Big Data

Jose A. Arroyo Meoño, Bryan S. Feng Feng, Juan P. Rodríguez Cano, Gennell C. Rodriguez Hall,

y Jose A. Vargas Torres

Instituto Tecnológico de Costa Rica

CE3101 Bases de Datos

Prof. Marco Rivera Meneses

31 de octubre del 2024

Índice

Introducción

Xuewei Li y Xueyan Li (2018, p. 76) nos indican que desde el 2005, debido a que los trabajos que Google publicó con la idea de sistemas de computación paralela basada en mapas reducidos fueron aplicados a través de Hadoop, el concepto de big data se comenzó a utilizar ampliamente en Estados Unidos. Años más tarde, el gobierno de este país y las Naciones Unidas comenzaron una serie de análisis basados en big data de algunas áreas como las redes sociales, y en la administración del presidente Barack Obama se haría una inversión de 200 millones de dólares en este tipo de sistemas.

Además, los autores explican que el big data es diferente a las tecnologías de recolección de información normales, pues hoy en día tienen tres objetivos específicos: el procesamiento de la información creada por organizaciones sobre el comportamiento de sus usuarios, la extracción del significado de datos no estructurados y la lectura de contenido científico y de producción para fomentar la cooperación entre el desarrollo científico y de compañías (p.77).

A pesar de que esta tecnología emergente muestra un gran potencial impuslado por las inversiones de los gobiernos y empresas a nivel mundial, Yojna Arora y el Dr. Dinesh Goyal (2016, p.229) señalan que enfrenta numerosos desafíos: se necesita reducir el volumen de datos mediante técnicas de muestreo ya que los algoritmos de recolección no son lo suficientemente escalables al crecimiento masivo de los datos; no todos los problemas pueden resolverse a través de computación paralela basada en mapas reducidos; retos en la seguridad de la gestión de los datos; dificultades en el procesamiento de datos no estructurados; entre otros.

Es por la importancia del big data en la actualidad que en esta investigación se exploran los conceptos relacionados con esta nueva tecnología, con el objetivo último de formular un ejercicio introductorio a la interpretación de datos de esta magnitud. Para esto se explicarán más a fondo las herramientas, de entre las cuales se escogieron Google Dataset Search y Dask en Python para realizar el taller, que se utilizan para desarrollar algoritmos de big data, los beneficios que estas muestran en la industria y posibles desafíos que puedan presentar.

Marco Teórico

Según Elnasri (2016), “Desde la llegada de la World Wide Web en 1994, la cantidad de datos a nivel mundial ha crecido exponencialmente” (p.911). Este crecimiento fenomenal dio lugar al concepto de “Big Data”, el cual, según Zendesk (2023), “consiste en un proceso que analiza e interpreta grandes volúmenes de datos, tanto estructurados como no estructurados. Su finalidad es que las empresas utilicen estos datos remotos como base para su toma de decisiones" (párraf.2).

Según Zendesk (2023), “El Big Data funciona en base a las llamadas ‘5 Vs’: volumen, variedad, velocidad, veracidad y valor”. Estos aspectos permiten manejar eficientemente los datos y maximizan su impacto en las organizaciones.

El volumen se refiere a la enorme cantidad de datos que se generan a diario. Desde interacciones en redes sociales hasta registros de transacciones financieras, la cantidad de datos disponibles ha crecido exponencialmente, requiriendo sistemas de almacenamiento robustos como las soluciones en la nube y bases de datos distribuidas para gestionar este flujo masivo de información.

La variedad se refiere a la diversidad de fuentes de datos, que plantea el desafío de integrar diferentes formatos y tipos en un mismo sistema de análisis. Las empresas deben adaptarse para manejar datos provenientes de redes sociales, dispositivos móviles, registros de clientes, entre otros, lo cual incrementa la complejidad pero también las oportunidades de obtener información más rica y detallada.

