
Universidad
Tecmilenio®

Universidad Tecmilenio

Ingeniería de datos masivos

Técnicas de distribución y procesamiento de datos

Semana 7

Introducción

Mercado libre es una plataforma de comercio electrónico en Latinoamérica que se ocupa de dar asistencia a más de 200 millones de usuarios en 18 países. A través de la plataforma, sus clientes tienen la posibilidad de comprar y vender productos mediante subastas o precios fijos, además de hacer pagos seguros y tener acceso a un crédito sin la necesidad de una tarjeta bancaria.

Al modernizar su infraestructura para almacenar datos en la nube con AWS, obtuvo un ahorro del 20% en sus costos durante el primer mes de uso (AWS, s.f.).

Explicación

Infraestructura de datos

Es una infraestructura digital para habilitar el consumo e intercambio de datos (Schulz, 2017), que se diseña para satisfacer los requerimientos de uso de los datos, promoviendo la eficiencia y la productividad del entorno en el que se emplea.

Los **ingenieros de datos** se encargan de definir, diseñar e implementar la infraestructura de datos, al diseñar los espacios de almacenamiento y la arquitectura para almacenar, procesar, explotar y administrar el flujo de los datos para su procesamiento en la organización. Además, definen las políticas de almacenamiento y gobernanza para garantizar el flujo y que se realicen visualizaciones, análisis y modelos de los mismos.

Se construye cuando se habilitan los sistemas de almacenamiento con hardware local o en la nube, usando bases de datos, herramientas ETL, *data warehouses* y *data lakes*.

Explicación

Cómputo y almacenamiento en la nube

Es la entrega bajo pedido de servicios de procesamiento de cómputo, almacenamiento de datos, aplicaciones y otros recursos de TI a través del Internet con precios de pago por su uso.

Las organizaciones pueden consumir servicios de computadoras bajo demanda, así como recursos de almacenamiento en lugar de construir, operar y mejorar la infraestructura por sí mismos.

Las organizaciones pueden consumir servicios de computadoras bajo demanda, así como recursos de almacenamiento en lugar de construir, operar y mejorar la infraestructura por sí mismos.

El almacenamiento en la nube depende del tipo de datos que se quieran almacenar, el objetivo de su uso y el volumen de los mismos.

Los diferentes **tipos de almacenamiento** son (Confluent, s.f.):

- **Base de datos.**
- **Data warehouse.**
- **ETL** (extract, transform and load).
- **Data lake.**

Explicación

Base de datos

Es una colección organizada de información estructurada o datos almacenados de forma electrónica en computadoras.
Utilizan lenguaje de consulta estructurado (SQL) para escribir y consultar datos.

Data warehouse

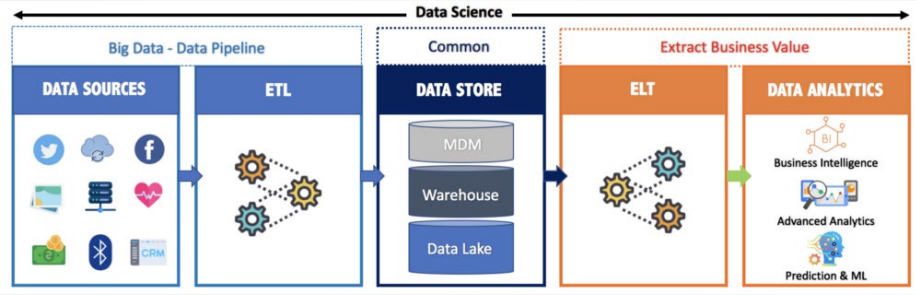
Depósito central de información que se usa para analizar y tomar decisiones.
Los datos fluyen desde sistemas transaccionales, bases de datos relacionales y otras fuentes.

ETL

Herramientas de software y procesos para hacer el movimiento y transformación de datos, permitiendo moverlos desde múltiples fuentes, reformatearlos y cargarlos en el data warehouse para analizarlos.

Data lake

Repositorio centralizado que permite almacenar datos estructurados y no estructurados a cualquier escala sin tener que crear una estructura para almacenarlos.



