encarga de convertir el formato de los datos y ajustarlo a los requisitos de la fuente de destino.

Actualmente en el mercado, existen herramientas muy sofisticadas para la transformación de datos, (Szell, 2020) menciona algunas de las más usadas en el sector empresarial:

- Apache Hadoop: Framework de código abierto que permite a los usuarios procesar y almacenar grandes volúmenes de datos de manera distribuida.Hadoop realiza la transformación de datos a través de sus componentes: HDFS y MapReduce.
- Apache Spark: Framework de código abierto más utilizado en las industrias para el procesamiento y transformación de datos.
- Apache Hive: Motor de consulta SQL para Hadoop, permite interactuar con grandes volúmenes de datos.
- Apache Pig: Motor de programación, permite el procesamiento de grandes volúmenes de datos en Hadoop.
- Apache Flink: Apache Framework para el procesamiento de los datos en tiempo real.
- Google Cloud Dataflow: Herramienta de procesamiento de datos basada en la nube de Google.

En el transcurso del análisis de las diversas herramientas existentes actualmente, se determina que la herramienta más eficiente es Spark debido a su soporte para diferentes sistemas operativos y sobre cualquier plataforma en la nube, pudiendo ser: Amazon EC2/S3 o Google Cloud. Spark ofrece análisis de datos para campañas de marketing, sensores de IoT, aprendizaje automático y sitios de redes sociales en tiempo real. Además, Spark es que es compatible con una amplia variedad de lenguajes de programación, lo que lo hace accesible a una amplia gama de usuarios. Por otra parte, Spark ofrece una API muy completa y fácil de usar para el procesamiento de datos, lo que lo hace ideal para una amplia gama de aplicaciones, desde la ciencia de datos hasta la inteligencia artificial. En general, Spark es una de las mejores herramientas para el procesamiento y transformación de datos, gracias a sus características avanzadas y a la amplia gama de funcionalidades que ofrece.

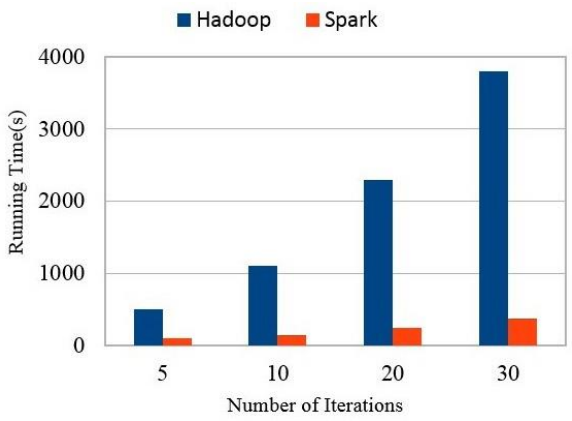


Gráfico 2: Comparación de regresión logística entre Hadoop y Spark.
Realizado por: Srikanth & Reddy (2016).

Análisis de datos

Debido a la masiva y diversa cantidad de información, se deben adoptar diversas tecnologías y técnicas analíticas, necesarias para la extracción y procesamiento de datos relevantes. Dentro de este marco existen varias técnicas estadísticas, reconocimiento de patrones, algoritmos matemáticos, algoritmos de machine learning, sin embargo, se identifica como principal técnica a la minería de datos, debido a que permite implementar algoritmos que son capaces de extraer datos masivos y con mucha calidad.

Para Giner (2018), la minería de datos se caracteriza por combinar métodos de estadística y machine learning con la gestión de bases de datos con el fin de identificar patrones en grandes conjuntos de datos. Dentro de la minera de datos, existen diversos algoritmos que son capaces de crear distintas formas de captar y extraer los datos para optimizar así el proceso de análisis. Un objetivo principal de esta investigación es identificar los algoritmos más utilizados y eficientes en el mundo de la Big Data y acorde a (Gutiérrez & Vigo, 2021), estos son:

- *MapReduce:* Algoritmo distribuido que permite procesar grandes volúmenes de datos en paralelo.
- *Algoritmos de aprendizaje automático:* Algunos de estos son: Naive Bayes, Árboles de Decisión, Random Forest, SVM (Support Vector Machines), entre otros, estos algoritmos se mejoran automáticamente basándose en la experiencia.
- *Regresión Lineal:* Algoritmo es muy útil para predecir valores futuros a partir de datos históricos.