Por otro lado, la velocidad es crucial en el mundo actual, donde la capacidad de procesar datos en tiempo real determina la eficiencia de los procesos. Tecnologías como el streaming de datos permiten que las empresas tomen decisiones basadas en datos frescos, sin retrasos, mejorando así la experiencia del cliente y la eficiencia operativa.

Otro factor importante es la veracidad, que se refiere a la calidad y fiabilidad de los datos. En muchas ocasiones, los datos pueden ser incompletos, inconsistentes o erróneos, lo que

compromete su utilidad. Para asegurar que las decisiones basadas en Big Data sean correctas, es necesario implementar procesos de validación y limpieza de los datos.

Finalmente, el valor se refiere al potencial que tiene el Big Data para transformar los datos en conocimiento útil. Esto implica la capacidad de analizar y utilizar efectivamente estas grandes cantidades de información, mediante herramientas avanzadas como la inteligencia artificial (IA) y el aprendizaje automático (machine learning).

Con la gran cantidad de datos que manejan los sistemas de Big Data, es crucial contar con un método eficiente para su procesamiento. Este objetivo se logra mediante los procesos ETL (Extract, Transform, Load). Según Elena Bellos (2022), “Los procesos ETL (Extract, Transform, Load) hacen referencia a un conjunto de técnicas, herramientas y tecnologías que permiten extraer datos de varias fuentes, transformarlos de forma que sean veraces y útiles, y cargarlos en otros sistemas…”.

El primer proceso, conocido como Extract, consiste en extraer los datos desde uno o múltiples sistemas de origen. Esta información puede provenir de archivos comunes, ERP, bases de datos, etc.

El segundo proceso, Transform, se refiere a la transformación de los datos. En el contexto de Big Data, es un paso esencial, ya que este procedimiento convierte lo recopilado en información útil y disponible para su análisis posterior.

El tercero y último proceso, Load, se encarga de cargar los datos en su respectiva base de datos.

De las tres etapas que componen el proceso ETL, la fase de "Transformación" es la que más suele variar, ya que aquí se aplican todos los cambios a los datos extraídos para unificarlos como si provinieran de una única fuente. Una vez transformados, estos datos se almacenan en un “Data Warehouse” o almacén de datos (Bellos, 2022).

Las herramientas ETL permiten filtrar, seleccionar, ordenar y almacenar todos los datos obtenidos, eliminando la información irrelevante y facilitando el trabajo de los analistas en las

empresas. Es importante resaltar que la relevancia de los datos depende de los intereses específicos de cada compañía.

Algunas ventajas del proceso ETL para decodificar las grandes cantidades de datos o Big Data incluyen la migración efectiva de datos entre aplicaciones, evitando la alteración en la integridad de los mismos. También permite replicar datos con el objetivo de conseguir copias de seguridad y para el análisis de redundancia existente. Por otro lado, posibilita que la información sea almacenada en una base de datos, clasificada y transformada en campos como Inteligencia de Negocios o información de alto valor (CEUPE, s.f., párraf.6).

El proceso de Big Data no solo implica el análisis masivo de datos, sino también el uso de tecnologías avanzadas como la inteligencia artificial y el aprendizaje automático. Estas herramientas permiten identificar patrones y tendencias ocultas en grandes conjuntos de datos, lo que ayuda a las empresas a anticiparse a las demandas del mercado, mejorar la eficiencia operativa y personalizar las experiencias de sus clientes.

Finalmente, con los datos recopilados, seleccionados y almacenados, se procede al análisis de la información. Este análisis depende del objetivo de la organización, pero en la mayoría de los casos, los analistas de datos son quienes determinan la utilidad de la información para la empresa. El Big Data suele emplearse en áreas como el marketing, la lucha contra el crimen, el deporte y la política, entre otras. Cualquier sector relacionado con la estadística, de alguna manera, utiliza Big Data.

