Sistemas de Big Data

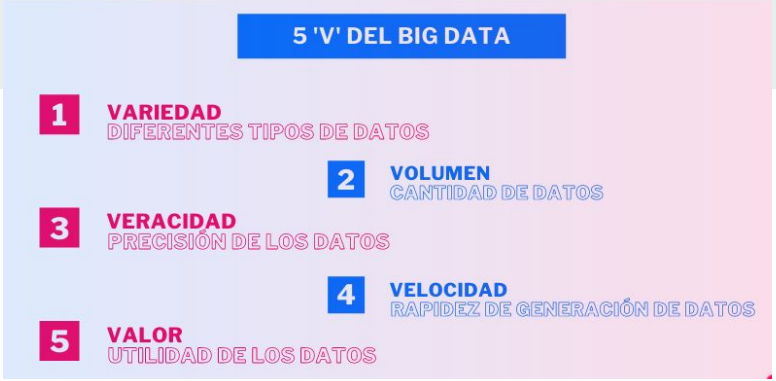
Curso de especialización en Inteligencia Artificial y Big Data.

Programación

Introducción - Big Data

No existe una definición precisa del término pero los términos de datos masivos o grandes volúmenes de datos hacen referencia al big data. Por este motivo, a menudo el concepto de big data es definido en función de las características que poseen los datos y los procesos que forman parte de este nuevo paradigma de computación. **Esto es lo que se conoce como las Vs del Big Data.**

Introducción - Big Data

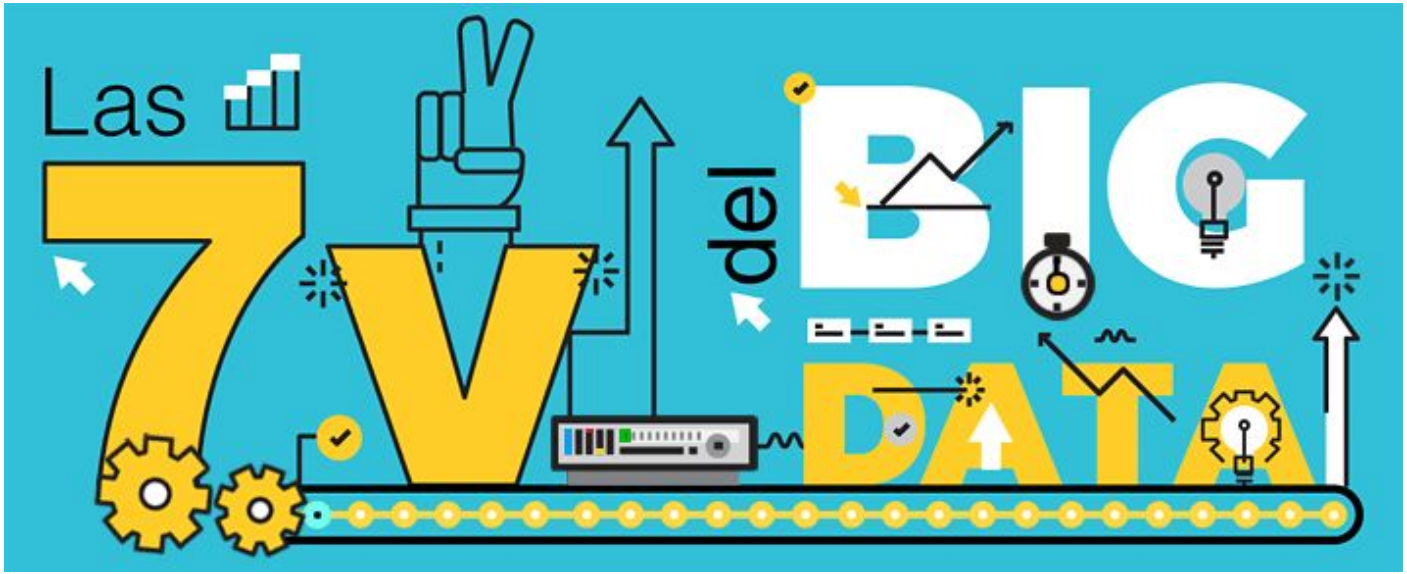


Las 5Vs del Big Data:

- **Volumen:** La cantidad masiva de datos generados a cada segundo.
- **Velocidad:** La rapidez con la que los datos se crean y procesan.
- **Variedad:** Diferentes tipos de datos (estructurados, no estructurados, semi-estructurados).
- **Veracidad:** La fiabilidad de los datos para garantizar su utilidad.
- **Valor:** El potencial valor que los datos pueden ofrecer al ser procesados y analizados.

Empresas como Amazon y Google usan Big Data para personalizar sus servicios y tomar decisiones de negocio.

Introducción - Big Data



Modelos de negocio basado en datos

El modelo de negocio se refiere a cómo una organización crea y captura valor. Los modelos de negocio basados en Big Data se encuentran actualmente en una fase incipiente.

Tabla 1.3: Tipología de BDBM (fuente [WSM20])

Tipo	Fuente de valor (ejemplo)
Usuarios de datos	• Analisis BD para apoyar la toma de decisiones estratégicas • Uso de BD para mejorar los procesos internos • Enriquecer productos, servicios y experiencia de clientes mediante BD • Desarrollo de nuevos productos y servicios mediante BD
Proveedoras de datos	• Recopilando datos primarios y vendiéndolos a terceros • Agregando datos y empaquetando datos internos para la venta
Facilitadores de BD	• Ofreciendo la infraestructura a las anteriores tipos de empresa necesaria para realizar BD • Consultoría relativa a BD • Subcontratación de técnicas analíticas para BD (ejemplo en la nube)

Modelos de negocio basado en datos. Empresas orientadas a datos

Una empresa orientada a los datos es aquella que utiliza datos como un recurso estratégico para la toma de decisiones. La adopción de tecnologías de Big Data y análisis de datos **permite a las empresas mejorar sus procesos y tomar decisiones basadas en información precisa y en tiempo real.** Desde empresas emergentes que usan datos para personalizar productos hasta grandes organizaciones que optimizan sus operaciones mediante análisis avanzados.

Evolución de las Empresas hacia el Uso de Big Data:

El uso de Big Data se ha convertido en un factor diferenciador para las empresas competitivas. Las empresas han pasado de **procesos tradicionales a modelos centrados en el análisis de datos,** permitiendo predicciones y mejoras en la eficiencia. **Causas del cambio**:

- Incremento en la cantidad y variedad de datos disponibles.
- Necesidad de tomar decisiones más rápidas.
- Uso de tecnología para gestionar grandes volúmenes de información.

Modelos de negocio basado en datos. Empresas orientadas a datos

Proceso de Transformación hacia una Empresa Basada en Datos:

- **Etapa 1: Exploración** - Identificación de oportunidades y planificación de estrategias basadas en datos.
- **Etapa 2: Implementación** - Adopción de herramientas y plataformas de Big Data.
- **Etapa 3: Optimización** - Uso de análisis avanzados para optimizar procesos.
- **Etapa 4: Innovación continua** - Transformación de la cultura organizativa para aprovechar nuevas tecnologías.

Impacto en la toma de decisiones

Big Data permite a las empresas basar sus decisiones en datos objetivos. **Como por ejemplo Walmart usa análisis de Big Data para gestionar el inventario en tiempo real.**

Tecnologías: Herramientas como Apache Hadoop y Apache Spark permiten el procesamiento eficiente de grandes volúmenes de datos.

Actividad de debate

Impacto en la toma de decisiones

- **Importancia de los datos en Walmart**: Walmart utiliza los datos para mejorar la experiencia de compra de más de 240 millones de clientes cada semana, tanto en sus tiendas físicas como en su plataforma online.
- **Infraestructura y manejo de datos**: Walmart procesa multiterabytes de nuevos datos cada día y gestiona petabytes de datos históricos que cubren millones de productos y cientos de millones de clientes. (Teradata, NoSQL, SAS y Hadoop).
- **Limpieza y privacidad de los datos**
- **Equipos y democratización del acceso**: El “Equipo de Datos Grandes y Rápidos” facilita el acceso y uso de los datos para desarrolladores, científicos de datos y analistas, lo que permite la rápida creación de prototipos y soluciones. → acceso simultáneo a datos anónimos sin procesos burocráticos.
- **Uso de los datos para aplicaciones de cliente**: Han desarrollado aplicaciones que mejoran la experiencia de compra en tiendas físicas, como la búsqueda de artículos, la gestión de recetas y la localización de productos dentro de la tienda.
- **Estrategia de valor basada en datos**: Walmart recomienda identificar qué datos recopilar y enfocar los esfuerzos en áreas que proporcionen mayor valor para el negocio y el cliente.

Complejidad computacional para el análisis de datos

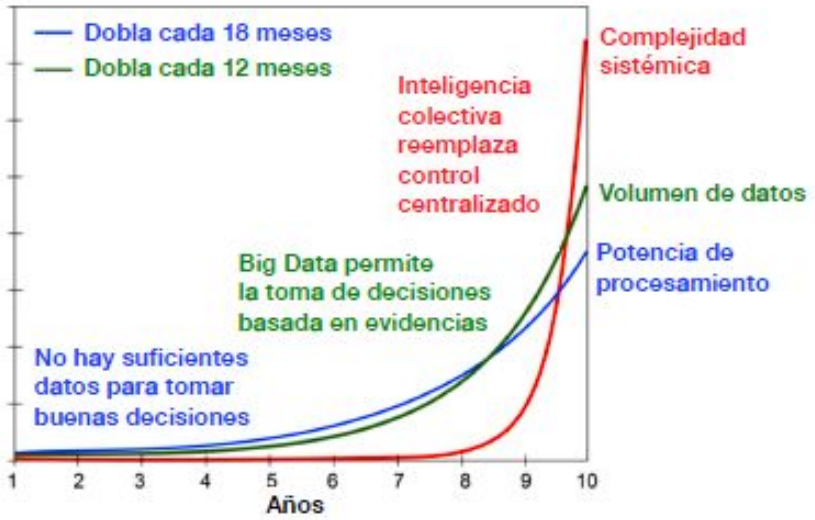
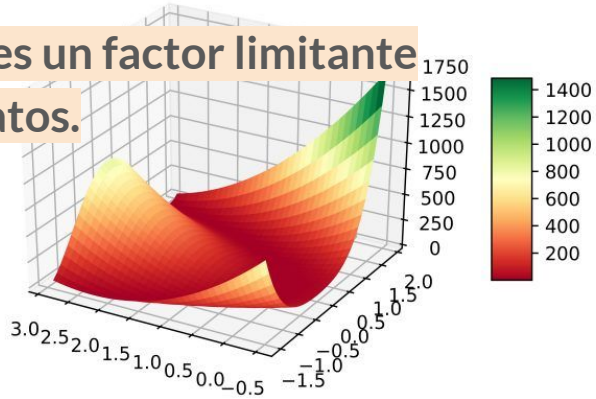


Figura 1.1: Modelo de crecimiento digital, fuente [HPG+19].