- *Algoritmos de Clustering:* Algoritmos que permiten agrupar objetos o personas por similitud y se utilizan para descubrir patrones en grandes conjuntos de datos.
- *Algoritmos de minería de datos:* Técnicas que combinan métodos de estadística y machine learning con la gestión de bases de datos para identificar patrones en grandes conjuntos de datos.

Luego de investigar de manera exhaustiva sobre los diferentes algoritmos para el análisis de datos, se procede a identificar a los algoritmos de Clustering como los más ampliamente utilizados y eficientes. Estos algoritmos funcionan reuniendo objetos o personas similares en grupos o clústeres, con el objetivo de que los miembros del clúster compartan características similares y los clústeres sean lo más diferenciados posible. El proceso de Clustering se realiza a través de la identificación de patrones en los datos, lo cual permite agrupar objetos similares en un solo clúster y objetos diferentes en clústeres distintos. Además, los algoritmos de búsqueda son ampliamente utilizados en la actualidad y son los que están mejor probados para establecer patrones a partir de datos previamente establecidos.

Algoritmos de clustering

Algoritmo K-Means

Es uno de los más populares y se utiliza ampliamente en la industria, este algoritmo trabaja dividiendo los datos en k grupos o clústeres, donde k es un número previamente establecido.

Ramírez (2023), define a este algoritmo como uno de los más utilizados y consiste en dividir el conjunto de datos en k clústeres, donde k es un número predefinido por el usuario. El algoritmo K-Means es uno de los algoritmos de clustering más utilizados y ampliamente conocidos en la industria de la minería de datos. Este algoritmo funciona dividiendo un conjunto de datos en k grupos (clústeres) basados en las características similares de los objetos incluidos en ellos. La finalidad del algoritmo es maximizar la diferencia entre los clústeres y minimizar la similitud dentro de ellos.

El algoritmo K-Means se realiza en dos etapas principales. En la primera etapa, se asignan los centroides a cada uno de los clústeres. Estos centroides representan el centro geométrico de los objetos incluidos en cada clúster. En la segunda etapa, se reasignan los objetos a los clústeres en función de su distanciaa los centroides.

Pol. Con. (Edición núm. 80) Vol. 8, No 4, Abril 2023, pp. 656-675, ISSN: 2550 - 682X

El algoritmo K-Means utiliza un enfoque iterativo para mejorar la asignación de los objetos a los clústeres. Cada iteración se realiza hasta que no se produzcan más cambios en la asignación de los objetos a los clústeres. Este proceso se repite hasta que se alcance una solución óptima.

Una de las mayores ventajas del algoritmo K-Means es su eficiencia en términos de tiempo de procesamiento y capacidad de manejar grandes conjuntos de datos. El algoritmo es fácil de implementary es muy escalable, lo que lo hace ideal para una amplia gama de aplicaciones en la minería de datos.

Sin embargo, existen algunas desventajas del algoritmo K-Means. Una de las más comunes es su dependencia de la elección de los valores iniciales para los centroides. Si los valores iniciales son pocorepresentativos, el algoritmo puede no producir los resultados deseados. Además, el algoritmo K-Meanstambién puede ser sensible a la presencia de outliers en los datos, lo que puede afectar negativamente ala calidad de los resultados.

Otro factor importante a tener en cuenta al utilizar el algoritmo K-Means es la necesidad de especificarel número de clústeres deseados (k) antes de comenzar el proceso. Si se escoge un valor de k que no refleje adecuadamente la estructura de los datos, los resultados pueden ser pobres. Por lo tanto, es importante seleccionar un valor de k adecuado para el conjunto de datos en cuestión.