Una vez comprendida la base del Big Data y los procesos ETL, es importante conocer algunas de las principales herramientas para la manipulación de los datos.

Principales Herramientas

Apache Hadoop

Apache Hadoop consiste en software libre utilizado para la gestión de Big Data. Esta herramienta almacena los datos por medio de clusters, lo cual es una técnica que permite separar los datos según sus similitudes, con el fin de que no se pierdan. Esta herramienta utiliza un sistema de archivos llamado Hadoop Distributed File System, el cual procesa los datos mediante

dos fases, conocido como MapReduce. "Hadoop es un marco que permite el almacenamiento y procesamiento de grandes conjuntos de datos de forma distribuida" (Shvachko, Kuang, Radia, & Chansler, 2011). Apache Hadoop se usa usualmente en plataformas como Google Cloud y Amazon EC2/S3.

MongoDB

En el caso de MongoDB esta herramienta almacena datos en formato BSON lo cual significa que guardará los datos en forma de documentos, solo que de manera binaria por eso es BSON y no de tipo JSON. MongoDB es reconocido ya que es especialmente útil en entornos que requieren escalabilidad como además se menciona en (Nand & Malhotra, 2019): "MongoDB permite una escalabilidad horizontal mediante técnicas como la replicación y el sharding, facilitando así el manejo de grandes volúmenes de datos" . Gracias a estas características, podemos construir sistemas que escalen sin necesidad de que esto se vuelva un dolor de cabeza.

Elasticsearch

Elasticserch es una de las herramientas dentro del Big Data cuando se habla de por supuesto manejar grandes flujos de datos y además complejos pero sobretodo poder tener acceso a esta informacion de manera actual y poder hacer uso de esta información para realizar consultas ya que estos datos están indexados correctamente lo cual facilita en el acceso a esto.

Además Elasticsearch puede realizar búsquedas de texto complejas, visualizar el estado de los nodos y escalar fácilmente si se necesita más potencia. "Elasticsearch permite el procesamiento de datos a gran escala para búsquedas de texto completo y visualización, facilitando una gestión eficiente de grandes conjuntos de datos" (Magomedov, 2018).

Beneficios del Big Data en la industria

Mejores decisiones

Gracias al análisis de datos recolectados y el análisis de datos tomados, es posible una toma de decisiones más informada y estudiada, ya que existen parametros y estándares que cumplir y seguir.

Clientes felices

Con la recolección de datos se puede estár más cerca de los pensamientos y opiniones de los clientes lo cual permite tener una mayor cercanía a los mismos y así reducir fuentes de error en procesos dejando grandes beneficios tanto para el cliente como para la empresa.

Aumento en la seguridad

La estructura que lleva esta recolección de datos es de gran exactitud y orden por medio de tablas, documentos e imagenes, además esta información se encuentra respaldada y asegurada correctamente lo cual permite observar un panorama más completo, permitiendo la rápida identificación de amenazas.

Desafios

Almacenamiento

Cada vez es más la información que existe y debe de ser almacenada dejando la interrogación de a dónde se guardará esta información.

Velocidad

Otra cosa a tomar en cuenta es la velocidad de gestion de la información, ya que entre menos tiempo de analisis pase en algunas ocasiones, es más sencillo resolver y proponer soluciones.

Análisis

Una vez que se recolecta la información a analizar es importante que esta gestión y entendimiento de datos sea la correcta ya que de no serlo podría significar muchos problemas.

Veracidad

La validez de la información también es de suma importancia ya que si esta información recolectada no representa a la población podrían llegarse a soluciones o análisis erróneos.

Desarrollo

Contexto del taller y selección del caso de estudio

Objetivo del taller

El taller tiene como objetivo aplicar conceptos teóricos de Big Data en un caso simulado, de esta manera ayudando a facilitar la comprensión de cómo estas técnicas impactan la forma de decisiones en un contexto práctico.