Fuente: Ahmed, I. (2018). *Big Data Science in 5 Minutes.* Recuperado de https://medium.com/@peterjaberau/big-data-science-in-5-minutes-a99372117d55

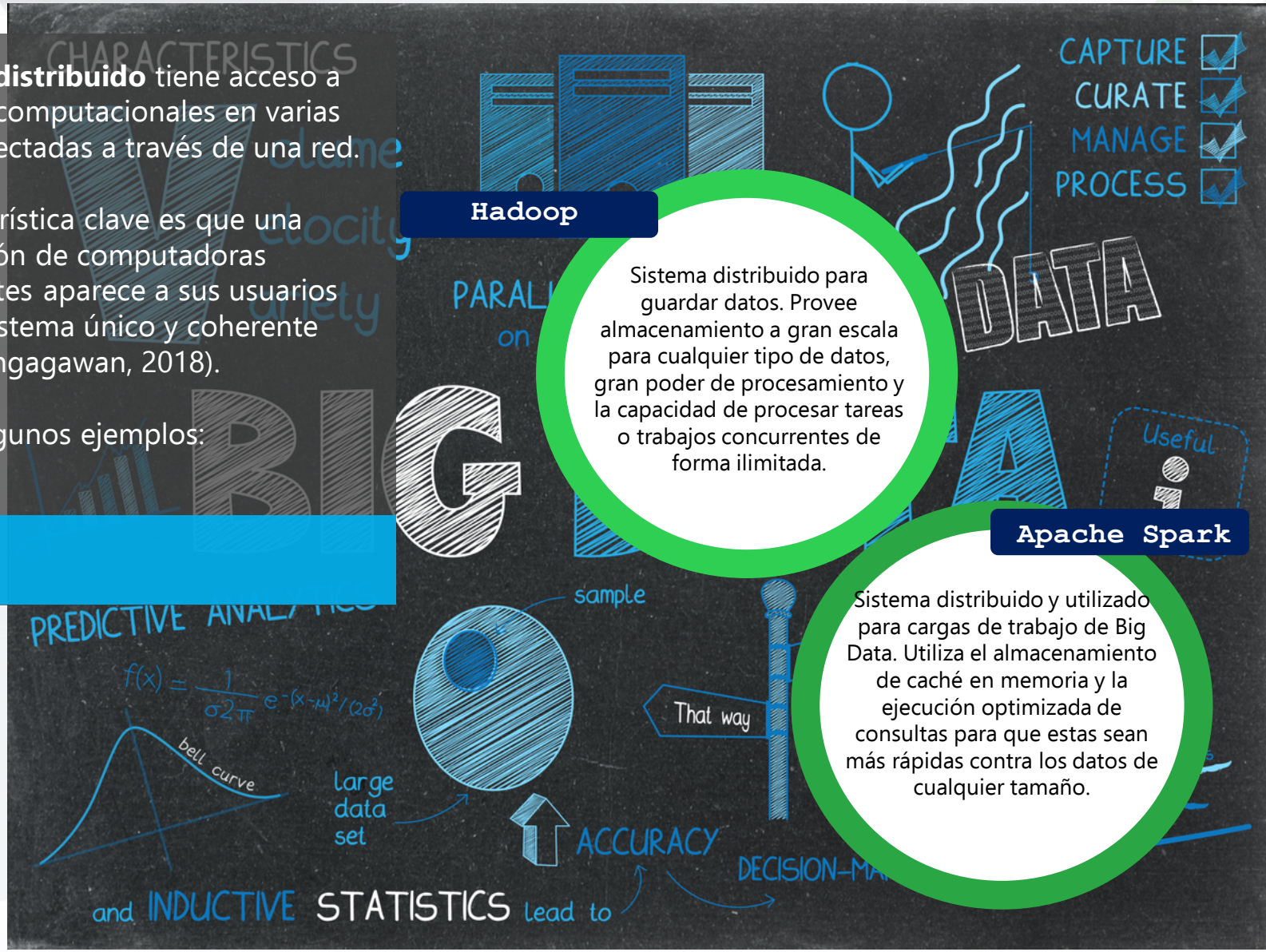
Por su volumen, los **datos** requieren de una infraestructura más sofisticada, almacenamiento en la nube y herramientas que permitan su control. Los **data warehouses** y los **data lakes** son complementarios, ya que el primero sirve para saber qué pasó, cómo pasó y cómo se comportan los indicadores de desempeño, mientras que el segundo almacena Big Data.