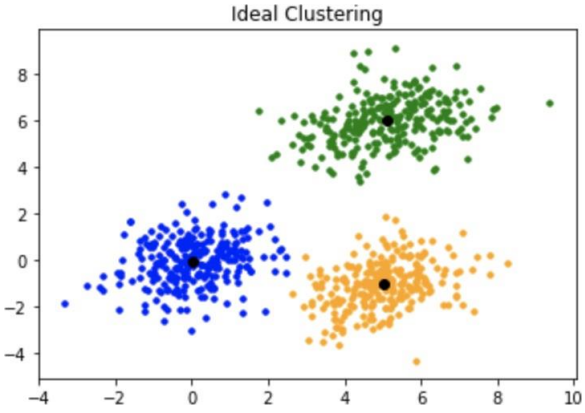


Gráfico 3: Algoritmo K-means.

Realizado por: Gupta et al (2023).

Algoritmo por agrupamiento jerárquico

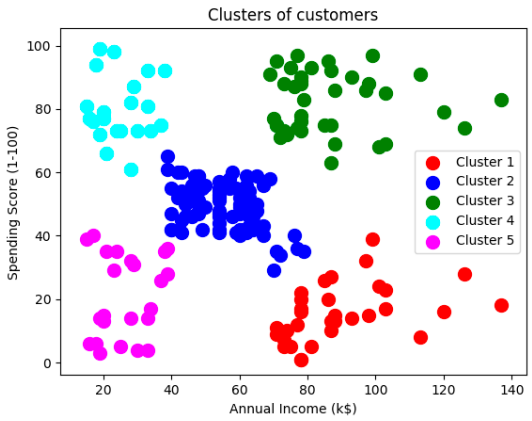
Según Jianan (2022), el clúster por jerarquías es un algoritmo que se utiliza para agrupar objetos en clústeres basados en la relación de contención entrelos objetos. Otro algoritmo de clustering comúnmente utilizado es el algoritmo de agrupamiento jerárquico. Este algoritmo utiliza un enfoque diferente al de k-means, ya que utiliza una técnica de agrupamiento que recursivamente

agrupa objetos similares en un solo clúster. El algoritmo de agrupamiento jerárquico produce una estructura de árbol que representa la relación entre los clústeres, lo que lo hace ideal para visualizar y explorar los datos.

El algoritmo de agrupamiento jerárquico es una técnica de clustering que se utiliza para clasificar objetos en grupos basados en su similitud y relación. Es uno de los algoritmos de clustering más antiguos y ampliamente utilizados. El algoritmo de agrupamiento jerárquico se divide en dos enfoques principales: aglomerativo y divisivo. En el enfoque aglomerativo, se inicia con cada objeto en un clúster separado y, a continuación, se combinan gradualmente clústeres hasta que todos los objetos se agrupan en un solo clúster. Por otro lado, en el enfoque divisivo, se inicia con todos los objetos en un solo clúster y, a continuación, se divide gradualmente en clústeres más pequeños hasta que cada objeto se encuentre en un clúster separado.

Este algoritmo utiliza una representación de árbol llamada dendrograma para representar los clústeres resultantes. Cada nivel del dendrograma representa un clúster diferente, y los nodos en el dendrograma representan la fusión o la división de clústeres. El dendograma también permite visualizar la relación entre los clústeres y su evolución a lo largo del tiempo.