Descripción del caso de estudio

Análisis de datos de la correlación de la popularidad y las características de una canción de spotify; y también de la correlación de las distintas características de las canciones.

Relevancia para el Big Data

Este caso de estudio ilustra cómo se pueden utilizar grandes volúmenes de datos para resolver problemas prácticos, especialmente en áreas clave como marketing, salud o finanzas.

Proceso de recolección de datos

Identificación de fuentes de datos

Para buscar los datos se utilizó Google Dataset Search, utilizando la palabra clave de “Spotify” se encuentra una dataset de spotify que contiene características de las canciones, año de publicación, artista, entre otros.

Herramientas de recolección

Para la recolección de datos se utilizó Dask, una biblioteca de Python que permite trabajar con grandes conjuntos de datos que no caben en memoria.

Proceso de limpieza y preparación de los datos

Preparación y estandarización de los datos

Para esta parte se seleccionó las columnas numéricas de interés para el análisis. (popularity, valence, acousticness, danceability, energy, instrumentalness, liveness, loudness, speechiness, tempo)

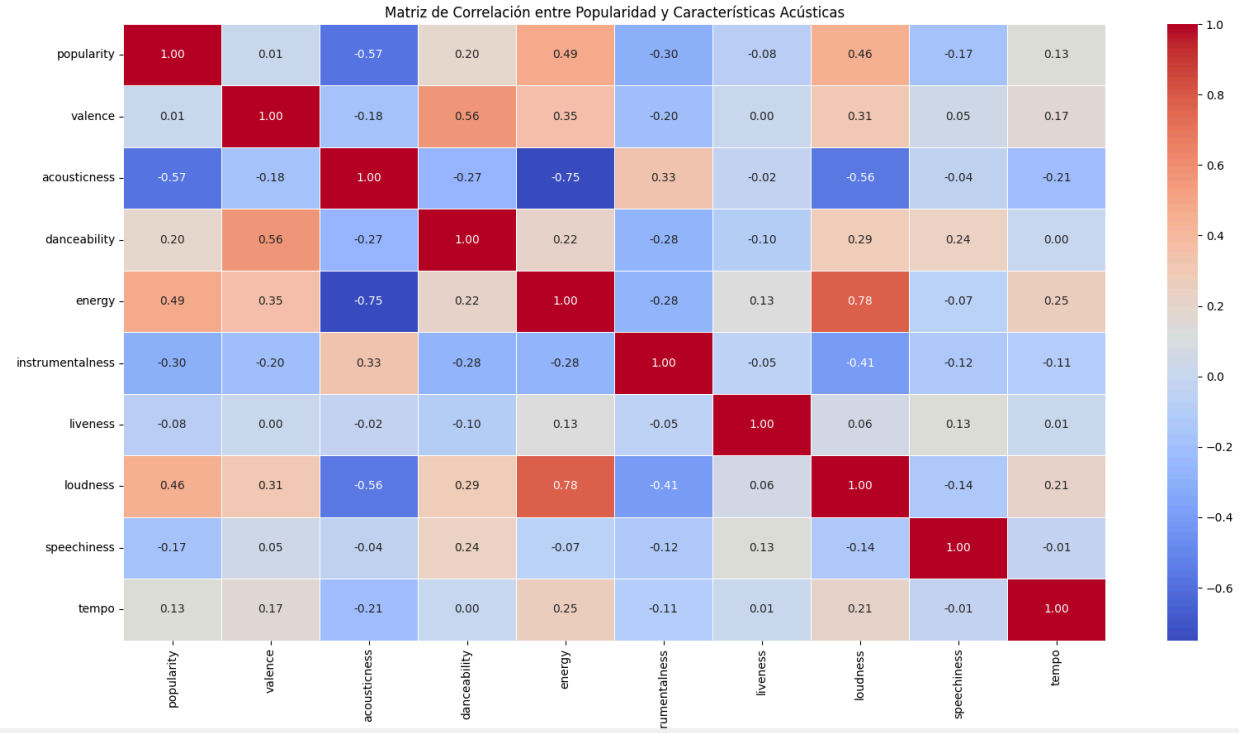
Herramientas utilizadas

Se utilizó Dask para el proceso de limpieza y preparación y luego de limpiar y preparar los datos, se calcula la matriz de correlación. Para lo anterior se creó el siguiente código en python.

```python
import dask.dataframe as dd
import seaborn as sns
import matplotlib.pyplot as plt

# Cargar los datos desde un archivo CSV usando Dask
data = dd.read_csv(r'C:\Users\PC\PycharmProjects\TallerDatos\data.csv\data.csv')

# Seleccionar las columnas numéricas de interés
numeric_cols = ['popularity', 'valence', 'acousticness', 'danceability',
                'energy', 'instrumentalness', 'liveness', 'loudness', 'speechiness', 'tempo']

# Calcular la matriz de correlación
corr_matrix = data[numeric_cols].corr().compute()  # .compute() para calcular el resultado

# Visualizar la matriz de correlación como un mapa de calor
plt.figure(figsize=(10, 8))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt=".2f", linewidths=0.5)
plt.title("Matriz de Correlación entre Popularidad y Características Acústicas")
plt.show()
```