Explicación

Un **sistema distribuido** tiene acceso a los recursos computacionales en varias máquinas conectadas a través de una red.

La característica clave es que una colección de computadoras independientes aparece a sus usuarios como un sistema único y coherente (Tungagawan, 2018).

Algunos ejemplos:



Hadoop

Sistema distribuido para guardar datos. Provee almacenamiento a gran escala para cualquier tipo de datos, gran poder de procesamiento y la capacidad de procesar tareas o trabajos concurrentes de forma ilimitada.

Apache Spark

Sistema distribuido y utilizado para cargas de trabajo de Big Data. Utiliza el almacenamiento de caché en memoria y la ejecución optimizada de consultas para que estas sean más rápidas contra los datos de cualquier tamaño.

Explicación

Universidad Tecmilenio

Los diferentes proveedores de cómputo en la nube ofrecen servicios de analítica. Sin embargo, también es posible utilizar los notebooks de **Jupyter** en estas plataformas, ya que los datos y los notebooks de Jupyter están en la nube.

Algunos ejemplos de estas herramientas:

Otra opción a las herramientas señaladas es correr localmente los notebooks de Jupyter para tener acceso a los datos que estén almacenados en la nube.

Cierre

El propósito de la infraestructura de datos es protegerlos y convertirlos en información. Además, si se utiliza adecuadamente, se traducirá en una reducción significativa de los costos operativos en las organizaciones, por ejemplo, el caso de Mercado Libre y el uso de la infraestructura en la nube de Amazon Web Services.

Bibliografía

- Ahmed, I. (2018). *Big Data Science in 5 Minutes.* Recuperado de https://medium.com/@peterjaberau/big-data-science-in-5-minutes- a99372117d55
- AWS. (s.f.). *Simple Change Cuts Mercado Libre Compute Costs 31% Without Hurting Performance.* Recuperado de https://aws.amazon.com/es/solutions/ca se-studies/mercadolibre- ec2/?trk=cr_card
- Schulz, G. (2017). *What's a datainfrastructure?* Recuperado de https://www.networkworld.com/article/3 171257/whats-a-data-infrastructure.html
- Confluent. (s.f.). *Differences Between a Data Warehouse, Data Lake, and a Database.* Recuperado de https://www.confluent.io/learn/database-data-lake-data-warehouse-compared/
- Tungagawan, E. (2018). *On Distributed Systems Setup and Architecture Planning.* Recuperado de https://medium.com/cermati-tech/on- distributed-systems-setup-and- architecture-planning-8ad9954fe2c2

Universidad Tecmilenio

Ingeniería de datos masivos

Métodos de agrupaciones de datos

Semana 7

Introducción

La tarea de encontrar agrupaciones significativas para *customizar* acciones se vuelve cada vez más importante como estrategia para vender y operar en las empresas.

Teniendo el registro de los comportamientos de compra y las variables demográficas de los clientes, el análisis de agrupamiento puede ayudar a responder la siguiente pregunta:
¿Cuáles son las agrupaciones naturales de los clientes?

Explicación

Agrupación de datos

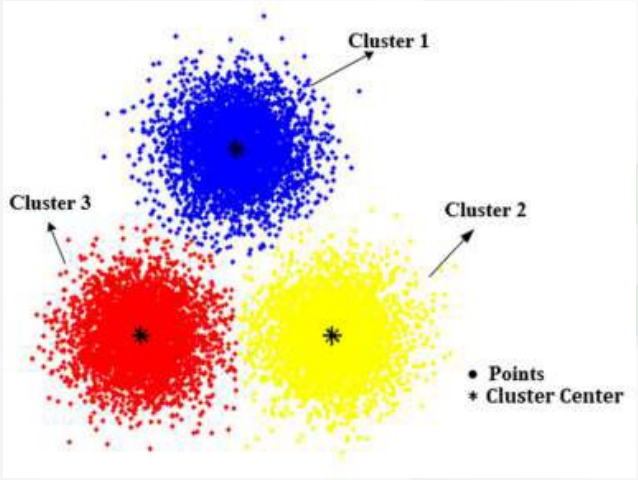
Es un modelo de análisis que consiste en dividir todos los datos en grupos (clústeres) en función de los patrones de los datos. Asimismo, estos comparten características similares, ya que tienen una distancia similar a una referencia común de todos los elementos del clúster.
A menudo se usa como un análisis previo para hacer una clasificación.

La agrupación es un modelo de análisis no supervisado (Packt, 2016), es decir, dentro de los datos no hay categorías previamente definidas que sirvan de referencia para aprender de ellas y así poder asignar una categoría a nuevos datos. El problema central que se tiene que resolver al usar una agrupación es analizar la estructura de los datos y determinar la mejor manera de agruparlos.

Los modelos de agrupación más usados son el **K-Means** (que es un modelo basado en centroides o puntos centrales dentro de los clústeres) y la **agrupación jerárquica**.

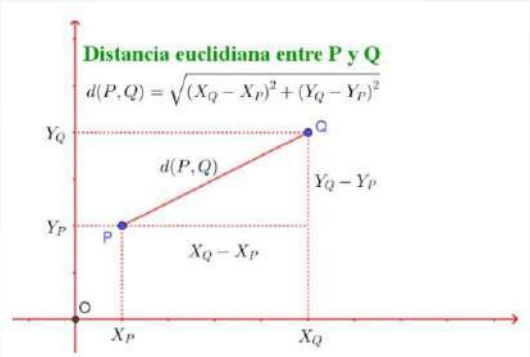
Explicación

K-Means (Lopez, 2019) es el algoritmo en el cual cada clúster está asociado a un centro llamado centroide. El objetivo principal del algoritmo es minimizar la suma de las distancias de los elementos agrupados en un clúster con su centroide, siendo este un algoritmo iterativo.



Fuente: Zhang, J., Chen, W., Gao, M., y Shen, G. (2017). *K-means-clustering- based fiber nonlinearity equalization techniques for 64-QAM coherent opticalcommunication system.* Recuperado de https://www.osapublishing.org/oe/fulltext.cfm?uri=oe-25-22-27570&id=375887

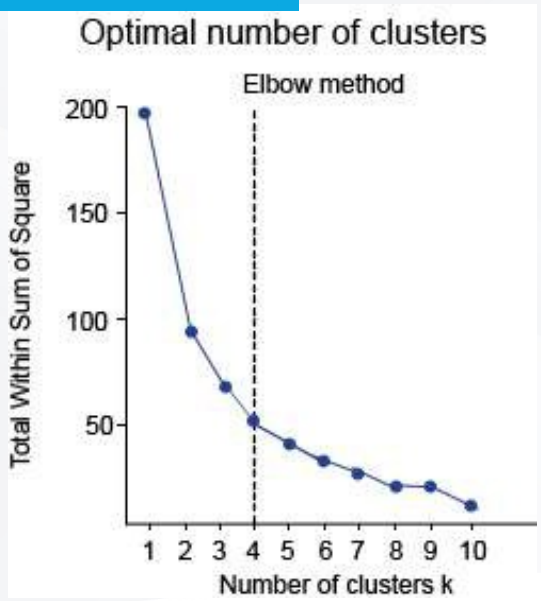
La **distancia euclidiana** es la distancia entre dos puntos que se deduce a partir del teorema de Pitágoras. La distancia euclidiana entre dos puntos: P(XP, YP) y Q(XQ, YQ) es la siguiente:



Fuente: Pérez, R. (s.f.). *Distancia euclidiana: concepto, fórmula, cálculo, ejemplo.* Recuperado de https://www.lifeder.com/distancia-euclidiana/

Explicación

El **método Elbow** sirve para determinar un número óptimo de clústeres, el cual consiste en hacer una gráfica del número de clústeres y el error, para después analizar la gráfica y determinar el número de clústeres en el que el error es menor.

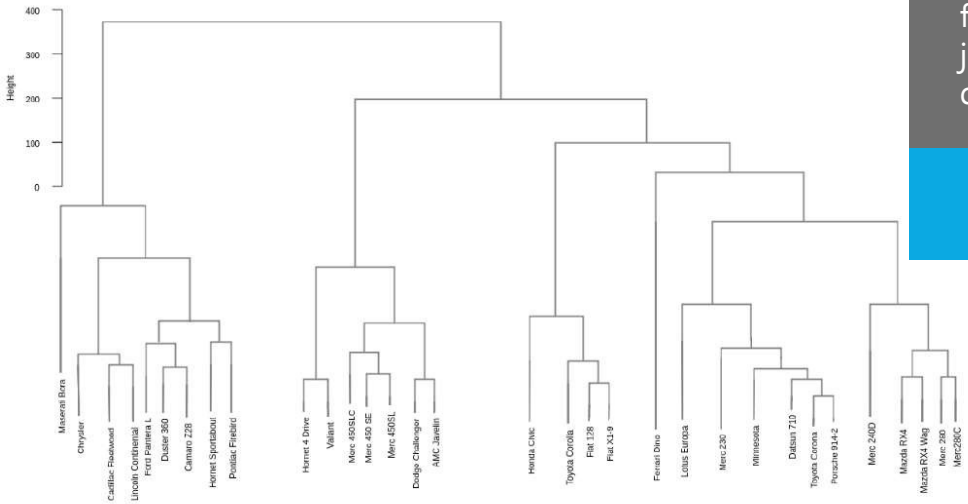


Fuente: Educba. (s.f.). *K- Means Clustering Algorithm.* Recuperado de https://www.educba.com/k-means-clustering-algorithm/

Explicación

Agrupación jerárquica

Es un modelo que construye una jerarquía de agrupaciones, la cual inicia con todos los datos asignados a un clúster propio para que después los clústeres más cercanos se fusionen en un mismo cluster, construyendo una jerarquía de clústeres en una gráfica conocida como dendrograma (Reddy, 2018).



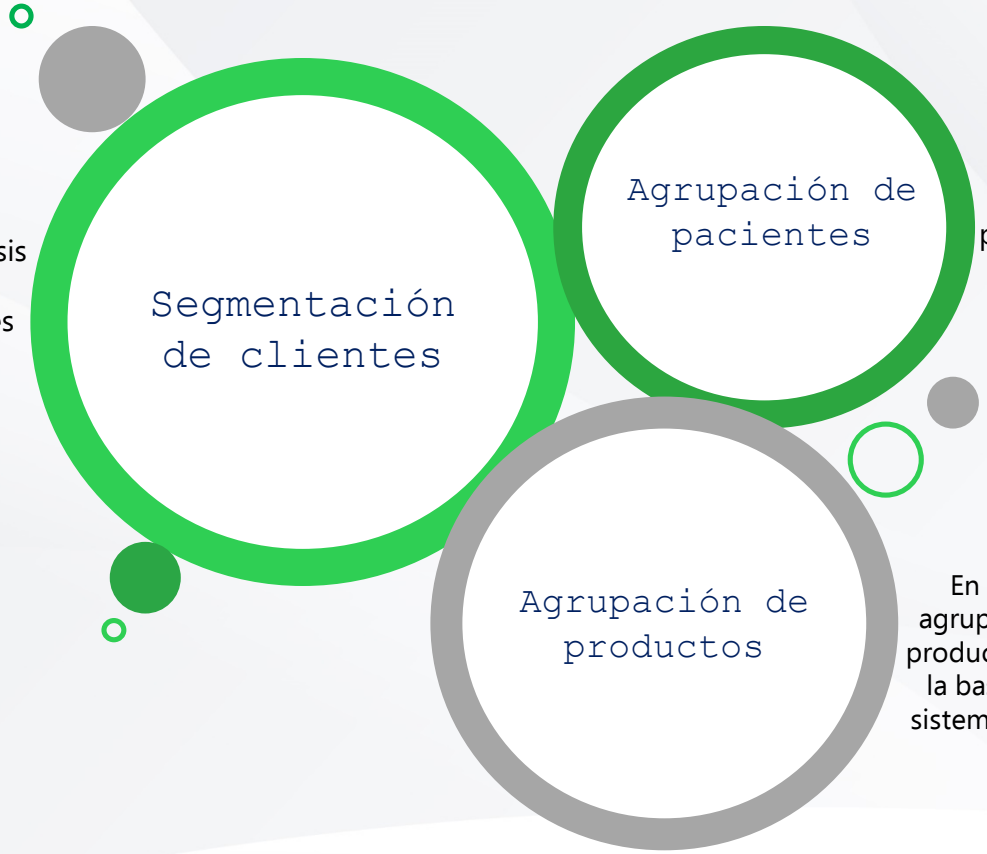
Fuente: Visual Paradigm. (s.f.). *Dendrogram Example: Cars Cluster Dendrogram*. Recuperado de https://online.visual-paradigm.com/diagrams/templates/dendrogram/cars-cluster-dendrogram/

Para conocer un ejemplo de agrupación de datos en Python, ve el siguiente video:

Dirección de Producción de Contenidos. (2020, 3 de agosto). *Tema 7. Agrupación de datos de datos en Python*. [Archivo de video]. Recuperado de https://www.youtube.com/watch?v=quoT4mJsYRY&feature=youtu.be

Explicación

Usos de agrupación de datos



Proceso de hacer el análisis para encontrar las similitudes en los clientes para que estos puedan agruparse, definiendo segmentos.

Útil en biología para la clasificación de plantas y animales, así como en la genética humana, ya que diversos atributos del paciente se pueden utilizar para hacer agrupaciones.

En una selección con agrupación se encuentran productos similares que son la base del análisis de un sistema de recomendación.

Cierre

La agrupación de datos o *clustering* es un modelo de análisis no supervisado que permite localizar y clasificar grupos de datos (también llamados clústeres), donde los elementos que comparten características similares se ubican dentro de un mismo grupo.

Asimismo, los modelos de agrupación más utilizados son K-Means y la agrupación jerárquica (*hierarchical clustering*).

Bibliografía

- Educba. (s.f.). *K- Means Clustering Algorithm.* Recuperado de https://www.educba.com/k-means- clustering-algorithm/
- Dirección de Producción de Contenidos. (2020, 3 de agosto). *Tema 7. Agrupación de datos de datos en Python.* [Archivo de video]. Recuperado de https://www.youtube.com/watch?v= quoT4mJsYRY&feature=youtu.be
- Packt. (2016). *Introduction to Clustering and Unsupervised Learning.* Recuperado de https://hub.packtpub.com/introduction- clustering-and-unsupervised- learning/#more
- Pérez, R. (s.f.). *Distancia euclidiana: concepto, fórmula, cálculo, ejemplo.* Recuperado de https://www.lifeder.com/distancia- euclidiana/
- Lopez, D. (2019). *A complete guide to K- means clustering algorithm.* Recuperado de https://www.kdnuggets.com/2019/05/gui de-k-means-clustering-algorithm.html
- Reddy, C. (2018). *Understanding the concept of Hierarchical clustering Technique.* Recuperado de https://towardsdatascience.com/underst anding-the-concept-of-hierarchical- clustering-technique-c6e8243758ec
- Visual Paradigm. (s.f.). *Dendrogram Example: Cars Cluster Dendrogram.* Recuperado de https://online.visual-paradigm.com/diagrams/templates/den drogram/cars-cluster-dendrogram/

Universidad Tecmilenio

Ingeniería de datos masivos

Técnicas de visualización y sumas numéricas

Semana 7

Introducción

La visualización de datos consiste en la exposición de los mismos en un formato gráfico que permite a las personas que toman decisiones contemplar el análisis representado en forma visual.

En la actualidad, tenemos herramientas que facilitan la creación de visualizaciones de una gran cantidad de datos, los cuales pueden ser de alto nivel, como la sumarización, la agrupación y la contextualización.

Explicación

Antes de comenzar el tema, plantéate las siguientes preguntas de reflexión:

¿Qué es visualización de datos?

¿Cómo se hace una visualización de datos?

¿Para qué se usa la visualización de datos?

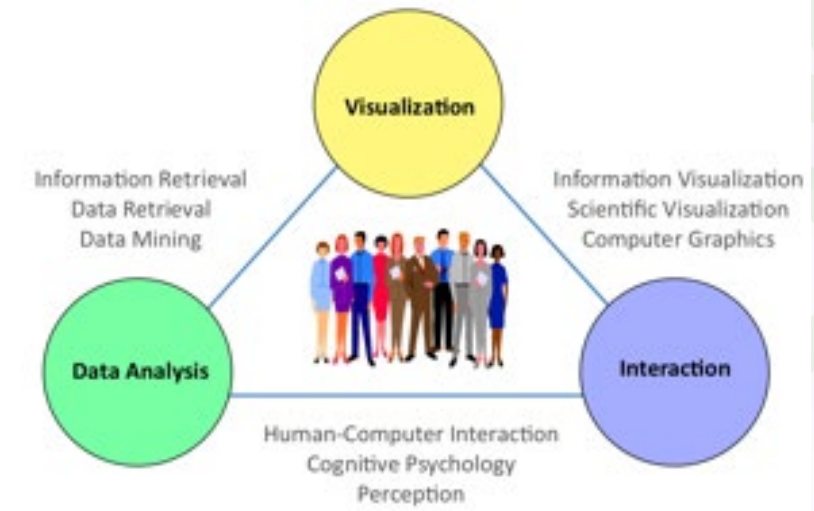
Hoy en día existen diversas herramientas para la visualización de datos, por ejemplo:

- **Qlick**
- **Power BI**
- **Tableau**

Explicación

La **visualización de datos** es una representación gráfica de la información y de los datos. Mediante el uso de elementos visuales, como gráficos y mapas, la visualización de datos ofrece una manera accesible para detectar y comprender las tendencias, los valores atípicos y los patrones en los datos (Tableau, s.f.).

Se puede realizar analítica de datos usando solamente visualizaciones de datos, para lo cual existe el **visual analytics**, que es la ciencia del análisis racional soportado por una interfaz visual e interactiva (Bose, 2018).



Fuente: Visual-Analytics.eu. (s.f.). *What is Visual Analytics?* Recuperado de https://visual-analytics.eu/faq/

Explicación

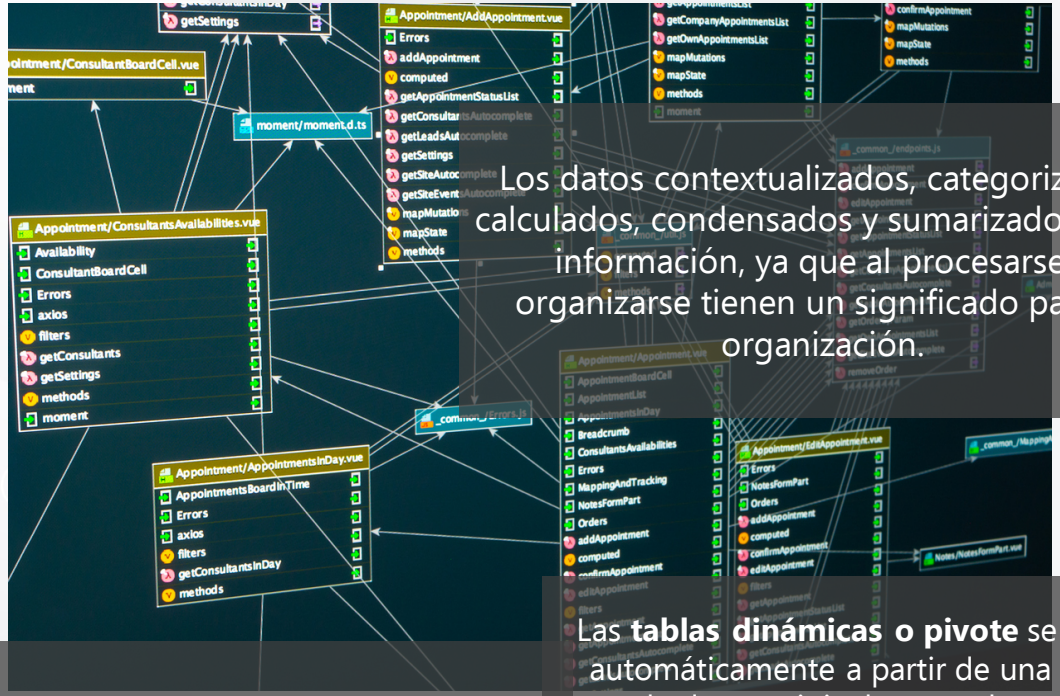
Las visualizaciones de datos se han convertido en un nuevo lenguaje en los negocios, ya que una buena visualización puede comunicar la fuerza y el impacto potenciador que la información le da a las ideas.

El objetivo de las visualizaciones de datos es ayudar a comprenderlos. Además, se usan en el análisis exploratorio de los datos, que sirve para visualizar distribuciones, medidas centrales, desviaciones, correlaciones, valores atípicos, así como para descubrir nuevos patrones o relaciones en los datos.

Las mejores prácticas para crear buenas visualizaciones (con el objetivo de comunicar) consideran aspectos de percepción visual, neurociencia, estética, comunicación efectiva y las reacciones, así como los sentimientos que se generan al verlas.

Inteligencia de negocios (*business intelligence*) es hacer visualizaciones y *dashboards* para crear información a partir de datos, es decir, hacer analítica descriptiva con dashboards de datos históricos para *sumarizar*, resumir y describir.

Explicación



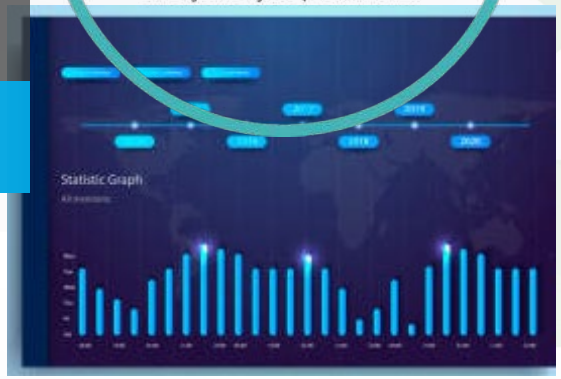
Los datos contextualizados, categorizados, calculados, condensados y sumarizados crean información, ya que al procesarse y organizarse tienen un significado para la organización.

Un **dashboard** (King, 2018) es una representación gráfica de los indicadores más sobresalientes (KPI) que intervienen para lograr los objetivos de una organización, en donde los ejecutivos reciben todas las herramientas necesarias para realizar un análisis profundo, logrando tomar decisiones para potencializar la estrategia de la organización.

Las **sumas numéricas** permiten sumarizar datos para presentar totales agrupados y filtrados.

Las **tablas dinámicas o pivote** se crean automáticamente a partir de una tabla de datos originales, creada para clasificar, filtrar, contar, totalizar o hacer operaciones numéricas con los datos agrupados y filtrados.

Un **KPI** (Key Performance Indicator) es un indicador clave que mide el desempeño y es una medida del nivel del rendimiento de un proceso. El valor del indicador está directamente relacionado con un objetivo fijado previamente.



Explicación

Cuando el objetivo de las visualizaciones de datos es comunicar resultados, ideas, hallazgos o persuadir contando una historia para influenciar la toma de decisiones, se usan las recomendaciones de *storytelling* para crearlas.

Data storytelling es comunicar ideas y hallazgos de los datos y se involucran tres elementos (Dykes, 2016):

- Datos
- Narrativa
- Visualizaciones

Una de las mejores prácticas para construir esa narrativa es seguir la estructura que se presenta a continuación:

- ✓ Describir el contexto (*setup*).
- ✓ Presentar el reto o situación a resolver (*conflict*).
- ✓ Presentar la solución o conclusión (*resolution*).

Recomendaciones para construir este tipo de visualizaciones

1. Entender el concepto.
2. Usar visualizaciones efectivas.
3. Eliminar los elementos confusos, excesivos o distractores.
4. Llevar la atención a lo que es importante.
5. Pensar y usar la estética de un diseñador.
6. Contar una historia.

Cierre

La visualización de datos consiste en la presentación de datos de forma gráfica o ilustrada, permitiendo a las personas que toman decisiones en las organizaciones ver la analítica presentada de forma visual, logrando identificar más fácilmente nuevos patrones, comprender tendencias y valores atípicos, ya que los datos tienen un gran valor al poder ser visualizados, permitiendo a los profesionales comunicar eficazmente su significado mediante el storytelling, involucrando los tres elementos clave: datos, visualizaciones y narrativa.

Bibliografía

- Bose, B. (2018). *What is Visual Analytics: Key Concepts*. Recuperado de https://www.digitalvidya.com/blog/visual-analytics/
- Visual-Analytics.eu. (s.f.). *What is Visual Analytics?* Recuperado de https://visual-analytics.eu/faq/
- Dykes, B. (2016). *Data Storytelling: The Essential Data Science Skill Everyone Needs.* Recuperado de https://www.forbes.com/sites/brentdyke s/2016/03/31/data-storytelling-the- essential-data-science-skill-everyone-needs/#35aaecc852ad
- Tableau. (s.f.). *Guía de visualización de datos: definición, ejemplos y recursos de aprendizaje.* Recuperado de https://www.tableau.com/es- es/learn/articles/data-visualization