

Tema 1: Big Data

1. Big Data

¿Qué es el Big Data?

Grandes volúmenes de datos que son **difíciles de procesar y analizar** con herramientas tradicionales debido a su tamaño, complejidad y rapidez de generación.

La popularidad de Internet y las redes sociales han hecho que los usuarios de Internet sean a la vez consumidores y productores de contenido. Lo cual genera una gran cantidad de datos.

¿Para qué se usa Big Data?

- **Mejora en la toma de decisiones.** Análisis de grandes volúmenes de datos para identificar patrones, tendencias y hacer predicciones precisas. Ej: Predicción de demanda en empresas de retail.
- **Optimización de procesos empresariales.** Mejorar la eficiencia operativa y reducir costos. Ejemplo: Optimización de la cadena de suministro o logística.
- **Personalización de experiencias.** Para personalizar productos y servicios según las preferencias de los clientes. Ej: Recomendaciones de productos en plataformas como Amazon o Netflix.
- **Investigación y desarrollo.** Impulsa avances en investigación científica y tecnológica al procesar grandes cantidades de datos experimentales. Ej: Descubrimiento de nuevos medicamentos o terapias genéticas

Estadística

Es la ciencia que se encarga **de recolectar, analizar, interpretar y presentar datos**. Utiliza métodos matemáticos para inferir propiedades sobre una población a partir de una muestra, con el objetivo de entender patrones y tendencias.

- **Estadística descriptiva:** Nos ayuda a describir nuestros datos. Se apoya en el Análisis Exploratorio de Datos
- **Estadística inferencial:** Hacer predicciones sobre una población basándose en una muestra.

IA vs ML vs Estadística

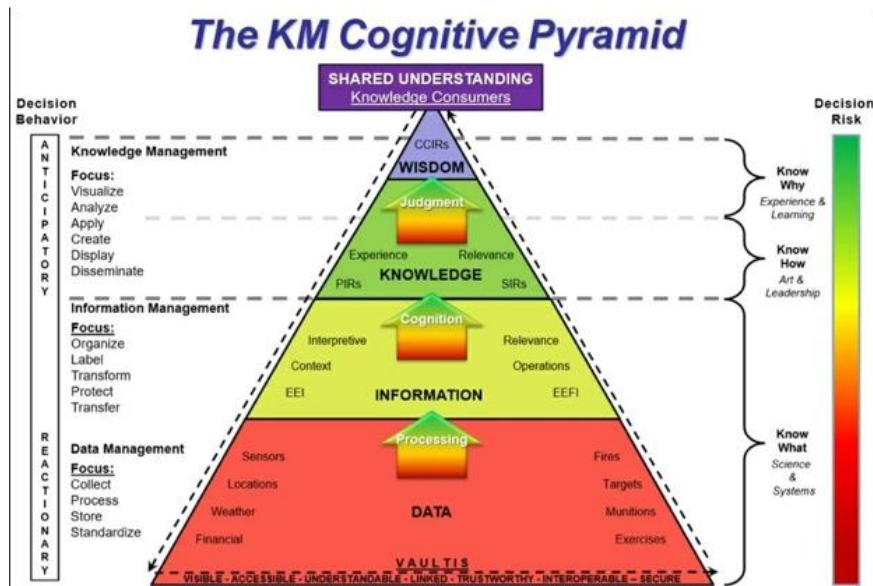
- **Machine Learning:** utiliza técnicas estadísticas para extraer patrones de datos y mejorar los modelos con más experiencia (datos).
- **Big Data:** se utiliza como una fuente de datos masivos para aplicaciones tanto de estadística como de Machine Learning.
- **Inteligencia Artificial:** es el campo más amplio que abarca muchas tecnologías, incluyendo Machine Learning y Big Data, que le permiten mejorar y automatizar

2. Del Dato a la Información

¿Cómo funciona la pirámide DIKW?

Representa un proceso ascendente, donde **los datos se transforman en información, luego en conocimiento y finalmente en sabiduría**. Cada nivel construye sobre el anterior, y la sabiduría es el resultado de la integración de todos los niveles

- **Datos:** Cada registro de venta es un dato individual (fecha, producto, cantidad, precio).
- **Información:** Al analizar los datos, puedes generar informes que muestren las ventas totales por mes, los productos más vendidos o los clientes más recurrentes.
- **Conocimiento:** A partir de estos informes, puedes identificar tendencias en las ventas, predecir la demanda futura y ajustar tus estrategias de marketing.
- **Sabiduría:** Utilizando este conocimiento, puedes desarrollar nuevos productos, mejorar la experiencia del cliente y tomar decisiones estratégicas para el crecimiento de tu negocio.



A medida que ascendemos en la pirámide nos enfrentamos con conceptos menos programables y susceptibles de ser manipulados mediante algoritmos.

2.1 Caracterización del Dato

Dato en cuanto al Tipo

- **Tipos simples:** Representan un **único valor**. Estos se dividen en lógicos (booleanos), caracteres y numéricos (enteros o flotantes), cada uno de ellos representado por un número de bits.
- **Tipos complejos:** Son el resultado de la **combinación de los tipos simples** y aparecen tipos compuestos, que representan un conjunto de valores a modo de estructura. Con tipos compuestos podemos representar la imagen de la matrícula de un coche, en forma de matriz de bits, datos multimedia como imágenes y vídeos, información geoespacial o imágenes de calor etc.

Dato en cuanto al Formato

- **Datos estructurados:** Los datos se presentan en un **modelo o esquema organizativo**. Ejemplo: una base de datos SQL, donde las colecciones se hacen presentes mediante tablas y las relaciones como referencias. Por ejemplo: un cliente en una tabla puede estar almacenado como tipo compuesto en una tabla, constituyendo un registro y estará compuesto por un conjunto de datos simples.
- **Datos semiestructurados:** Estos datos no son estructurados, pero que **presentan cierta organización**. Tienen etiquetas o identificadores que indican su significado. Por ejemplo: MIME (Multipurpose Internet Mail Extension), XML y JSON.
- **Datos no estructurados:** Se caracterizan por su **formato variable** y la **dificultad para almacenarlos** en bases de datos relacionales tradicionales. Ejemplo: Mensajes de Blog, Imágenes de redes sociales, Datos geoespaciales, Lecturas de sensores de una ciudad, Etc...

Dato en cuanto al Generador

- **Datos generados por personas:** Son los que más rápidamente están creciendo por el volumen de interacciones en las redes sociales y el comercio electrónico. Ejemplos: Información sobre nuestro día a día online, Evaluaciones online de productos, Blog y vídeos subidos por usuarios, Publicaciones en redes sociales, Texto en la red social X, Información recopilada de la apps móviles...
- **Datos generados por máquinas:** Información creada, capturada y transmitida por dispositivos, sensores y sistemas automatizados, sin intervención humana directa. Ejemplos: Lectura de sensores, Registros de transacciones comerciales online, Registros de accesos web, Datos generados por dispositivos IoT, Etc...

Dato en cuanto al Tamaño

El **Impacto del Tamaño** en la Tecnología:

- **Sistemas de almacenamiento:** Los sistemas deben ser capaces de manejar volúmenes masivos, desde GB hasta PB.
- **Capacidad de procesamiento:** Se necesita hardware potente para acelerar el acceso a datos.
- **Bases de datos NoSQL:** Las BBDD tradicionales no pueden escalar horizontalmente ni manejar datos no estructurados o de grandes volúmenes.

El **Impacto del Tamaño** en el ANÁLISIS de DATOS:

- **Complejidad del análisis:** Los datos son más diversos.
- **Necesidad de escalabilidad:** Requiere infraestructuras escalables tanto horizontal (añadir más maquinas) como verticalmente (mejorar los recursos de la máquina).
- **Importancia de la visualización:** Representar y entender los datos es abrumador sin técnicas de visualización.

Dato en cuanto al Rol

- **Datos maestros:** se refieren a un conjunto de información consistente y unificada sobre **entidades clave** de una organización. Estos datos son esenciales para las operaciones comerciales y se comparten entre diferentes áreas y sistemas.

Características de los datos maestros:

- Consistencia.
- Uniformidad.
- Centralización.
- Actualización.

Beneficios de los datos maestros:

- Mejora de la calidad de los datos.
- Facilitación de la integración de datos.
- Obtención de una visión unificada del negocio.
- Toma de decisiones más informada.

Ejemplos: Clientes, Productos, Proveedores, Empleados, Ubicaciones

- **Datos operacionales:** En el contexto empresarial, se refieren a la información generada por las **operaciones diarias del negocio**.
Ejemplos: Datos de transacciones de clientes, Datos de inventario, Datos de navegación web.
- **Datos externos:** datos generados y recopilados **fuera del propio negocio**, pero que tienen relación él, siendo susceptibles de **influir y aportar valor**. Fuentes:
 - Redes sociales
 - Blogs.
 - Sitios webs.
 - Dispositivos móviles.
 - Información financiera.
- **Datos analíticos:** El dato analítico se genera a partir de los **datos operacionales dentro del contexto de los datos maestros**, denominados **dimensiones**. Se utilizan para generar información procesada que puede apoyar decisiones estratégicas.

Dato en cuanto a la Latencia

Se refiere al tiempo que transcurre **entre la solicitud de un dato y su disponibilidad para su uso**. Es un factor crítico en muchas aplicaciones, especialmente en aquellas que requieren respuestas en tiempo real o procesamiento de alta frecuencia

- **Datos en tiempo real:** son aquellos que se procesan y **analizan a medida que se generan**, permitiendo tomar **decisiones y acciones inmediatas**. El procesamiento en tiempo real es una característica inherente a los **datos operacionales**. Ejemplos: Monitorización de seguridad en redes, Comercio electrónico, Redes sociales, Sistemas de transporte, ETC...

Beneficios de los datos en tiempo real:

- Toma de decisiones más rápida e informada.
 - Mejor experiencia del cliente.
 - Mayor eficiencia operativa.
 - Innovación.
- **Datos en lotes:** Los datos son almacenados en un lote (batch) cuando son recibidos, permaneciendo así durante un periodo de tiempo o hasta que alcancen un volumen determinado. Por ejemplo, una empresa puede procesar las transacciones de ventas del día al final del día, o los datos de sensores cada hora.

Ventajas:

- Eficiencia para grandes volúmenes de datos.
- Simplificación de la infraestructura.

Desventajas:

- Mayor latencia.
- Menor flexibilidad.

Dato en cuanto a la Sensibilidad

Son los datos clasificados en términos de **privacidad e intimidad**. Esto marca quien accede a estos datos y durante cuánto tiempo, estableciéndose también las condiciones en que deben ser almacenados.

Las empresas suelen categorizar estos datos en cuanto a su sensibilidad y acceso, pero además estas tienen en cuenta tanto **datos recogidos de las interacciones con los clientes y proveedores, como aquellos generados por el propio negocio**.

Varían de un país a otro en función de las **regulaciones** que se apliquen:

- GDPR.
- SOC2.
- HIPAA.
- PCI SSC.

Hay tres **clases de datos** en términos del riesgo que supone su manipulación indebida:

- Riesgo alto.
- Riesgo medio.
- Riesgo bajo.

Riesgo bajo	Riesgo medio	Riesgo alto
<ul style="list-style-type: none">• Ofertas de empleo.• Nota de prensa.• Material de marketing aprobado para uso público	<ul style="list-style-type: none">• Identificadores internos de empleados• Datos de desarrollo, investigación y patentes no publicadas.• Contenidos con propiedad intelectual licenciada de tercero o restringida por contrato.• Datos de recursos humanos relacionados con los empleados.	<ul style="list-style-type: none">• Identificadores personales (números de la Seguridad Social, DNI, pasaporte)• Contraseñas y claves de acceso a sistemas, aplicaciones, etc.• Información identificable sobre salud o pólizas.• Números de tarjetas de crédito o débito.

3. Roles

Chief Data Officer (CDO)

Es responsable de **maximizar el valor estratégico de los datos** en una organización. A través de la gobernanza y explotación eficiente de los datos, el CDO **impulsa la toma de decisiones basada en datos**.

Además, se encarga de la calidad de los datos, la seguridad y alineación con la estrategia de negocio.

Gobierno del dato

Es el **conjunto de políticas, procedimientos y estándares** que aseguran la **calidad, seguridad y uso adecuado de los datos** en una organización. Su objetivo es garantizar que los datos sean **consistentes, confiables y accesibles** para su uso en la toma de decisiones. Sus roles son:

- **Data Owner:** es la persona o equipo responsable de los datos en su ciclo de vida. Establece reglas sobre **cómo deben gestionarse y es responsable de su calidad y seguridad**. Tiene una visión estratégica de los datos dentro de su dominio.
- **Data Steward:** es el encargado operativo de **garantizar que las políticas de gobierno de datos se cumplan**. Se asegura de que los datos sean precisos, estén disponibles y cumplan con los estándares definidos por los Data Owners y el equipo de gobernanza. (Que las bases de datos no tengan direcciones duplicadas o datos inválidos).

Arquitectura de datos

Para diseñar, implementar y mantener la infraestructura que soporta el manejo de datos. Sus roles son:

- **Data Architect:** Diseña la estructura general de los sistemas de datos. Define cómo se almacenarán, accederán y procesarán los datos en la organización. Su enfoque es tanto estratégico como técnico.

Responsabilidades: Diseñar modelos de datos, seleccionar tecnologías y establecer políticas de almacenamiento y acceso.

- **Solutions Architect:** diseña soluciones técnicas que cumplan con los requisitos de negocio. Colabora con el Data Architect para asegurarse de que la infraestructura de datos soporte las aplicaciones empresariales

Responsabilidades: Integrar tecnologías y herramientas adecuadas para cumplir con las demandas de las aplicaciones de negocio.

- **Cloud Architect:** Diseña y gestiona la arquitectura de datos en entornos en la nube. Selecciona las plataformas de nube (AWS, Azure, GCP) y define cómo los datos serán almacenados, protegidos y procesados en estos entornos.

Responsabilidades: Configurar infraestructuras en la nube, garantizar la escalabilidad y la seguridad de los datos

Data Engineer

El Ingeniero de Datos (Data Engineer) es responsable de **construir y mantener la infraestructura que permite el flujo, almacenamiento y procesamiento eficiente de datos dentro de una organización**. Su trabajo principal es crear **pipelines de datos**. Sus responsabilidades son:

- **Diseño y creación de pipelines de datos:** Implementa flujos de trabajo que capturan, procesan y transforman datos desde diversas fuentes hacia un destino, como un Data Warehouse o Data Lake.
- **Optimización del rendimiento:** Se asegura de que los pipelines y las infraestructuras de datos funcionen de manera eficiente, minimizando tiempos de procesamiento y maximizando el rendimiento

- **Garantizar la calidad de los datos:** Implementa procesos para limpiar, validar y transformar los datos para asegurar su calidad antes de que lleguen a los analistas o científicos de datos.
- **Gestión de herramientas y tecnologías:** Selecciona y trabaja con tecnologías como Hadoop, Spark, Kafka, bases de datos NoSQL, entre otras, para el procesamiento masivo de datos

Data Scientist

Un Data Scientist es un profesional que **analiza grandes volúmenes de datos** para obtener insights valiosos que ayuden en la toma de decisiones estratégicas. Combina conocimientos en estadística, matemáticas y programación para crear modelos predictivos, analizar patrones y resolver problemas de negocio mediante datos.

Otros roles

- **Machine Learning Engineer:** Especializado en diseñar, construir y desplegar modelos de machine learning en producción.
- **Data Governance Manager:** Responsable de implementar y supervisar las políticas de gobierno de datos.
- **Data Visualization Specialist:** Enfocado en crear visualizaciones efectivas para comunicar insights de datos.
- **Data Analyst:** Fundamental para transformar datos en insights accionables. Es el rol más directamente relacionado con el análisis y la generación de informes.
- **Data Translator:** Crucial en organizaciones donde hay una brecha significativa entre equipos técnicos y de negocio. Facilita la alineación y la comunicación efectiva.
- **Business Owner:** Vital para asegurar que las iniciativas de datos estén alineadas con las metas estratégicas del negocio y para proporcionar dirección y recursos.

4. Etapas de análisis en la exploración de la información

Analítica descriptiva

Proporciona **información sobre el rendimiento pasado del negocio y su contexto**. Este tipo de análisis se vale de informes, de cuadros de mando (dashboards) que permiten al usuario **consultar y navegar por la información en modo autoservicio**. Responde a preguntas como:

- Cuál fue el número de piezas defectuosas en cada una de las fábricas durante el último trimestre.
- Cómo ha variado la rentabilidad promedio por metro cuadrado de las tiendas respecto al último año.
- Qué relación existe entre los días de lluvia y el incremento en la venta de paraguas.

Analítica prescriptiva

Tiene un enfoque más operativo y de proceso, ya que **busca detallar la mejor solución para una situación determinada**. Por ejemplo:

- Organizar los turnos y las rotaciones de las tripulaciones en una compañía aérea teniendo en cuenta restricciones operativas y laborales.
- Establecer las ubicaciones más adecuadas para situar una serie de centros logísticos con el fin de abastecer los puntos de venta los más rápido posible, incurriendo en los mínimos costes.
- Definir la estrategia más adecuada para el petróleo, considerando los niveles de producción en cada momento, la demanda y la situación geopolítica.

Analítica predictiva

Se basa en el **descubrimiento de patrones, tendencias y relaciones** que permiten explicar un comportamiento a partir de datos históricos con el fin de anticiparse a él en el futuro.

Analítica cognitiva

El objetivo es desarrollar sistemas con capacidad para entender, razonar e interactuar emulando a los seres humanos. Por ejemplo:

- Revelando patrones y relaciones entre los datos que son difíciles de detectar, sugiriendo nuevos cruces de información.
- Recomendando las mejores formas de representar y visualizar los datos, interpretando su significado y contexto.
- Respondiendo a preguntas en lenguaje natural, de forma clara y concisa, acelerando la navegación sobre los datos

5. Almacenamiento

Bases de datos

En Big Data se tiende a replicar la información, ya que **importa más el acceso al dato más que su almacenaje**, ya que esto último suele ser más barato. En la práctica las ETL's, se hacen con JOIN de distintas tablas quedando una tabla mucho más grande. Tipos:

- **NoSQL:** ofrecen modelos de esquema **flexibles** y **no dependen de tablas estructuradas ni de consultas SQL**. Están diseñadas para gestionar datos distribuidos a gran escala con distintos formatos.
 - **Document store:** Almacena documentos formato JSON
 - **Clave valor:** Los datos se guardan como pares clave valor
 - **Columnar:** Almacena el contenido en columnas en lugar de filas, diseñado para una eficiencia alta de lectura/escritura
 - **Grafos:** Se centran en las relaciones entre puntos de datos utilizando estructuras gráficas de nodos y aristas.

- **Series Temporales:** Para manejar datos de series temporales en las que se recopilan puntos de datos a lo largo del tiempo. Estas bases de datos están optimizadas para **leer y escribir secuencias de datos**, generalmente **con una marca de tiempo**. Usado en
 - IoT
 - Datos financieros
 - Monitorización de sistemas
- **De vectores:** Es una base de datos que puede almacenar vectores (listas de números de longitud fija) junto con otros elementos de datos. Generalmente implementan el algoritmo ANN o kNN para que uno pueda buscar en la base de datos con un vector de consulta para recuperar los registros de base de datos coincidentes más cercanos.