Primero se importó las librerías de dask; seaborn, biblioteca de visualización de datos y matplotlib.pyplot para crear gráficos. Luego se carga los datos del archivo CSV en un dataframe de Dask. Posteriormente se seleccionan las columnas de interés, en este caso las características de las canciones y su popularidad. Usando ".corr()" se obtiene la correlación de la matriz de columnas y ".compute()" para obtener el resultado final de la correlación. Finalmente, se crea una figura con un tamaño personalizado con "plt.figure(figsize=(10,8))" y se genera el mapa de calor de la matriz de correlación con "sns.heatmap()"; y después se detalla el título y se muestra el mapa de calor generado.

Resultados



Análisis

Se puede notar que una matriz de correlación es simétrica con respecto a la diagonal principal por lo que consta analizar un solo segmento del área dividida por dicha diagonal.

De inicio podemos ver como las correlaciones más fuertes son entre la acústica y energía con un valor de -0.75, aunque no es un resultado fuera de lo usual ya que es un resultado que se puede intuir fácilmente, al igual que el valor de 0.78 para la correlación entre volumen y energía. Pero antes de analizar los valores para la variable de popularidad que es la que más interesa y puede ser de utilidad, se encuentran resultados interesantes con correlaciones un poco más bajas. Por ejemplo, se puede ver que suele haber una leve correlación negativa entre la música instrumental y la música bailable, lo cual puede indicar que canciones que se quieran utilizar para un evento de fiesta filtre aquellas que llevan muchos componentes de instrumentos. Otro resultado que no es tan intuitivo es la correlación entre la música bailable y música positiva, con un valor de 0.35. Esto indica que a pesar de géneros como el rap, electrónica, trap etc. que

suelen tener mensajes más pesados o problemáticos que tienen popularidad en lugares de fiesta no tienen la misma calificación en cuanto a lo bailable.

Los resultados con más importancia se encuentran en la correlación entre la popularidad y los demás atributos. Se puede notar que la mayor correlación es con la acústica con un valor de -0.57, indicando que entre menos sonidos acústicos tenga, más popularidad tendrá. Se podría deducir que a los usuarios de spotify les importa más la energía y volumen de las canciones para repetir las visitas a dicha canción ya que estos atributos tienen valores de 0.49 y 0.46 respectivamente.