Complejidad computacional para el análisis de datos

El **crecimiento del coste de los procesos para analizar datos**, aspecto que incluyen los algoritmos, tiene un **crecimiento factorial** cuyo consecuencia inmediata es una gran cantidad de datos que no podrán ser analizados. Esta complejidad explosiva impone un modelo de inteligencia distribuida para poder afrontar los retos del mundo y de nuestras sociedades.

El coste computacional de la ejecución de un algoritmo es un factor limitante incluso cuando no se está ante grandes volúmenes de datos.



Complejidad computacional para el análisis de datos

Supongamos que:

- **el número de datos disponibles es n**
- **si el número de datos n aumenta entonces el tiempo de ejecución de un algoritmo para procesarlos también aumente**.

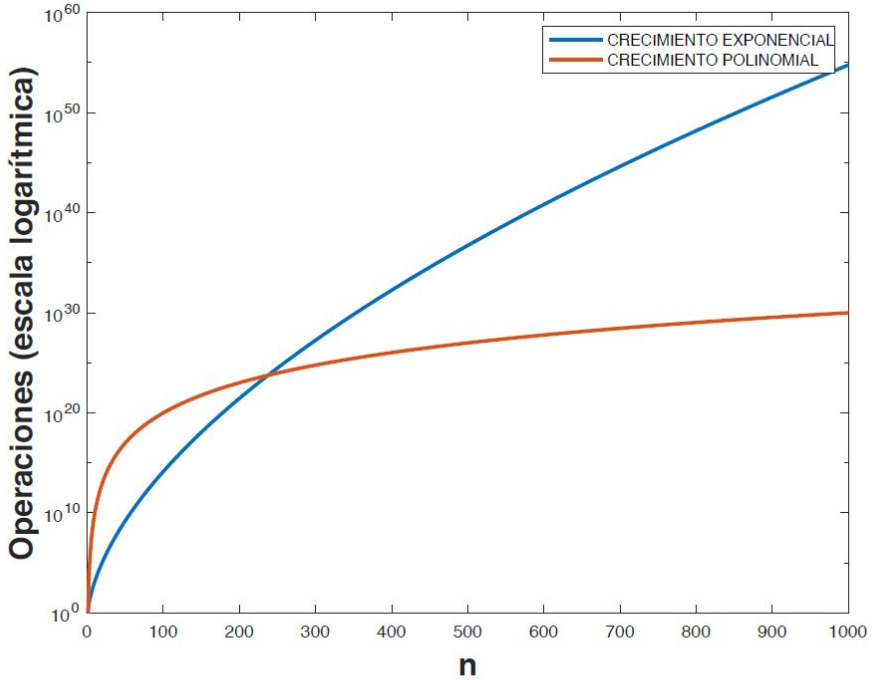
Por tanto cualquier algoritmo tiene la propiedad que **si n crece hacia infinito su coste computacional también crece hacia infinito**. La cuestión esencial es que existen varias velocidades de crecer a infinito, unas significativamente más rápidas que otras. La relación matemática

$$\lim_{n \to +\infty} \frac{P(n)}{a^n} = 0$$

para cualquier a > 1 y para cualquier polinomio P(n) indica que aunque tanto el numerador como el denominador tienden a infinito el denominador crece muchísimos más rápido que el numerador. El algoritmo se volverá cada vez más lento y puede ser impracticable para conjuntos de datos grandes

Complejidad computacional para el análisis de datos

Crecimiento polinomial vs exponencial



Complejidad computacional para el análisis de datos

Supongamos que disponemos del ordenador de IBM *Roadrunner* que es capaz de superar un *petaflop* de operaciones por segundo. Es capaz de realizar $1{,}105 * 10^{15}$ operaciones por segundo. ¿Cuánto tiempo emplearía en factorizar un número de 50 cifras?

$$\text{VelocidadCPU} = 1{,}105 * 10^{15} \text{ operaciones / segundos}$$

$$\frac{f_{\mathcal{A}}(50)}{\text{VelocidadCPU}} = 1{,}4907 * 10^{-6} \text{ segundos}$$

El resultado muestra que tardaría menos de una milésima de segundo. Y si nos preguntamos cual es el máximo número de cifras n que *Roadrunner* puede factorizar durante un año. Planteamos el siguiente código (ver Listado 1.1) en Python para resolver la cuestión.

Complejidad computacional para el análisis de datos

Listado 1.1: Código en python para calcular el máximo número de cifras de una clave RSA que se puede romper con el ordenador Roadrunner

```
# Carga de la libreria math para calcular log10
from math import *
# Operaciones/segundo realizadas por Roadrunner
VelocidadCPU = 1.105 * 10 ** 15
# número de dígitos
n=1
# número de operaciones
fA=1
# ¿ nº operaciones requeridas <nº operaciones en un año?
while  fA <= (VelocidadCPU *24 * 60 * 60 *365):
    n = n+1
    fA = 10 ** ( pow( n * log(n,10) , 0.5 ) )

print('Numero maximo de cifras  =' , n-1)
```

El resultado de ejecutarlo es $n = 217$. Hemos pasado de factorizar un número de 50 cifras en una millonésima de segundo a necesitar un año para factorizar un número con un poco mas del cuádruple de cifras. Nos podemos plantear la cuestión

Complejidad computacional para el análisis de datos

La duplicación de los recursos computacionales no es una estrategia por sí sola que sea capaz de abordar el coste computacional de los problemas.

Este ejemplo muestra que el un tiempo de ejecución exponencial es inabordable para cualquier máquina actual. De hecho a efectos prácticos una complejidad computacional superior a n3 es dramática.

Complejidad computacional para el análisis de datos

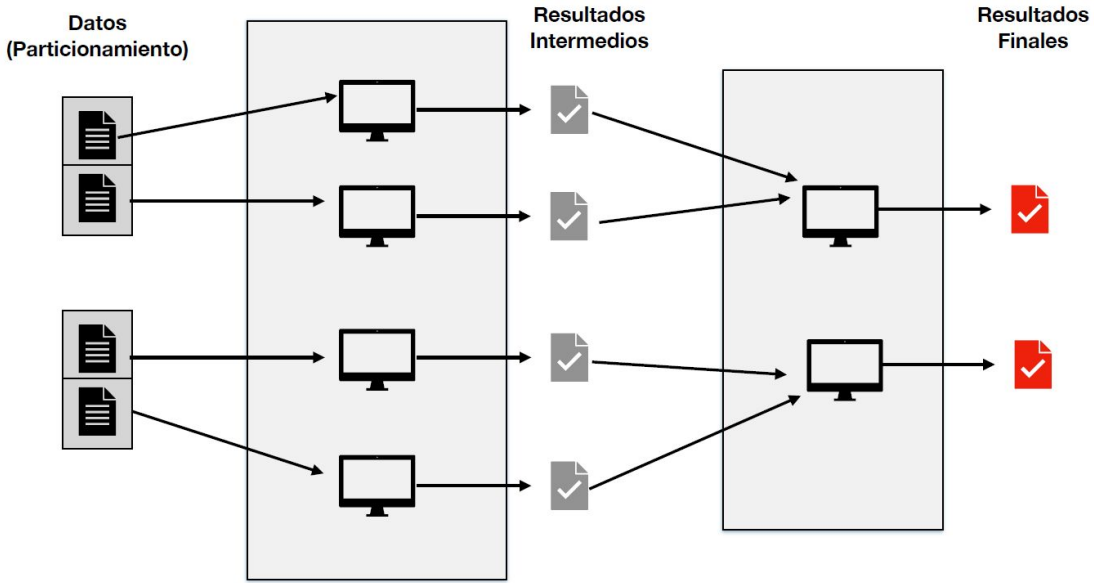


Figura 1.3: Computación distribuida.

Complejidad computacional para el análisis de datos

Velocidad de ejecución de los lenguajes de programación. enlace

Puesto	Lenguaje	Características Clave	Ventajas	Desventajas
1	C	Cercano al ensamblador, simple, muy rápido	Máxima velocidad	Poco seguro, propenso a errores de búfer
2	C++	Orientado a objetos, potente, STL	Muy rápido, versátil	Complejo, puede ser menos seguro que C
3	Rust	Seguro, concurrente, sin recolector de basura	Rápido, seguro, moderno	Comunidad más pequeña que otros lenguajes
4	Fortran	Especializado en cálculos científicos	Alto rendimiento en cálculos numéricos	Menos versátil que otros lenguajes
5	Julia	Alto rendimiento, dinámico, fácil de usar	Rápido, adecuado para ciencia de datos y machine learning	Comunidad más pequeña
6	Ada	Seguro, confiable, orientado a objetos	Muy confiable, adecuado para sistemas críticos	Menos popular que otros lenguajes
7	Java	Plataforma cruzada, orientado a objetos	Muy popular, gran ecosistema	Puede ser más lento que C o C++ en algunas tareas
8	C#	Orientado a objetos, moderno, .NET	Versátil, potente	Puede ser más lento que C++ en algunas tareas
9	F#	Funcional, .NET, interoperable con C#	Bueno para programación funcional	Comunidad más pequeña que C#
10	Pascal	Procedural, tipado estático	Rápido, sencillo	Menos popular en la actualidad

Aplicación de técnicas de integración, procesamiento y análisis de la información

El rápido avance de las tecnologías digitales en las últimas décadas ha llevado a una **generación exponencial de datos**. **Se estima que el 90% de todos los datos disponibles actualmente fueron generados en los últimos 2 años**. Para transformar estos enormes volúmenes de datos en información útil **surgen las técnicas modernas de procesamiento y análisis de información.**

Estas técnicas se aplican en diversas industrias:

- en **investigación científica** permite analizar grandes conjuntos de datos experimentales para **obtener nuevos insights**;
- en **salud** posibilita el procesamiento de historiales clínicos para la **detección temprana de enfermedades**;
- en **comercio** viabiliza la personalización de contenidos y productos de acuerdo a los **intereses de los usuarios**; etc.

Técnicas modernas de tratamiento de datos

Algunas de las técnicas más relevantes son:

- **Big Data: enfocada en el almacenamiento, administración y procesamiento de enormes conjuntos de datos, tanto estructurados como no estructurados**. Utiliza tecnologías como Apache Hadoop, Spark y bases de datos NoSQL.

- **Data Mining: busca descubrir patrones repetitivos y relaciones entre variables en grandes bases de datos,** utilizando algoritmos estadísticos y de machine learning. Algunas herramientas son Orange, Weka y KNIME.

- **Business Intelligence:** conjunto de estrategias y herramientas **enfocadas en el análisis de datos empresariales** para facilitar una mejor toma de decisiones. Algunas soluciones populares son Tableau, Qlik Sense y Microsoft Power BI.

Técnicas y procesos de extracción de la información de los datos

El proceso típico de extracción de información consta de:

- **Recolección:** obtención de los datos desde diferentes fuentes.
- **Limpieza:** corrección de errores, manejo de valores faltantes, etc.
- **Transformación:** conversión a formatos adecuados para el análisis.
- **Modelado y análisis:** aplicación de algoritmos para descubrir patrones y relaciones. Por ejemplo, la clasificación mediante árboles de decisión.
- **Interpretación:** traducción de los resultados del modelo analítico en insights aplicables para la toma de decisiones.

Técnicas y procesos de extracción de la información de los datos

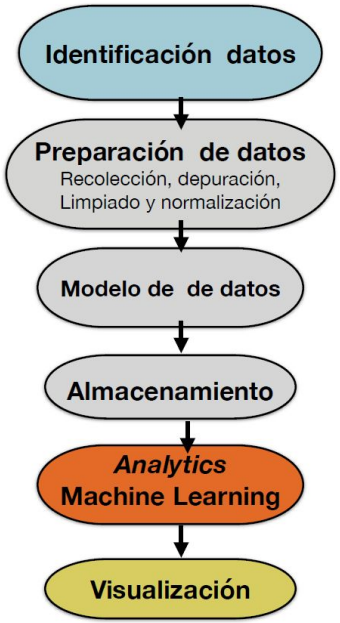


Figura 1.4: Proceso de extracción de información de los datos.

Técnicas y procesos de extracción de la información de los datos. Caracterización del dato

El análisis y la caracterización del dato son fundamentales para comprender su utilidad y cómo debe ser procesado en un entorno de Big Data. Los datos pueden ser clasificados de acuerdo a diferentes características como el tipo, formato, generador, tamaño, rol, latencia y sensibilidad.

Datos en cuanto al Tipo

- Datos Simples: Son aquellos que representan un único valor y no están compuestos por múltiples elementos.
- Datos Compuestos: Conformados por múltiples elementos que pueden ser de diferentes tipos.

Técnicas y procesos de extracción de la información de los datos. Caracterización del dato

Datos en cuanto al Formato

- Datos Estructurados: Tienen un formato fijo y predefinido, generalmente organizados en **tablas con filas y columnas**. Su almacenamiento se registra en tablas. Si el volumen es grande se **almacenan en Data Warehouse,** en otros casos se emplean bases de datos relacionales o simples hojas de cálculo. En los modelos de bases de datos relacionales toda la información se almacena en tablas en las que se especifican el tipo de campos que tienen y cómo se relacionan entre ellas.

- Datos Semi-Estructurados: Tienen una **estructura flexible y no fija,** pero contienen **etiquetas o delimitadores que los organizan**. Ejemplos de este tipo de datos son las páginas web o servicios de correos electrónicos donde se almacenan los mensajes y ficheros adjuntos. **Los metadatos permiten clasificarlos y realizar búsquedas por palabras clave.**

- Datos No Estructurados: Carecen de un formato predefinido y no se ajustan fácilmente a tablas. El **almacenamiento de este tipo de datos debe realizarse de forma organizada a través de una base de datos no relacional (NoSQL).**

Técnicas y procesos de extracción de la información de los datos. Caracterización del dato

Datos en cuanto al Generador

- **Datos Generados por Personas**: Provienen de actividades humanas y reflejan comportamientos, interacciones o preferencias.
- **Datos Generados por Máquinas**:Producidos automáticamente por dispositivos, sensores o procesos industriales.

Datos en cuanto al Tamaño

- Datos **Pequeños**: Conjuntos de datos que pueden ser almacenados y procesados en un solo sistema.
- Datos **Medianos**: Pueden almacenarse en una sola máquina, pero su procesamiento es más costoso.
- Datos **Grandes**: Superan la capacidad de procesamiento de una sola máquina y requieren clústeres distribuidos.
- Datos Muy Grandes: Escalan a nivel de petabytes o exabytes.

Técnicas y procesos de extracción de la información de los datos. Caracterización del dato

Datos en cuanto a su Rol

- **Datos Maestros**: Son los datos esenciales que describen los elementos más importantes de un negocio (clientes, productos, empleados).
- **Datos Operacionales**: Datos que registran las operaciones diarias de una organización.
- **Datos Externos**: Datos que provienen de fuera de la organización y se integran para enriquecer el análisis.
- **Datos derivados de la transformación y el análisis de datos operacionales o externos**.

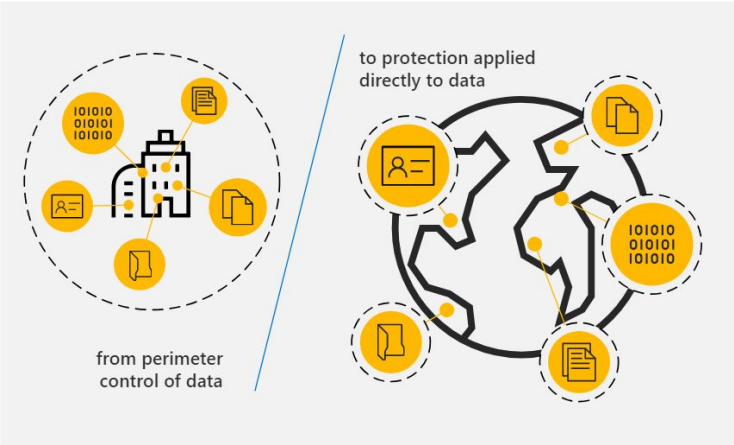
Datos en cuanto a su Latencia

- **Datos en Tiempo Real**:Se procesan y analizan en el momento en que se generan, con mínima latencia.
- **Datos por Lotes**: Se almacenan y luego se procesan en conjunto a intervalos definidos.

Técnicas y procesos de extracción de la información de los datos. Caracterización del dato

Datos en cuanto a su Sensibilidad

- Datos con **Riesgo Alto:** Datos altamente confidenciales que requieren protección estricta.
- Datos con **Riesgo Medio**: Datos con valor para la organización, pero no críticos.
- Datos con **Riesgo Bajo:** Datos que no requieren protección especial y su pérdida no afecta significativamente a la organización.



Técnicas y procesos de extracción de la información de los datos. Recopilación, extracción de datos.

El primer paso es **determinar las fuentes de datos que pueden ser útiles**.

Estas fuentes abarcan los **datos generados en el propio entorno empresarial** como aquellas que son **ajenas a la organización**, como redes sociales, web, bases de datos compradas a otras compañías, datos geográficos, datos económicos sectoriales, etc. Tras este proceso se dispondrá de un conjunto de datos que abarcarán una o varias de las siguientes tipologías: estructurados, no estructurados y semi-estructurados.

Herramientas de extracción de datos

Extracción de datos de fuentes no estructuradas:

- **ParseHub:** permite extraer datos de sitios web mediante un interfaz visual de arrastrar y soltar.
- **Import.io:** extrae datos de web scraping mediante un proceso guiado por el usuario.
- **MonkeyLearn:** extrae información de texto plano utilizando modelos de machine learning.

Para datos estructurados y bases de datos:

- **Knime:** plataforma de integración y análisis de datos con módulos para ETL, mining e informes.
- **Trifacta:** herramienta de preparación de datos que limpia, transforma y enriquece datasets.
- **Pentaho:** solución open source de business intelligence con capacidades de ETL y análisis de datos.

Análisis en tiempo real

El análisis en **tiempo real** tiene aplicaciones como:

- **Detección de fraudes:** analiza transacciones financieras para identificar en segundos comportamientos anómalos o sospechosos.

- **Monitoreo industrial:** sensores IoT envían continuamente datos de fábricas que son analizados en tiempo real para detectar fallas o necesidades de mantenimiento.

- **Personalización de contenidos:** las interacciones de los usuarios en un sitio web se analizan al instante para ofrecer resultados de búsqueda, productos y anuncios adaptados a sus intereses.

Costes y calidad

- Los costes del análisis de datos pueden incluir: licencias de software especializado, infraestructura en la nube, consultoría de expertos, capacitación de personal, que pueden variar entre algunos miles a cientos de miles de euros dependiendo de la complejidad.

- **La calidad de los datos fuente es crucial. Los problemas más comunes son:**
  - datos incompletos,
  - ruidosos,
  - duplicados,
  - desactualizados,
  - incorrectamente formateados.

Si no se detectan y corrigen pueden sesgar los resultados del análisis.

Actividad

Indica qué aplicaciones de Big Data podemos encontrarnos en cada uno de los siguientes sectores:

1. Distribución
2. Marketing
3. Logística
4. Telefonía
5. Energía
6. Salud
7. Seguridad
8. Medios de Comunicación
9. Finanzas
10. Seguros
11. Servicio al Cliente

Modelos de datos y almacenamiento. Recopilación en bruto de datos y su preprocesamiento.

1. Limpieza de Datos (Data Cleaning): Durante la recopilación de datos, es común encontrar datos erróneos o inconsistentes. La limpieza de datos implica la identificación y corrección de estos problemas.

Esto podría incluir la eliminación de registros duplicados, la corrección de errores tipográficos o la estandarización de formatos de datos.

Modelos de datos y almacenamiento. Recopilación en bruto de datos y su preprocesamiento.

1. Limpieza de Datos (Data Cleaning):

- **Datos duplicados:** Utilizar algoritmos de deduplicación para identificar y eliminar los registros duplicados, manteniendo solo una entrada por cliente.

- **Errores tipográficos:**Implementación de diccionarios para corregir automáticamente los errores ortográficos o crear reglas de validación para detectar y corregir estos errores manualmente.

- **Formatos de datos inconsistentes:**Establecer un formato estándar para las fechas de nacimiento y convertir todos los datos a ese formato.

- **Valores faltantes:** Eliminar los registros con valores faltantes críticos (si no hay suficiente información), imputar los valores faltantes utilizando métodos estadísticos (por ejemplo, la media o la mediana), o marcar los valores faltantes como "desconocido".

- **Datos inconsistentes:** Realizar una verificación de consistencia entre los diferentes campos para identificar y corregir estas discrepancias.

Modelos de datos y almacenamiento. Recopilación en bruto de datos y su preprocesamiento.

2. Detección de Outliers: Los outliers son **valores atípicos que se desvían significativamente del resto de los datos y pueden distorsionar los análisis.** Se utilizan métodos estadísticos para identificar estos valores atípicos y se decide si deben ser ignorados o tratados de alguna manera (por ejemplo, reemplazándolos por valores más representativos o eliminándolos si son errores evidentes).

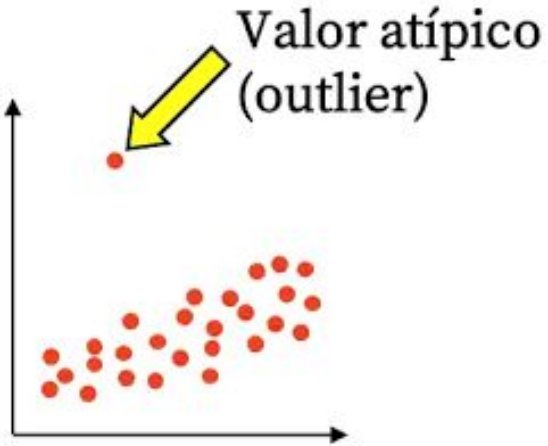
Modelos de datos y almacenamiento. Recopilación en bruto de datos y su preprocesamiento.

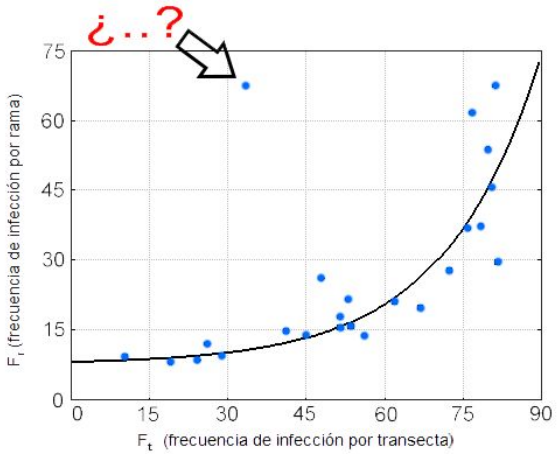
2. Detección de Outliers

Métodos para detectar outliers:

- **Gráficos:**
  - **Diagrama de caja:** Muestra la distribución de los datos y permite identificar fácilmente los valores que se encuentran fuera de los límites superior e inferior.
  - **Histograma:** Muestra la frecuencia de diferentes valores. Los outliers suelen aparecer como barras aisladas en los extremos del histograma.
- **Estadísticas descriptivas:**
  - **Rango intercuartílico (IQR):** Se utiliza para calcular los límites superior e inferior y determinar si un valor es un outlier.
  - **Desviación estándar:** Mide la dispersión de los datos. Los valores que se encuentran a más de 3 desviaciones estándar de la media suelen considerarse outliers.

Modelos de datos y almacenamiento. Recopilación en bruto de datos y su preprocesamiento.






```
Salario (en dólares):
50000
52000
55000
60000
2000000  <-- Outlier
57000
59000
```

Modelos de datos y almacenamiento. Recopilación en bruto de datos y su preprocesamiento.

3. Manejo de Valores Faltantes (Missing Values): Los valores faltantes en los datos son comunes y deben ser abordados adecuadamente. Esto implica tomar decisiones sobre cómo tratar esos valores. Algunas opciones incluyen:

- Relleno con la media o mediana de la columna.
- Estimación basada en otros datos relacionados.
- Eliminación de registros con valores faltantes si la cantidad de datos perdidos es pequeña y no crítica.

Datos originales:

```
Nombre | Edad | Puntuación
Alice  | 35   | 8
Bob    | N/A  | 7
Charlie| 42   | 9
```

Manejo de valores faltantes (usando la media):

```
Nombre | Edad (Rellenado) | Puntuación
Alice  | 35               | 8
Bob    | 38 (media)       | 7
Charlie| 42               | 9
```

Modelos de datos y almacenamiento. Recopilación en bruto de datos y su preprocesamiento.

4. Estandarización y Normalización: En muchos casos, es importante estandarizar o normalizar los datos para que tengan una escala similar. Esto es especialmente relevante en algoritmos basados en distancias o gradientes.

5. Codificación de Categorías: Si los datos contienen variables categóricas (como colores o categorías de productos), **es necesario codificarlas en valores numéricos para que los algoritmos de machine learning puedan utilizarlos.** Esto se hace mediante técnicas como la codificación one-hot.

Modelos de datos y almacenamiento. Recopilación en bruto de datos y su preprocesamiento.

6. Selección de Características: En ocasiones, es necesario seleccionar un subconjunto de las características (columnas) más relevantes para el análisis o el modelo. Esto puede mejorar la eficiencia computacional y reducir la complejidad del modelo.

7. División en Conjuntos de Entrenamiento y Prueba: Finalmente, **los datos se dividen en conjuntos de entrenamiento y prueba para evaluar el rendimiento del modelo. El conjunto de entrenamiento se utiliza para entrenar el modelo, mientras que el conjunto de prueba se utiliza para evaluar su capacidad de generalización**.

Modelos de datos y almacenamiento

“El modelo de datos se puede entender como un lenguaje o conjunto de herramientas que utilizamos para hablar sobre cómo almacenamos y organizamos nuestros datos en una base de datos. **Es como un conjunto de reglas y conceptos que nos permite describir los datos de una manera sistemática y comprensible.**

Modelos de datos y almacenamiento

Imagina que estás construyendo una casa. Para que todos los trabajadores entiendan cómo debe ser la casa, necesitas un plano. Este plano es como un modelo de datos para tu casa.

El plano (modelo de datos) te dice:

- **Cómo son las habitaciones:** Cuántas hay, de qué tamaño son, qué cosas contienen (como habitaciones, baños, cocina). Esto es como la estructura de los datos.
- **Qué cosas no pueden faltar:** Por ejemplo, cada habitación debe tener una puerta. Estas son las restricciones de integridad.
- **Cómo se usan las habitaciones:** Puedes entrar a la cocina, dormir en el dormitorio, etc. Estas son las operaciones.

Entonces, en una base de datos:

- **El modelo de datos es el plano de la base de datos.**
- La estructura de los datos son las "habitaciones" donde guardamos la información.
- Las restricciones de integridad son las "reglas" que aseguran que la información sea correcta y consistente.
- **Las operaciones son las formas de "usar"** la información, como buscar un dato o actualizarlo.

Modelos de datos y almacenamiento

Por ejemplo: Imagina una base de datos de clientes de una tienda. El modelo de datos podría decir:

- Estructura: Cada cliente tiene un nombre, una dirección y un número de teléfono.
- Restricciones: El número de teléfono debe tener 10 dígitos y la dirección no puede estar vacía.
- Operaciones: Puedes buscar a un cliente por su nombre, agregar un nuevo cliente o cambiar la dirección de un cliente.

Modelos de datos y almacenamiento

El modelo de datos se puede entender como un lenguaje o conjunto de herramientas que utilizamos para hablar sobre cómo almacenamos y organizamos nuestros datos en una base de datos. **Es como un conjunto de reglas y conceptos que nos permite describir los datos de una manera sistemática y comprensible.**

Este contiene tres elementos:

- **Notación para la descripción de la estructura de los datos:** Esto significa que tenemos una forma de representar cómo se ven nuestros datos, qué tipo de información contienen y cómo están relacionados entre sí.
- **Restricciones de integridad:** reglas que indican que deben cumplir los datos para ser considerados válidos y precisos.
- **Operaciones para actualizar y recuperar los datos:** el modelo de datos también nos proporciona formas de agregar, modificar o eliminar datos, así como de buscar y obtener información específica de la base de datos.

Modelos de datos y almacenamiento

Existe tres niveles de abstracción a la hora de analizar un modelo de datos:

1. **Modelo de datos conceptual.**
2. **Modelo de datos lógico.**
3. **Modelo de datos físico.**

Modelos de datos y almacenamiento

Existe tres niveles de abstracción a la hora de analizar un modelo de datos:

1. Modelo de datos conceptual. Es el nivel más alto de abstracción y en este nivel se define lo que el sistema contiene, estableciendo su organización, finalidad y reglas y conceptos del negocio.

- En una empresa de ventas, el modelo conceptual podría definir entidades como "Cliente", "Producto" y "Pedido", y las relaciones entre ellas (un cliente puede realizar muchos pedidos, un pedido puede contener varios productos).

2. Modelo de datos lógico. Está situado en el nivel intermedio de abstracción y define cómo el sistema debería estar implementado, independiente del tipo de base de datos que se empleará. El objetivo es desarrollar un mapa técnico para las reglas y la estructura de los datos.

3. Modelo de datos físico. Es el nivel más bajo de abstracción y describe cómo el sistema será implementado usando una base de datos específica.

Modelos de datos y almacenamiento

Existe tres niveles de abstracción a la hora de analizar un modelo de datos:

1. Modelo de datos conceptual. Es el nivel más alto de abstracción y en este nivel se define lo que el sistema contiene, estableciendo su organización, finalidad y reglas y conceptos del negocio.

- En una empresa de ventas, el modelo conceptual podría definir entidades como "Cliente", "Producto" y "Pedido", y las relaciones entre ellas (un cliente puede realizar muchos pedidos, un pedido puede contener varios productos).

2. Modelo de datos lógico. Está situado en el nivel intermedio de abstracción y define cómo el sistema debería estar implementado, independiente del tipo de base de datos que se empleará. El objetivo es desarrollar un mapa técnico para las reglas y la estructura de los datos.

- En el mismo ejemplo, el modelo lógico definiría las claves primarias y foráneas, los tipos de datos de los atributos (numérico, texto, fecha), y las restricciones de integridad (por ejemplo, un pedido debe tener al menos un producto).

Modelos de datos y almacenamiento

Existe tres niveles de abstracción a la hora de analizar un modelo de datos:

-

3. Modelo de datos físico. Es el nivel más bajo de abstracción y describe cómo el sistema será implementado usando una base de datos específica.

- En mismo ejemplo, el modelo físico definiría las tablas, índices, vistas y otros objetos de la base de datos, así como las características de almacenamiento (discos, tamaño de bloques, etc.).

Modelos de datos y almacenamiento

Técnicas de modelización de datos empleadas. Estas técnicas de modelización son aplicables tanto a modelos relacionales como no relacionales, y se describe a continuación:

1. **Diagrama de entidades-relaciones (ERD).** Esta técnica visual es la opción por defecto empleada en la modelización y diseño de bases de datos relacionales. Incorpora el uso de entidades, atributos, relaciones, cardinalidades, restricciones entre otros elementos, así como notación simbólica.

2. **Diagramas de clase Unified Modeling Language (UML).** Este es una notación standard para modelizar y diseñar sistemas de información empresarial.

3. **Diccionario de datos.** Es una representación tabular de los conjuntos de datos y sus atributos, que contienen elementos como la descripción de los elementos, relaciones entre tablas, restricciones (unicidad, valores por defecto, valores válidos).

Modelos de datos y almacenamiento

Diagrama de entidades-relaciones (ERD). Ejemplo

Supongamos que estamos diseñando una base de datos para una biblioteca. Utilizamos un diagrama de Entidades-Relaciones para representar la estructura de datos. Ejemplo simplificado:

- Entidades:
  - Libro
  - Autor
  - Usuario
- Atributos:
  - Libro: Título, ISBN, Fecha de Publicación
  - Autor: Nombre, Apellido, Nacionalidad
  - Usuario: Nombre, Apellido, Número de Tarjeta
- Relaciones:
  - Un libro puede ser escrito por uno o varios autores (relación "Escrito por").
  - Un usuario puede tomar prestado uno o varios libros (relación "Prestado a").

El diagrama mostrará visualmente cómo se relacionan las entidades (Libro, Autor, Cliente) y sus atributos, así como las relaciones entre ellas (quién escribió qué libro, quién ha tomado prestado qué libro).

Modelos de datos y almacenamiento

2. Diagramas de clase Unified Modeling Language (UML). Ejemplo

Diseño para un sistema de gestión de pedidos en línea para una tienda, para ello utilizaremos un diagrama de clase UML para modelar el sistema. Ejemplo simplificado:

- Clases:
  - Pedido
  - Cliente
  - Producto
- Atributos:
  - Pedido: ID, Fecha de Pedido, Total
  - Cliente: ID, Nombre, Dirección
  - Producto: ID, Nombre, Precio
- Relaciones:
  - Un pedido tiene un cliente asociado (relación "Pertenece a").
  - Un pedido puede contener uno o varios productos (relación "Contiene").

Este diagrama UML muestra cómo se relacionan las clases (Pedido, Cliente, Producto) y sus atributos, así como las relaciones entre ellas. También puede incluir métodos y funciones que describen el comportamiento de las clases.

Modelos de datos y almacenamiento

3. Diccionario de datos. Ejemplo

Supongamos que estamos documentando una base de datos para una empresa de gestión de proyectos. Utilizamos un diccionario de datos para describir los elementos de la base de datos. Ejemplo simplificado:

Tablas:

- Proyectos
- Empleados
- Tareas

Atributos (ejemplo de la tabla "Proyectos"):

- ID (Clave Primaria)
- Nombre del Proyecto
- Fecha de Inicio
- Estado del Proyecto

Relaciones (ejemplo de la tabla "Tareas"):

- Una tarea está relacionada con un proyecto (relación "Pertenece a Proyecto").
- Una tarea está asignada a un empleado (relación "Asignada a Empleado").

Este diccionario de datos proporciona una representación tabular de las tablas y sus atributos, incluyendo detalles como las claves primarias y las relaciones entre tablas. También puede incluir descripciones de atributos, restricciones (como valores únicos) y otros metadatos que son útiles para entender y gestionar la base de datos.

Modelos de datos y almacenamiento

Taxonomía de modelos de datos: La taxonomía de modelos de datos es útil para comprender y categorizar los diversos enfoques utilizados para representar y almacenar datos en sistemas de información.

Tipo	Modelo de datos	Técnica
Basada en registros	•Jerárquico: Los datos se almacenan en registros. Los registros tienen estructura de árbol donde cada registro tiene un único *padre* (nodo superior) • Red: Elaborado sobre un modelo jeráquico pero permitiendo que los registros tengan múltiples padres • Registros: Se especifica la estructura de toda la base de datos, definiendo los tipos de registros. Tienen un número determinado de campos con una longitud fijada. • Multidimensional: la estructura de datos contenida en la base de datos está mezclada con los propios datos. Es útil para fuentes de datos basadas en la web y para relacionar bases de datos de diferentes tipos.	ERD, Dic
Relacional	• Relacional: los datos están segmentados con la ayuda de tablas. Se emplea *Structured Query Language* (SQL) como su lenguaje canónico de base de datos. • Modelo de entidades-relación: describe la relación entre cosas de interés en un dominio de conocimiento. Un modelo básico ER está compuesto por tipo de entidades y especifica las relaciones que pueden existir entre estas entidades. • Modelo de datos relacional extendido (ERDM) es un híbrido del modelo relacional añadiendo funcionalidades de modelos orientados a objetos.	ERD, Dic

Modelos de datos y almacenamiento

Taxonomía de modelos de datos: Ayuda a los profesionales de la informática y las bases de datos a identificar cuál es el modelo más adecuado para una tarea o aplicación específica.

	orientados a objetos.	
Basado en objetos	• Modelos orientados a objetos: consiste de objetos que poseen sus características y métodos. Estos modelos se pueden considerar que son post relacionales debido a que no se limitan a tablas aunque las empleen. • Modelo dato por contexto: este modelo incorpora varios modelos de datos según se necesitan, como relacional, orientado a objetos, semi estructurado entre otros. Este modelo permite varios tipos de usuarios que se diferencia en el modo de interactuar con la base de datos	UML ER
NoSQL	• Modelo de grafo. Se emplea una estructura de grafo con nodos, aristas y propiedades para representar los datos almacenados. • Modelo dato multievaludado: Este modelo permite respecto al modelo relacional que cada atributo en lugar de contener un dato unitario almacene una lista de datos. • Modelo de datos documento: este tipo permite almacenar y gestionar documento o dato semi-estructurados mas que datos atómicos.	Diccionar

Tabla 1.4: Taxonomía de modelos de datos

Técnicas y procesos de extracción de la información de los datos.

Etapas de Análisis en la Explotación de la Información y Madurez Analítica en el Negocio

La explotación de la información se refiere al proceso de analizar datos para extraer conocimientos que apoyen la toma de decisiones empresariales. En un entorno competitivo, la capacidad de aprovechar los datos y transformarlos en información útil se ha convertido en un diferenciador clave para las empresas.

Objetivos:

- Identificar patrones y tendencias.
- Optimizar procesos y recursos.
- Anticiparse a los cambios del mercado.
- Apoyar la toma de decisiones en tiempo real.

Etapas de Análisis en la Explotación de la Información

1. **Recolección de Datos:**

- Definir las fuentes de datos (internas y externas).
- Establecer métodos de recopilación (sensores, encuestas, registros, etc.).
- Integrar y estructurar los datos para el análisis.

Técnicas y procesos de extracción de la información de los datos.

Etapas de Análisis en la Explotación de la Información y Madurez Analítica en el Negocio

Etapas de Análisis en la Explotación de la Información

2. **Preparación de los Datos:**

   - Limpieza de datos para eliminar errores e inconsistencias.
   - Transformación y normalización de datos.
   - Almacenamiento en estructuras que faciliten el acceso y análisis (bases de datos relacionales, almacenamiento en la nube, etc.).

3. **Análisis de Datos:**

   - Aplicación de técnicas estadísticas, minería de datos, algoritmos de machine learning y técnicas de visualización.
   - Identificación de patrones, correlaciones y relaciones.

Técnicas y procesos de extracción de la información de los datos.

Etapas de Análisis en la Explotación de la Información y Madurez Analítica en el Negocio

Etapas de Análisis en la Explotación de la Información

4. **Interpretación de Resultados:**
   - Traducción de resultados analíticos en términos de negocio.
   - Evaluación del impacto en la estrategia empresarial.
   - Presentación a las partes interesadas para apoyar la toma de decisiones.

5. **Implementación y Monitoreo:**
   - Incorporación de los resultados en la estrategia empresarial.
   - Definición de métricas clave de rendimiento (KPIs).
   - Monitoreo continuo para ajustar y optimizar los resultados.

Técnicas y procesos de extracción de la información de los datos.

Estadios de Madurez Analítica dentro del Negocio

1. **Analítica Descriptiva (Nivel Básico):**
   - Proporciona una visión retrospectiva: ¿Qué ha sucedido?
   - Responde preguntas como: "¿Cuántas ventas se realizaron el último trimestre?"
   - Se basa en el uso de gráficos, tablas y reportes para mostrar información histórica.

2. **Analítica Diagnóstica (Nivel Intermedio):**
   - Explica las causas de un fenómeno: ¿Por qué sucedió?
   - Utiliza análisis de correlación y patrones para entender las razones detrás de los datos.
   - Responde a preguntas como: "¿Por qué las ventas han disminuido en una región específica?"

3. **Analítica Predictiva (Nivel Avanzado):**
   - Predice eventos futuros: ¿Qué sucederá?
   - Usa modelos estadísticos y algoritmos de machine learning para anticiparse a tendencias.
   - Ejemplo: "¿Cuántas ventas se proyectan para el próximo trimestre?

Técnicas y procesos de extracción de la información de los datos.

Estadios de Madurez Analítica dentro del Negocio

4. **Analítica Prescriptiva (Nivel Experto):**
   - Proporciona recomendaciones: ¿Qué debería hacerse?
   - Integra técnicas de optimización y simulación para proponer la mejor acción posible.
   - Responde a preguntas como: “¿Cuál es la mejor combinación de productos para maximizar las ventas?”

5. **Analítica Cognitiva (Nivel Innovador):**
   - Simula el razonamiento humano para tomar decisiones complejas.
   - Usa inteligencia artificial avanzada y procesamiento de lenguaje natural (NLP).
   - Responde preguntas como: “¿Cómo respondería un experto a una consulta compleja sobre datos?”

Técnicas y procesos de extracción de la información de los datos.

Estadios de Madurez Analítica dentro del Negocio

Tipo de Analítica	Nivel	Descripción	Ejemplo de Pregunta
Analítica Descriptiva	Básico	Proporciona una visión retrospectiva sobre los datos históricos: ¿Qué ha sucedido?	"¿Cuántas ventas se realizaron el último trimestre?"
Analítica Diagnóstica	Intermedio	Explica las causas de un fenómeno: ¿Por qué sucedió?	"¿Por qué las ventas han disminuido en una región específica?"
Analítica Predictiva	Avanzado	Predice eventos futuros: ¿Qué sucederá?	"¿Cuántas ventas se proyectan para el próximo trimestre?"
Analítica Prescriptiva	Experto	Proporciona recomendaciones y la mejor acción a tomar: ¿Qué debería hacerse?	"¿Cuál es la mejor combinación de productos para maximizar las ventas?"
Analítica Cognitiva	Innovador	Simula el razonamiento humano para tomar decisiones complejas utilizando IA avanzada y NLP.	"¿Cómo respondería un experto a una consulta compleja sobre datos?"

Técnicas de aprendizaje automático para la extracción de información de los datos estructurados y no estructurados

Para el análisis de datos se aplican **técnicas de aprendizaje automático o machine learning.**

En el aprendizaje automático, cuando analizamos datos, tenemos dos tipos de información:

Atributos (Características): Estos son los detalles o características específicas de los datos que estamos estudiando.

- Por ejemplo, si estamos analizando información sobre coches, los atributos podrían incluir el color, la marca, el modelo y la velocidad máxima de cada coche.

Variable Respuesta (Etiqueta): Esta es la información que queremos predecir o entender. En nuestro ejemplo de coches, la variable respuesta podría ser si un coche es "seguro" o "peligroso" en función de sus atributos.

Técnicas de aprendizaje automático para la extracción de información de los datos estructurados y no estructurados

Cómo clasificamos los problemas de aprendizaje automático depende de si tenemos o no la variable respuesta (etiqueta) en nuestros datos:

Aprendizaje Supervisado: Cuando tenemos la variable respuesta en nuestros datos, estamos en un problema de aprendizaje supervisado. Significa que podemos entrenar a un modelo de aprendizaje automático utilizando los atributos y las etiquetas conocidas para predecir o entender futuros datos.

Aprendizaje No Supervisado: Si no tenemos la variable respuesta en nuestros datos, estamos en un problema de aprendizaje no supervisado. En este caso, estamos buscando patrones o estructuras ocultas en los datos sin utilizar etiquetas preexistentes.

La diferencia principal es si tenemos o no la información que queremos predecir o entender (la variable respuesta) en nuestros datos. Si la tenemos, es un problema de aprendizaje supervisado; si no, es un problema de aprendizaje no supervisado.

Técnicas de aprendizaje automático para la extracción de información de los datos estructurados y no estructurados

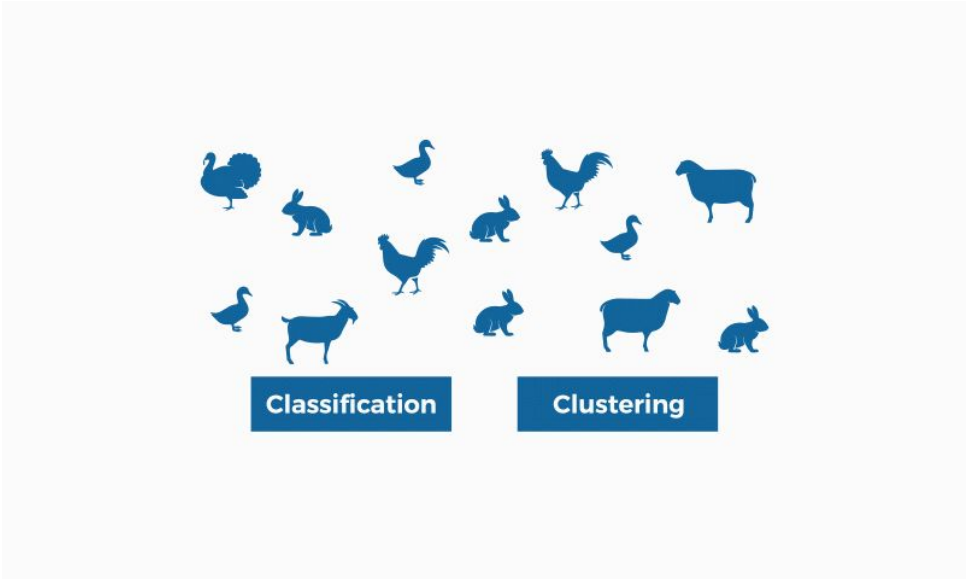
Los tres tipos de problemas que aparecen son:

Clustering. En este tipo de problema, tratamos de identificar grupos o conjuntos de datos que se parecen entre → **patrones**. Esto es, encontrar subconjuntos de observaciones que son similares entre sí. Por ejemplo, si tenemos información sobre personas y queremos agruparlas en diferentes categorías basadas en sus intereses

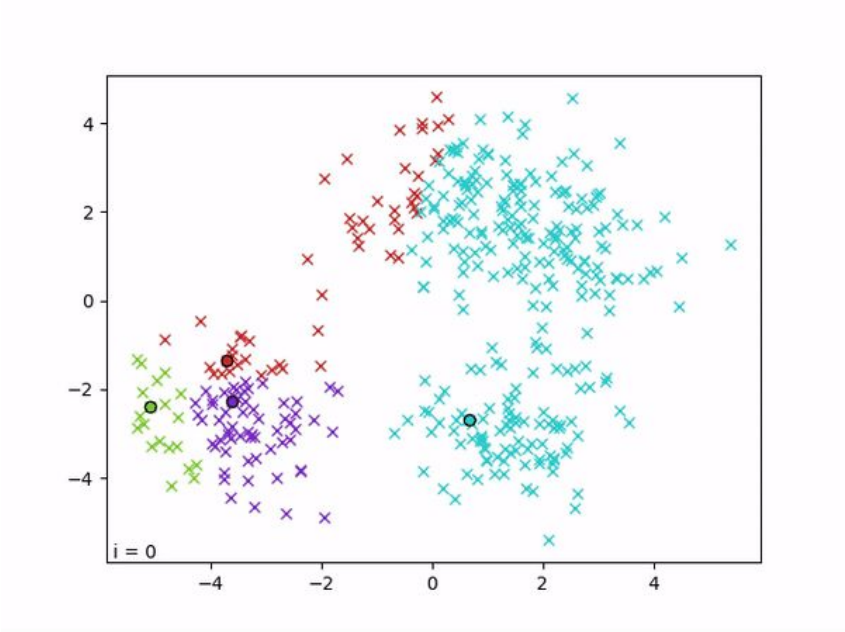
Regresión. Cuando tenemos un problema de regresión, lo que intentamos hacer es predecir un valor numérico o continuo basado en ciertas variables que ya conocemos. Por ejemplo, si tenemos información sobre el tamaño de casas (variable x) y sus precios (variable y), queremos encontrar una fórmula o modelo que nos permita predecir el precio de una casa en función de su tamaño.

Clasificación. En un problema de clasificación, nuestro objetivo es asignar una etiqueta o categoría a un conjunto de datos en función de ciertas características conocidas. Por ejemplo, si tenemos información sobre correos electrónicos y queremos etiquetarlos como "spam" o "no spam" en función de su contenido y otras características

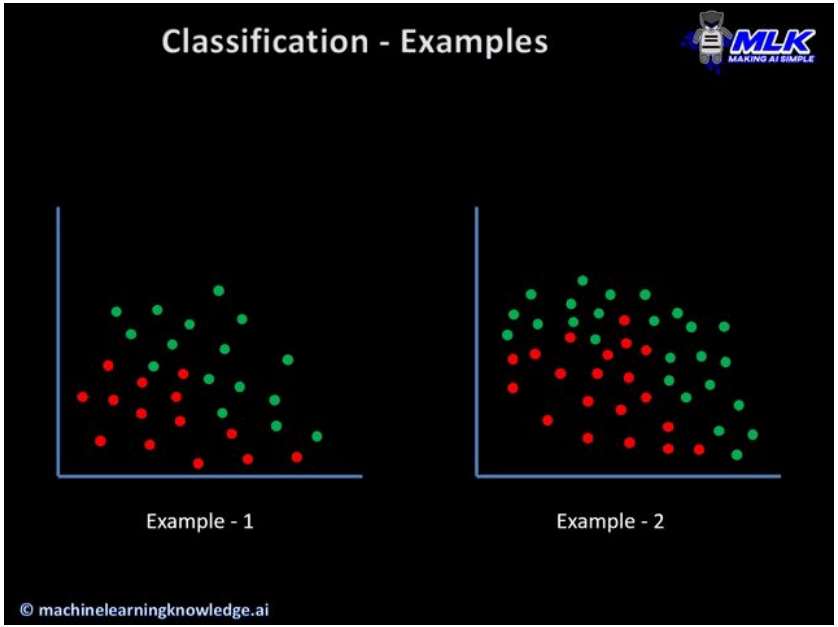
Técnicas de aprendizaje automático para la extracción de información de los datos estructurados y no estructurados



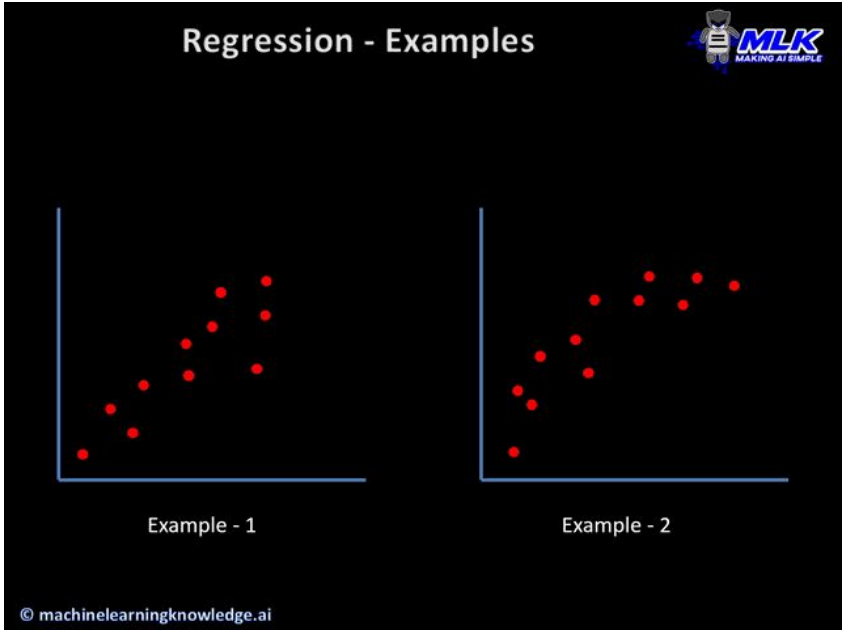
Técnicas de aprendizaje automático para la extracción de información de los datos estructurados y no estructurados



Técnicas de aprendizaje automático para la extracción de información de los datos estructurados y no estructurados



Técnicas de aprendizaje automático para la extracción de información de los datos estructurados y no estructurados



Técnicas de aprendizaje automático para la extracción de información de los datos estructurados y no estructurados

Las técnicas de aprendizaje automático se aplican tanto a datos estructurados como no estructurados. La diferencia esencial es que:

- en los datos estructurados las características ***x*** están completamente definidas a partir de los datos iniciales mientras que
- en los datos no estructurados hay que recurrir a procedimientos para extraer automáticamente estas características.

Técnicas de aprendizaje automático para la extracción de información de los datos estructurados y no estructurados

Figura 1.5: Clasificación de técnicas de extracción de información en datos no estructurados.

Técnicas de aprendizaje automático para la extracción de información de los datos estructurados y no estructurados

El análisis de cada tipo de dato no estructurado es en sí una disciplina distinta apareciendo el denominado procesamiento del lenguaje natural o minería de texto (texto), reconocimiento del habla (audio), reconocimiento de imágenes y procesamiento de vídeos.

Técnicas de aprendizaje automático para la extracción de información de los datos estructurados y no estructurados

- Aplicación en el análisis de texto incluyen:

  1. Análisis de sentimientos. Estas técnicas analizan automáticamente el texto en busca del sentimiento de quien lo escribe (positivo, negativo, neutro, etc.) Estas técnicas permiten a las empresas analizar miles de reseñas en línea o comentarios en las redes sociales sobre ciertos productos en cuestión de minutos.

  2. El reconocimiento de entidades con nombre. Se busca localizar e identificar las entidades con nombre dentro de un texto en categorías predefinidas como organizaciones, lugares, valores monetarios, abreviaturas, etc.

  3. Extracción de eventos. Ejemplos de estas tareas son detectar si los eventos del mundo real han sido reportados en artículos y posts o el seguimiento de acontecimientos similares en diferentes textos.

  4. Extracción de relaciones. Construcción de una base de datos con las interacciones entre fármacos a partir del análisis de texto bruto, o la determinación de relaciones entre personas con el objetivo de construir una base de conocimiento.

Técnicas de aprendizaje automático para la extracción de información de los datos estructurados y no estructurados

- En el análisis de imágenes algunos de los problemas estudiados son:

  1. Extracción de textos y objetos, como la matrícula de un vehículo o determinar la existencia de células cancerígenas en una imagen médica.

  2. Entendimiento de imágenes. Ejemplos en esta categoría son la clasificación semántica de las imágenes.

  3. Análisis de imágenes geoespaciales que van desde la distribución de cultivos a la localización de la pobreza.

  4. Reconocimiento facial, permite identificar la persona de una fotografía.

Técnicas de aprendizaje automático para la extracción de información de los datos estructurados y no estructurados

- Respecto al audio:

  1. Sistemas de reconocimiento de audio que transforma el audio a formato de texto. Sistemas como Google Assistant, Siri, Alexa, Cortana, etc. implementan este tipo de sistemas.

  2. Extracción de características. El habla contiene características prosódicas como el tono, velocidad, calidad, etc y su extracción suministra información sobre el emisor del mensaje. Hay sistemas que extraen estas características de las conversaciones en tiempo real.

Técnicas de aprendizaje automático para la extracción de información de los datos estructurados y no estructurados

- En el análisis de videos destacan:

  1. Creación de resúmenes automáticos en formato de texto o de imágenes en los que se identifican los momentos más destacados, por ejemplo, momentos en los el público aplaude o se pronuncian ciertas palabras clave. Estas técnicas permiten la navegación sobre los vídeos como si se trataran de colecciones de documentos escritos. También permiten la generación automática de subtítulos.

  2. Reconocimiento de lugares, objetos o acciones. Ejemplos de esta categoría son el reconocimiento automático de una infracción de tráfico o el allanamiento de la morada en un sistema de videovigilancia.

Técnicas de aprendizaje automático para la extracción de información de los datos estructurados y no estructurados

Realizar actividad: **Sentiment Analysis: Concept, Analysis and Applications**

https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17

Visualización

En el proceso de obtener información de los datos, la última etapa es la llamada visualización, esta debe ser entendida como una capa entre los resultados obtenidos y el decisor de modo que se facilite la interpretación y evaluación de los resultados.

La visualización es mucho más que una herramienta para comunicar resultados de una forma rápida y objetiva, esta permite además descubrir y comprender los patrones que se encuentran detrás de un conjunto de datos.

La visualización de datos es la parte más del proceso para corroborar hipótesis sobre fenómeno a estudiar, buscando fundamentalmente explicar los datos existentes y realizar predicciones sobre nuevos datos. Con la irrupción de BD, se ha expandido estos propósitos introduciendo aspectos exploratorios y descubrimiento de patrones, apareciendo aspectos como:

- **Resumir** bases de datos masivas para facilitar la toma de decisiones.
- **Identificación interactiva de patrones**
- **Identificación de datos** relevantes para un determinado fenómeno.

Visualización vs Big Data

La aparición de los datos no estructurados en la era de BD ha conducido a que las técnicas de visualización deban abordar las características de las 3 Vs inherentes en su definición:

1. Volumen. La gran cantidad de datos requiere a los métodos tener la posibilidad de la identificación de datos relevantes al fenómeno en cuestión. Un aspecto esencial es determinar el nivel adecuado de agregación que visualice los aspectos esenciales.

2. Variedad. Debido a la existencia de datos no estructurados provenientes de múltiples fuentes se **requiere la integración** de los mismos en una visión de análisis.

3. Velocidad. Exige la recolección y análisis en tiempo real, **exigiéndoles a los métodos la inmediatez pero a la vez generando información útil a la organización.**

Visualización

Los cuadros de mando integrales son visualizaciones que permiten monitorizar los parámetros esenciales de la empresa y disponer de una imagen en tiempo real de lo que ocurre dentro y fuera de la misma.

El método más sencillo para personalizar un cuadro de mando es el uso de asistentes de configuración que utilizan interfaces gráficas de usuario que facilitan la selección de los widgets y de los datos a mostrar.

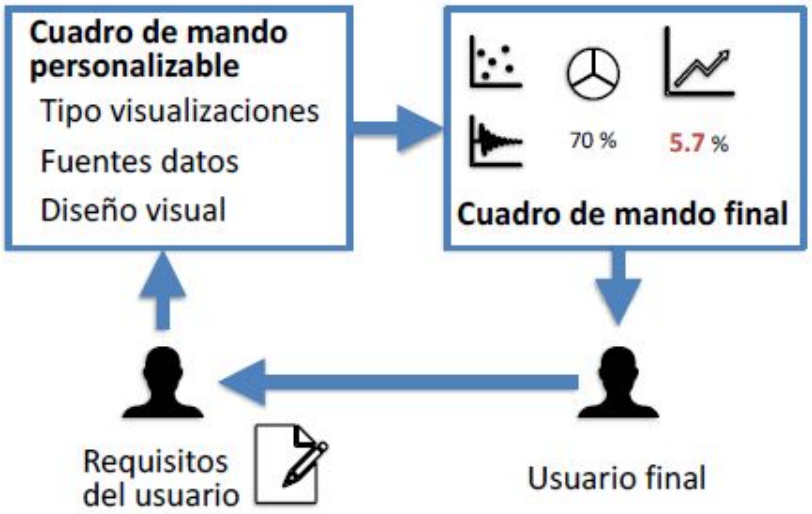


Figura 1.6: Flujo de trabajo para la elaboración de cuadros de mando

Visualización

Las características esenciales para el diseño de un cuadro de mando son:

1. **KPIs:** se debe seleccionar los KPIs esenciales (entre 7 y 10) para el negocio.

2. **Visualización:** deben ser fácilmente interpretables, hablar el mismo lenguaje del decisor y su representación gráfica la adecuada para los datos que representa y visualmente atractivos.

3. **Análisis:** además de las KPIs el cuadro de mando debe acompañar de un análisis sobre lo ocurrido, recomendaciones y su potencial impacto sobre el negocio.

Implicaciones éticas y legales en el uso del Big Data

Protección de Datos Personales:

- Con el uso masivo de datos, es fundamental proteger la privacidad de los usuarios.
- **GDPR (Reglamento General de Protección de Datos):** Es una normativa europea que regula cómo las empresas pueden recopilar, almacenar y utilizar los datos personales de los ciudadanos. Las multas por incumplimiento pueden ser significativas.

Implicaciones éticas y legales en el uso del Big Data

Ética en el uso de Big Data:

- **Problemas potenciales:**
  - **Discriminación algorítmica:** Los algoritmos de Big Data pueden aprender patrones sesgados o discriminatorios, reproduciendo desigualdades. Ejemplo: sistemas de contratación que penalizan a ciertos grupos por sesgos en los datos históricos.
  - **Manipulación de decisiones:** El uso de datos personales para influir en decisiones, como el caso de Cambridge Analytica, donde se utilizó información de Facebook para influir en votantes.

- **Consentimiento informado:** Las empresas deben obtener el consentimiento claro de los usuarios antes de recopilar sus datos y explicar cómo se utilizarán.

Implicaciones éticas y legales en el uso del Big Data

Responsabilidad Corporativa:

- **Transparencia:** Las empresas deben ser transparentes sobre cómo utilizan los datos. Los usuarios deben poder saber qué datos se recopilan y para qué propósito.
- **Seguridad de la información:** La protección de los datos también incluye la implementación de fuertes medidas de seguridad, como cifrado y control de accesos digitales (ciberseguridad) y aquellos mecanismos no digitales necesarios, para evitar brechas y ataques cibernéticos. Aparece **Zero Trust** como medida innovadora que permitirá el alojamiento de datos en la nube de forma segura.

Ejemplos de Incumplimiento y sus Consecuencias:

- Facebook fue multado con 5,000 millones de dólares por el escándalo de Cambridge Analytica.
- Equifax, una de las principales agencias de crédito, sufrió una violación de datos que expuso la información de 147 millones de personas, lo que llevó a multas y pérdida de confianza del público.

Desafíos y oportunidades del Big Data. Desafíos técnicos

Escalabilidad:

- A medida que las empresas recopilan más datos, necesitan infraestructuras que puedan escalar para manejar el volumen. Esto implica costos significativos en almacenamiento y procesamiento.

Velocidad de procesamiento:

- A medida que los datos aumentan, procesarlos en tiempo real se vuelve más complejo.

Seguridad:

- Las grandes bases de datos son un objetivo atractivo para ciberataques. Proteger esta información es esencial.

Desafíos y oportunidades del Big Data. Desafíos técnicos

Variedad de datos:

- Los datos provienen de múltiples fuentes (redes sociales, sensores IoT, transacciones, etc.) y en diferentes formatos (estructurados, no estructurados, semi-estructurados).

Veracidad de los datos:

- No todos los datos recopilados son precisos o útiles. Limpiar y filtrar los datos incorrectos es un desafío importante.

Desafíos y oportunidades del Big Data. Desafíos técnicos

Inteligencia Artificial y Machine Learning:

- Big Data **permite entrenar modelos de machine learning** que pueden generar predicciones precisas basadas en grandes cantidades de datos. Ejemplo: Predicciones de demanda en el sector retail, detección de fraudes en finanzas o diagnóstico de enfermedades en el sector salud.

Análisis Predictivo:

- **Utilizando datos históricos y actuales, las empresas pueden prever tendencias futuras** y tomar decisiones proactivas. Ejemplo: Las aerolíneas utilizan análisis predictivo para ajustar los precios de los billetes en función de la demanda y la disponibilidad.

Nuevas Aplicaciones:

- Big Data **abre nuevas oportunidades en áreas como la personalización del marketing, el análisis del comportamiento del consumidor y el desarrollo de productos basados en preferencias detectadas**. Ejemplo: Netflix utiliza Big Data para analizar los hábitos de visualización de sus usuarios y recomendar contenido personalizado.

Desafíos y oportunidades del Big Data. Desafíos técnicos

Computación en la Nube:

- Cada vez más empresas están moviendo sus sistemas de Big Data a la nube para aprovechar su flexibilidad, escalabilidad y menores costos.

Edge Computing:

- Procesamiento de datos en el mismo lugar donde se generan (sensores IoT, dispositivos móviles) para minimizar la latencia.

Tecnologías de privacidad diferencial:

- Nuevas técnicas que permiten el análisis de datos agregados mientras se garantiza la privacidad individual. Ejemplo: Google y Apple están adoptando estas tecnologías para anonimizar los datos de sus usuarios.

Técnicas de integración, procesamiento y análisis de la información en Big Data. Repaso, contexto. BIU.

El procesamiento de grandes volúmenes de datos en Big Data requiere técnicas específicas de integración, procesamiento y análisis, diferentes a las empleadas tradicionalmente. Se establece un proceso con varias etapas:

1. **Identificación de las fuentes de datos relevantes**, tanto internas como externas.
2. **Preparación y preprocesamiento de los datos crudos para mejorar su calidad**. Incluye limpieza, detección de outliers, manejo de valores faltantes, normalización y codificación.
3. **Definición de un modelo de datos adecuado para organizar y estructurar los datos**. Pueden emplearse modelos relacionales, no relacionales, documentales o gráficos según sea apropiado.
4. **Almacenamiento de los datos procesados en sistemas distribuidos (clúster, nube) para permitir su análisis.**
5. **Análisis de los datos mediante técnicas estadísticas, de aprendizaje automático, reconocimiento de patrones y minería de datos.**
6. **Visualización de resultados en cuadros de mando, reportes y otros formatos.**

Aplicaciones de los sistemas Big Data en empresas y organizaciones. Repaso, contexto. BIU.

Los sistemas Big Data tienen aplicaciones en diversos ámbitos:

- **Comercio electrónico:** Amazon analiza comportamientos de compra para recomendaciones y detección de fraude.
- **Entretenimiento:** Netflix analiza hábitos de visionado para recomendar contenidos y producir nuevas series.
- **Publicidad:** analítica web y redes sociales para segmentación de clientes y campañas de marketing.
- **Finanzas:** detección de fraudes en transacciones financieras mediante análisis en tiempo real.
- **Salud:** análisis de historiales clínicos para la detección temprana de enfermedades.
- **Industria:** mantenimiento predictivo de maquinaria según datos de sensores.
- **Smart cities:** procesamiento de datos de sensores en tiempo real para gestión de tráfico, energía, etc.
- **Investigación científica:** análisis de grandes conjuntos de datos experimentales para nuevos descubrimientos.

Desarrollo de proyectos Big Data: metodología, business case, evaluación de inversiones. Repaso, contexto. BIU.

Para desarrollar un proyecto Big Data se recomienda:

- Elaborar un **business case** que analice los costes, beneficios y riesgos esperados, para decidir su aprobación.
- **Calcular indicadores como VAN (valor actual neto), TIR (tasa interna de retorno) y periodo de recuperación.**
- **Detallar los hitos y duración del proyecto.** Considerar costes directos, de mantenimiento y actualizaciones.
- **Analizar los principales riesgos y sus probabilidades de ocurrencia.**
- **Apoyar el business case** en cifras e información como estudios de mercado, datos financieros e indicadores.
- Revisar y aprobar el documento entre los responsables de la organización.

Tipos de análisis: descriptivo, predictivo, prescriptivo
Contexto. BIU.

- **Análisis descriptivo:** resumen y visualización de datos para conocer el pasado y presente.
  - Ejemplo: Una empresa de comercio electrónico analiza el comportamiento de sus clientes, generando estadísticas de resumen sobre las categorías de productos más vendidas, el promedio de tiempo que los usuarios pasan en el sitio y las tasas de conversión.

- **Análisis predictivo:** encontrar patrones y tendencias para realizar predicciones sobre el futuro. Usa series de tiempo, correlaciones, clustering, clasificación.
  - Ejemplo: Una empresa de seguros utiliza datos históricos de accidentes automovilísticos para predecir las probabilidades de que un conductor tenga un accidente en el futuro y determinar las tarifas de seguros.

- **Análisis prescriptivo:** busca la mejor decisión ante múltiples alternativas. Usa optimización, algoritmos genéticos, heurísticas.
  - Ejemplo: Una aerolínea utiliza análisis prescriptivo para optimizar la programación de vuelos, considerando factores como el clima, la demanda de pasajeros y los costos operativos, lo que resulta en programaciones más eficientes y rentables.

Técnicas y herramientas para análisis en Big Data
Contexto. BIU.

Entre las principales técnicas y herramientas están:

- Plataformas como Hadoop, Spark o bases de datos NoSQL para almacenamiento y procesamiento distribuido.
- Lago de datos (data lake) para integrar y almacenar datos diversos.
- Técnicas de aprendizaje automático como redes neuronales, Random Forests o SVM.
- Algoritmos estadísticos como regresión lineal, clustering, series temporales.
- Procesamiento de lenguaje natural para textos. Reconocimiento de imágenes y voz.
- Visualización de datos con Tableau, Power BI, QlikView.
- Bibliotecas y APIs como TensorFlow, Scikit-Learn, NumPy o Pandas.

El profesional/gestor/científico/analista/técnico de datos y sus competencias

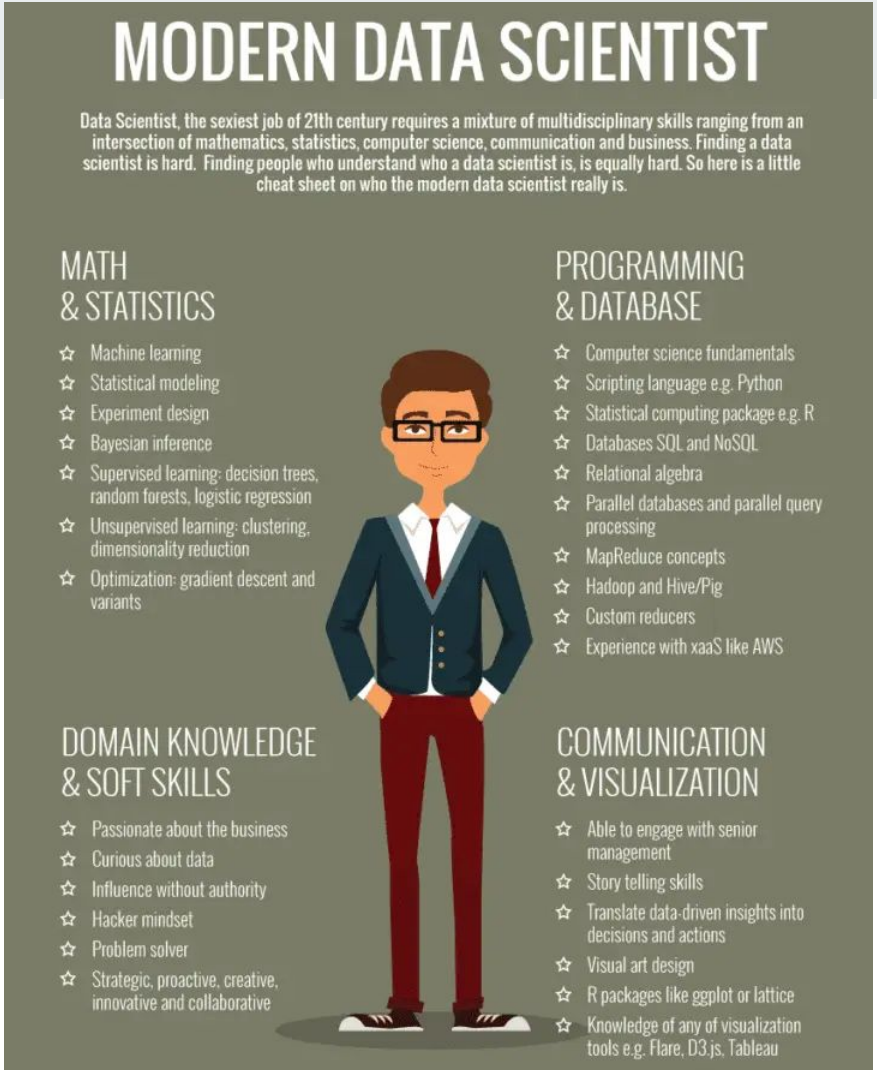
El científico de datos debe tener conocimientos de:

- Inteligencia artificial, aprendizaje automático, estadística.
- Programación y bases de datos.
- Visualización y comunicación de resultados.
- Dominio del problema: finanzas, marketing, medicina, etc.
- Pensamiento analítico y crítico.

Debe ser capaz de:

- Identificar las mejores fuentes de datos.
- Aplicar las técnicas de análisis más adecuadas según el problema.
- Interpretar y evaluar críticamente los resultados obtenidos.
- Comunicar hallazgos de forma comprensible para la toma de decisiones.
- Trabajar en equipos multidisciplinares y gestionar proyectos de análisis de datos.

El científico de datos y sus competencias



El científico de datos y sus competencias

Actividad: Artículo

- **Lee el artículo "How to Get a Job as a Data Scientist?"**
  (https://www.discoverdatascience.org/resources/jobs-in-data-science/).

- **Realiza un resumen del artículo que incluya los siguientes puntos:**
  - Las diferentes áreas de trabajo en ciencia de datos.
  - Las habilidades y conocimientos necesarios para cada área.
  - Saca tus propias conclusiones sobre el futuro de la ciencia de datos.

- **Contesta a las siguiente preguntas:**
  - ¿Qué áreas de trabajo en ciencia de datos crees que serán más demandadas en el futuro?
  - ¿Qué habilidades y conocimientos serán más necesarios para los científicos de datos del futuro?
  - ¿Qué desafíos crees que enfrentará la ciencia de datos en el futuro?