El algoritmo de agrupamiento jerárquico se basa en la medida de distancia entre los objetos. La medida de distancia más comúnmente utilizada es la distancia Euclidiana, aunque también se pueden utilizar otras medidas de distancia, como la distancia de Manhattan o la distancia de Mahalanobis. La medida de distancia se utiliza para determinar la similitud entre los objetos y para decidir cuándo combinar o dividir los clústeres. El algoritmo de agrupamiento jerárquico es flexible en cuanto a su capacidad para manejar diferentes tipos de datos, como datos continuos, categóricos o mixtos. Además, es una técnica robusta y resistentea la presencia de ruido o valores atípicos en los datos. Sin embargo, una de las desventajas del algoritmode agrupamiento jerárquico es que puede



Pol. Con. (Edición núm. 80) Vol. 8, No 4, Abril 2023, pp. 656-675, ISSN: 2550 - 682X

ser computacionalmente costoso, especialmente cuando se trabaja con grandes conjuntos de datos. Otra desventaja es que el algoritmo de agrupamiento jerárquicorequiere la selección de un criterio de corte para determinar el número final de clústeres.

Gráfico 4: Algoritmo de agrupamiento jerárquico

Realizado por: Berzal (2018).

Algoritmo Density-Based Spatial Clustering of Applications with Noise

Berzal (2020), lo define como un algoritmo de clustering basado en densidad, que se enfoca en la identificación de regiones densamente pobladas de datos. El algoritmo Density-Based Spatial Clustering of Applications with Noise (en adelante DBSCAN) es un algoritmo de clustering basado en densidad, diseñado para identificar regiones densamente pobladas de datos y encontrar estructuras en los datos. Este algoritmo es uno de los métodos de clustering más utilizados en la industria y se destaca por su capacidad para identificar clusters de forma automática, sin la necesidad de especificar previamente el número de clusters.

El algoritmo DBSCAN se basa en dos conceptos principales: la densidad y el radio. La densidad se refiere a la cantidad de puntos presentes en una región dada, y el radio se refiere a la distancia máximapermitida entre dos puntos para que se consideren parte del mismo clúster. A partir de estos conceptos, DBSCAN define un clúster como un conjunto de puntos en los que la distancia entre cualquier par de puntos es menor o igual que el radio.

El proceso de DBSCAN comienza eligiendo un punto al azar y buscando otros puntos dentro del radio.Si se encuentran suficientes puntos dentro del radio, se considera que existe un clúster. Luego, se eligenlos puntos encontrados en el primer clúster y se realiza el mismo proceso para ellos, hasta que ya no se encuentren puntos adicionales. Este proceso se repite hasta que se han explorado todos los puntos en elconjunto de datos.

Una de las ventajas de DBSCAN es que es capaz de identificar clusters de diferentes formas y tamaños,y puede manejar datos con un número variable de dimensiones. Además, DBSCAN es capaz de identificar clústeres no convexos y detectar puntos atípicos o ruido en los datos, lo que lo diferencia deotros algoritmos de clustering que solo pueden identificar clústeres convexos. Sin embargo, el algoritmo DBSCAN también tiene algunos desafíos. Uno de los mayores desafíos es encontrar el valor adecuado para el radio, ya que un radio demasiado pequeño puede resultar en la creación de demasiados clústeres, mientras que un radio demasiado grande puede combinar clústeres diferentes en un solo clúster. Además, DBSCAN puede tener dificultades para manejar

datos con una distribución no uniforme y con una concentración muy alta de puntos.

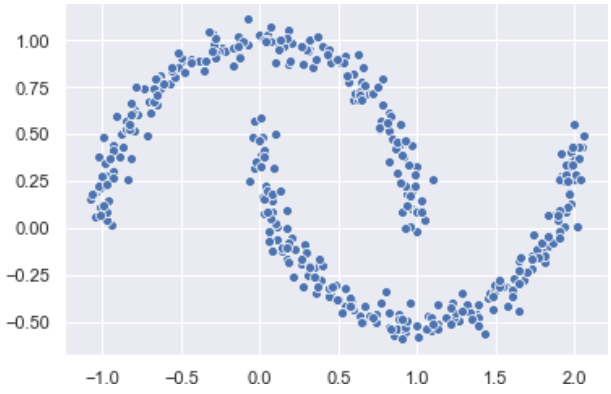


Gráfico 5: Algoritmo DBSCAN

Realizado por: Berzal (2020).

Se determina que de los tres algoritmos de clustering mencionados anteriormente, el más eficiente acorde a las necesidades específicas de la empresa, es K-Means por su ejecución rápida y eficiente. Por otra parte, el algoritmo Hierarchical Clustering es una buena opción si se requiere una visión general de la estructura de los clústeres y si se desea realizar un seguimiento de cómo los datos evolucionan a lo largo del tiempo. Finalmente, se recomienda el algoritmo DBSCAN cuando los clústeres tienen formas irregulares y si se requiere identificar regiones densamente pobladas de datos.

Se ha redactado de manera detallada toda la información para que las organizaciones tengana su disposición una guía de las mejores herramientas, algoritmos y métodos indispensables para el procesamiento de Big Data, debido a que las empresas deben automatizar sus procesos tradicionales. Con esta información las organizaciones podrán sacar el máximo provecho a los resultados obtenidos detrás del arduo proceso de limpieza, transformación y análisis de los datos.

Resultados

En esta sección de resultados se diseña una guía para las empresas acerca de la Big Data y la forma más adecuada para el procesamiento de información, limpieza y análisis de datos, con el objetivo de mejorar la selección de herramientas y algoritmos utilizados en el procesamiento de datos dentro de las organizaciones.

Guía para el procesamiento de información en las organizaciones

Limpieza de datos

Antes de comenzar el análisis de los datos, es importante que se realice una limpieza adecuada para eliminar cualquier dato inconsistente, duplicado o faltante. La limpieza de datos también puede incluirla corrección de errores y la normalización de los datos para hacerlos consistentes. Hay muchas herramientas disponibles en el mercado para ayudar en la limpieza de datos, incluyendo aquellas que se integran con sistemas de Big Data, como Apache NiFi y Apache Hive, así como soluciones de software independientes.

Melissa Data Clean Suite es una opción popular para muchas empresas debido a su capacidad para proporcionar una limpieza de datos eficiente y precisa. Con Melissa Data Clean Suite, las empresas pueden verificar y corregir la dirección postal, el correo electrónico y los números de teléfono de sus registros de datos. Además, la suite también ofrece la capacidad de eliminar duplicados y normalizar los datos para hacerlos consistentes.

Otro factor que hace que Melissa Data Clean Suite sea una opción atractiva para muchas empresas es su facilidad de uso. La suite es fácil de integrar con sistemas existentes y no requiere conocimientos técnicos avanzados para utilizarse. Esto significa que las empresas pueden comenzar a utilizar la suite de forma rápida y eficiente, sin tener que dedicar una gran cantidad de recursos a la formación de sus equipos.

Transformación de Datos

La transformación de datos implica la conversión de los datos en un formato que sea adecuado para su análisis. Esto puede incluir la agregación de datos, la creación de nuevas variables y la eliminación de variables irrelevantes. La transformación de datos es un paso crítico en el proceso de Big Data, ya quepermite a las empresas preparar los datos para su análisis. Esto incluye la reorganización de los datos, la eliminación de datos irrelevantes, la creación de nuevas variables y la combinación de datos de diferentes fuentes. La transformación de datos también es importante para corregir errores en los datosy hacerlos consistentes.

Spark es una de las opciones más populares para la transformación de datos en Big Data. Spark es un marco de procesamiento de datos en paralelo que ofrece una amplia gama de funciones para la transformación de datos, incluidas las operaciones de agregación, filtrado y reemplazo de datos. Sparktambién es una solución escalable, lo que significa que puede manejar grandes volúmenes de

datos con eficacia. Además, Spark es compatible con una amplia gama de lenguajes de programación, como Java,Python, Scala y R, lo que lo hace accesible para una amplia variedad de equipos de desarrollo.

En general, Spark es una opción popular para la transformación de datos en Big Data debido a su eficiencia, escalabilidad y compatibilidad con una amplia gama de lenguajes de programación. Las empresas pueden utilizar Spark para procesar y transformar grandes volúmenes de datos en un formatoadecuado para su análisis, lo que les permite obtener insights valiosos a partir de sus datos.

Análisis de Datos

Determinamos que de los tres algoritmos de clustering mencionados anteriormente, es posible decir queel mejor algoritmo depende de las necesidades específicas de la empresa. Sin embargo, en términos generales, se puede decir que el algoritmo K-Means es una buena opción si los clústeres tienen formasregulares y bien definidas, y si se requiere una ejecución rápida y eficiente. El algoritmo Hierarchical Clustering es una buena opción si se requiere una visión general de la estructura de los clústeres y si sedesea realizar un seguimiento de cómo los datos evolucionan a lo largo del tiempo. Finalmente, el algoritmo DBSCAN es una buena opción si los clústeres tienen formas irregulares y si se requiere identificar regiones densamente pobladas de datos.

Herramientas Tecnológicas

Las empresas pueden utilizar una amplia gama de herramientas y tecnologías para el procesamiento y análisis de Big Data. Algunas de las tecnologías más populares incluyen Apache Hadoop, Apache Spark, Apache Storm, Apache Flink y Apache Cassandra. También existen herramientas de análisis dedatos, como Tableau, Power BI, QlikView y SAS, que pueden ayudar a las empresas a visualizar y analizar sus datos de manera efectiva.

Además, existen plataformas en la nube, como Amazon Web Services (AWS) y Microsoft Azure, que ofrecen soluciones de Big Data asequibles y escalables. Estas plataformas pueden ayudar a las empresasa reducir los costos de hardware y de mantenimiento, y a tener una infraestructura de Big Data flexibley escalable.

Análisis e Interpretación

El procesamiento de datos se debe de manera general a tres importantes fases: Limpieza, Transformación y Análisis, por lo que de manera resumida se recomienda a las empresas la implementación de mecanismos que les permitan obtener resultados efectivos de información, esto con el objetivo de tomar decisiones correctas antes y después de las operaciones o actividades empresariales. También se debe invertir en infraestructura adecuada a las necesidades, en términos de escalabilidad como de costo. Las empresas deben contar con un equipo capacitado en procesamiento y análisis de Big Data, sobre todo aquellas empresas que manejan grandes volúmenes de información (ejemplo: La Fabril, CNT, Claro, Movistar, entre otros). Aunque la mayor parte de los servicios y aplicaciones que permiten la administración de Big Data son de pago, existen otras alternativas de software libre, tales como: Python, R, Power BI en su versión básica, entre otros.

Discusión

La guía para el procesamiento de datos, diseñada en esta investigación se encuentra actualizada, es una propuesta que garantiza a las organizaciones economizar costos en relación a la contratación de servicios. Lo que confirma (Flores & Villacís, 2017) como un factor muy importante dentro de las organizaciones, debido a que medida que han surgido aplicaciones de uso gratuito algunas empresas se han visto beneficiadas, sin obviar que la mayor parte de los servicios otorgados en estas licencias son básicos y no contienen todas las funciones, incluyendo la seguridad. Por lo tanto (García, 2022) asegura que es más conveniente pagar por las herramientas y así poder tener un control más amplio y efectivo de las mismas, en el sector bancario es elemental contar con todas las funcionalidades.

Las herramientas para el procesamiento de datos, pueden ser utilizadas por diversas organizaciones, aunque existan algunas que se encargan de desarrollarlas apoyándose de tecnologías de software libre, tales como Python, R u otras. Se vive en un mundo completamente digitalizado y el bien más importante para las organizaciones es la información, la cual debe ser procesada de manera correcta para tomar decisiones eficientes.

Conclusiones

La era digital en la que vivimos ha generado una cantidad masiva y creciente de datos que se producena una velocidad cada vez mayor. Estos datos pueden ser de diversos formatos y provienen

de múltiplesfuentes, incluyendo redes sociales, transacciones comerciales y sensores. La capacidad de procesar y analizar estos datos puede ser una ventaja competitiva para las empresas, ya que les permite obtener insights valiosos y tomar decisiones informadas basadas en datos.

El procesamiento de Big Data implica varios pasos, incluyendo la limpieza de datos, la transformaciónde datos y el análisis de datos. La limpieza de datos implica la eliminación de cualquier dato inconsistente, duplicado o faltante, así como la corrección de errores y la normalización de los datos para hacerlos consistentes. Hay muchas herramientas disponibles en el mercado para ayudar en la limpieza de datos, incluyendo Apache NiFi y Apache Hive, así como soluciones de software independientes como Melissa Data Clean Suite, que es la mejor opción para la mayoría de las empresasporque cada empresa es distinta y tiene distintas necesidades.

La transformación de datos implica la conversión de los datos en un formato adecuado para su análisis.Esto puede incluir la agregación de datos, la creación de nuevas variables y la eliminación de variablesirrelevantes. La transformación de datos es un paso crítico en el proceso de Big Data, ya que permite alas empresas preparar los datos para su análisis. Spark es una de las opciones más populares y de las mejores para la transformación de datos en Big Data.

El análisis de datos es un proceso clave para la obtención de información valiosa y útil para las empresas. El uso de técnicas de aprendizaje automático y estadísticas permite explorar los datos de manera efectiva y obtener insights. En cuanto a los algoritmos de clustering mencionados, K-Means, Hierarchical Clustering o también llamado Agrupamiento jerárquico y DBSCAN, cada uno de ellos tiene sus fortalezas y debilidades y la elección del mejor algoritmo dependerá de las necesidades específicas de la empresa. Sin embargo, en general, K-Means es una buena opción para clústeres con formas regulares y bien definidas, Hierarchical Clustering es adecuado para una visión general de la estructura de los clústeres y seguimiento a lo largo del tiempo, mientras que DBSCAN es útil para identificar regiones densamente pobladas de datos con formas irregulares. Es importante destacar que la elección del mejor algoritmo de clustering requiere un conocimiento profundo de los datos y una evaluación cuidadosa de las necesidades de la empresa.

En conclusión, el procesamiento de Big Data es un aspecto clave para las empresas que buscan mejorarsu toma de decisiones y aprovechar al máximo los datos que generan. La limpieza y transformación delos datos son pasos críticos en este proceso y deben ser realizados de manera eficiente y precisa. La utilización de herramientas especializadas en la limpieza y transformación de

datos, como Melissa DataClean Suite y Spark, puede ser de gran ayuda para las empresas en este proceso.

Se recomienda que las empresas inviertan en la capacitación de sus equipos en el uso de estas herramientas y en la comprensión de los conceptos y prácticas relevantes en el procesamiento de Big Data. También es importante que las empresas establezcan políticas y procedimientos claros para la gestión de datos y se aseguren de cumplir con las regulaciones y leyes aplicables en la protección de datos personales. En última instancia, un enfoque integral.

Referencias

1. Aguilar Aguilar, I., Cuevas Cruz, F., Duran Martínez, P., García Carmen, E., Hernández Romero, A., Mateos Casimiro, E., Ortega Sánchez, J., Retana Contreras, J., Ruiz Macedonio, J., Segundo Romero, C., Solís Colin, I., Ugalde Zaldivar, J., Vázquez Clemente, B., Vázquez Ramírez, A. & Yépez Martínez, D. (2022). Antología gestión y análisis de Big Data LIAD6 2022-A. Recuperado desde: http://ri.uaemex.mx/handle/20.500.11799/137910
2. Arias, W. (2019). BIG DATA: Extraer, transformar y cargar los datos, *Instituto Internacional de Ciencias de Datos*, https://i2ds.org/2016/05/04/big-data-extraer-transformar-y-cargar-los-datos/
3. Berzal, F. (2018). Clustering jerárquico. DECSAI. https://elvex.ugr.es/idbis/dm/slides/42%20Clustering%20-%20Hierarchical.pdf
4. Berzal, F. (2020). Clustering basado en densidad. DECSAI. https://elvex.ugr.es/idbis/dm/slides/43%20Clustering%20-%20Density.pdf
5. Camargo Vega, J., Camargo Ortega, J. & Joyanes Aguilar, L. (2015). Conociendo Big Data. *Revista Facultad de Ingeniería*, 24(38), pp. 63–77. http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0121-11292015000100006
6. Carranza, A. (2022). Data Cleansing: averigua cómo limpiar datos erróneos y conservar información valiosa, https://www.crehana.com/blog/transformacion-digital/data-cleansing/
7. García, I. (2022). Big Data en seguros. *NowoTech*. https://nowo.tech/formacion/big-data-en-seguros/

8. García, S., Ramírez, S., Luengo, J. & Herrera, F. (2016). Big Data: Preprocesamiento y calidad de datos, *Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada (España)*, 237(1), pp. 17-23, https://sci2s.ugr.es/sites/default/files/ficherosPublicaciones/2133_Nv237-Digital-sramirez.pdf
9. Giner, G. (2018). Minería de datos: ¿qué relación tiene con el Big data?. *Business Revista Digital*. https://www.escueladenegociosydireccion.com/revista/business/big-data/la-mineria-de-datos-en-el-big-data/
10. Gutiérrez, J. & Vigo, V. (2021). Modelo de aprendizaje automatizado del proceso de venta de productos financieros en un Call center. https://repositorio.ulima.edu.pe/handle/20.500.12724/14344
11. Hernández Leal, E., Duque Méndezy N., Moreno Cadavid, J. (2017). Big Data: una exploración de investigaciones, tecnologías y casos de aplicación. TecnoLógicas, 20(39), pp. 15-38. https://www.redalyc.org/journal/3442/344251476001/html/
12. Jianan, L. (2022). Understanding Mean Shift Clustering and Implementation with Python. *Towards Data Science*, https://towardsdatascience.com/understanding-mean-shift-clustering-and-implementation-with-python-6d5809a2ac40
13. Liarte Muñoz, J. (2019). Análisis de datos de las organizaciones. Big data. *TFG-Facultad de Ciencias de la Empresa,* pp. 1-45. https://repositorio.upct.es/bitstream/handle/10317/7754/tfg-lia-ana.pdf?sequence=1&isAllowed=y
14. Flores Avendaño, P. & Villacís Vera, A. (2017). Análisis comparativo de las herramientas de Big data en la Facultad de Ingeniería de la Pontificia Universidad Católica del Ecuador. http://repositorio.puce.edu.ec/handle/22000/14119
15. Rahman, A. (2019). What is Data Cleaning? How to Process Data for Analytics and Machine Learning Modeling?. *Towards Data Science,* https://towardsdatascience.com/what-is-data-cleaning-how-to-process-data-for-analytics-and-machine-learning-modeling-c2afcf4fbf45
16. Ramírez, L. (2023). Algoritmo k-means: ¿Qué es y cómo funciona?. *Business & Tech*. https://www.iebschool.com/blog/algoritmo-k-means-que-es-y-como-funciona-big-data/

17. Srikanth, B. & Reddy, V. (2016). Efficiency of Stream Processing Engines for Processing BIGDATA Streams. *Indian Journal of Science and Technology*. https://www.researchgate.net/publication/301797425_Efficiency_of_Stream_Processing_Engines_for_Processing_BIGDATA_Streams
18. Szell, C. (2020). Herramientas de la transformación digital – El ETL. *Conectamagazine*, https://www.conectasoftware.com/magazine/conector/herramientas-de-la-transformacion-digital-el-etl/
19. Viswarupan, N. (2017). K-Means Data Clustering. *Towards Data Science*, https://towardsdatascience.com/k-means-data-clustering-bce3335d2203
20. Zaheer A., Zaynah A. (2019). On big data, artificial intelligence and smart cities, *Cities*, 89(1), pp. 80-91, https://www.sciencedirect.com/science/article/pii/S0264275118315968