Conclusiones

Durante el trabajo se indagó sobre las características de las tecnologías big data, explicando las 5V's, volumen, variedad, velocidad, veracidad y valor; y las herramientas ETL. Además, se detallaron las principales herramientas de big data utilizadas en la industria, Apache Hadoop, MongoDB, Elasticsearch; se dieron a conocer sus beneficios como una mayor satisfacción del cliente y una mejor en la seguridad; y se examinaron sus desventajas entre las cuales se encuentras los desafíos de almacenamiento, la velocidad de análisis y la veracidad de los datos recolectados.

En el taller utilizó una ETL llamada Dask, escrita en Python, para transformar la información de 170 mil canciones en Spotify en un gráfico del cual se pudo visualizar que la popularidad es inversamente proporcional a la acústica y proporcional a la energía y el volumen y que las canciones bailables son menos instrumentales y más positivas.

Recomendaciones

Por último, observamos que al trabajar con conjuntos de datos tan grandes como los que se encuentran en Google Datasets y otras bases de datos de Big Data, se recomienda agilizar el proceso primero dividiendo los datos en varios archivos para acelerar su carga en paralelo

utilizando herramientas como Dask y, segundo, filtrar los datos para evitar cargar elementos que no se consideren necesarios en el análisis.

Link del video

https://youtu.be/Vk7FezJ6nUI

Referencias

Arora, Y., & Goyal, D. (2016). *Big Data: A Review of Analytics Methods & Techniques*. 2nd International Conference on Contemporary Computing and Informatics (IC3I), IEEE. doi:10.1109/IC3I.2016.7917965

Bello Elena. (2022). Guía de Procesos ETL: Qué son, cómo usarlos y herramientas clave. https://www.iebschool.com/blog/que-son-los-procesos-etl-big-data/

Elmasri, R., & Navathe, S. B. (2016). *Fundamentals of database systems* (7th ed.). Pearson.

Instituto de Investigación en Inteligencia Artificial (IIC). (n.d.). Herramientas Big Data para empresa. Recuperado de https://www.iic.uam.es/innovacion/herramientas-big-data-para-empresa/

Li, X., & Li, X. (2018). *Big Data and Its Key Technology in the Future*. *Computing in Science & Engineering*, IEEE CS y AIP. doi:10.1109/MCSE.2018.042781329

Nand, A., & Malhotra, S. (2019). A review on various aspects of MongoDB databases. *International Journal of Engineering Research and Technology*, 8(5). Recuperado de https://d1wqtxts1xzle7.cloudfront.net/60661268/a-review-on-various-aspects-of-mongodb-databases-IJERTV8IS050031201909 21-10051-a8lu8n-libre.pdf

Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2011). The Hadoop distributed file system. *USENIX Conference on Hot Topics in Cloud Computing*. Recuperado de https://www.usenix.org/legacy/publications/login/2011-06/openpdfs/Shvachko.pdf

Zendesk (2023, 6 octubre). *Big Data: ¿qué es y para qué sirve? [LOS 5 Vs QUE LO RIGEN]*. Recuperado de https://www.zendesk.com.mx/blog/big-data-que-es/

Centro Europeo de Postgrado. (2023, 27 marzo). ¿Cuáles son los procesos ELT y ETL en el Big Data? - Maestrías Online. CEUPE.

ttps://ceupe.com.ar/blog/cuales-son-los-procesos-elt-y-etl-en-el-big-data/#:~:text=%C2%BFCu%C3%A1l%20es%20el%20proceso%20ETL,datos%20cuyas%20fuentes%20son%20ilimitadas

Repartición de Tareas

Tabla de Contenidos, Introducción y Conclusiones (Jose Arroyo)

I parte Marco Teórico (José Vargas)

II parte Marco Teórica y Recomendaciones (Carolina)

Desarrollo- Taller- Descripción Trabajo Realizado (Bryan)

Análisis de resultados y bibliografías (Juan Pablo-)

Link Presentación Canva

https://www.canva.com/design/DAGUr0vyyrI/X6SI0NEjv_Lr_z3fRMp_eQ/edit?utm_content=DAGUr0vyyrI&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton