

Unidad 1

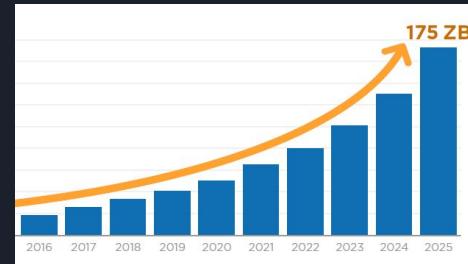
¿QUÉ ES LA MINERÍA DE DATOS?

Modelizado de Minería de Datos - 2Q2025

Nuevas necesidades

El crecimiento exponencial de los datos

En las últimas décadas, hemos experimentado un crecimiento exponencial en la generación y recopilación de datos, impulsado por la digitalización de nuestras vidas, el internet de las cosas (IoT) y las redes sociales.

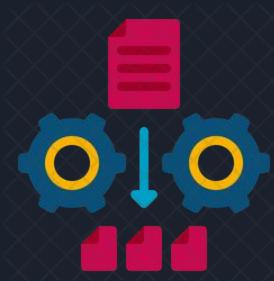


IDC says 175 ZB will be created by 2025 (Image courtesy IDC)
1ZB = 1.000 millones de TB

Extraer conocimiento

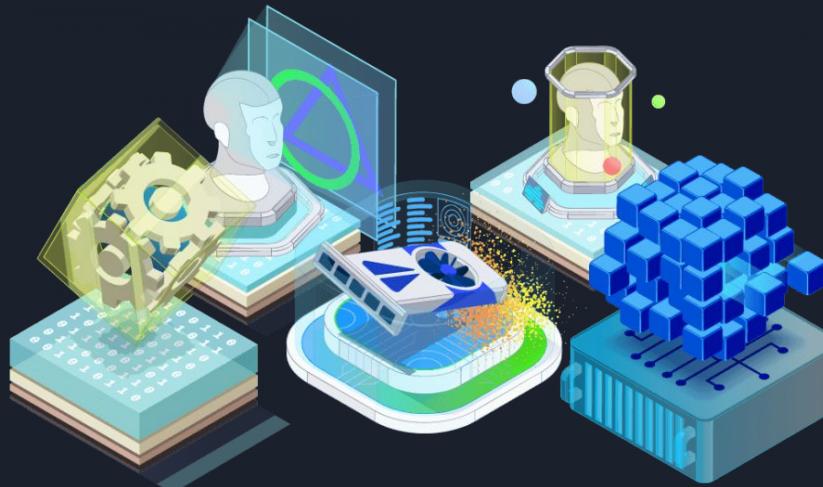
El análisis manual de datos es lento, costoso y subjetivo, especialmente con grandes volúmenes de datos. Esto puede llevar a tomar decisiones basadas en la intuición en lugar de en datos.

La gran cantidad de datos disponibles no garantiza automáticamente la obtención de información útil. Es necesario aplicar técnicas para identificar patrones, tendencias y relaciones ocultas en los datos.



¿Qué es la Minería de Datos?

Descubrir patrones ocultos a partir de la aplicación técnicas sobre grandes conjuntos de datos



→ EXTRACCIÓN DE INFORMACIÓN ÚTIL
(conocimiento)



El análisis tradicional

El análisis tradicional (SQL, OLTP) es rígido y poco escalable para grandes volúmenes de datos.

Las herramientas de inteligencia de negocios (OLAP, Data Warehouses) son eficientes para reportes, pero no descubren conocimiento nuevo.

La minería de datos supera estas limitaciones, ofreciendo conocimiento predictivo y patrones accionables que el análisis tradicional no puede encontrar.





El concepto de minería de datos

Extraer **conocimiento** valioso y comprensible que antes estaba oculto en grandes volúmenes de datos.

- **Datos:** La Materia Prima. Son la base cruda y sin procesar de la información.
- **Información:** El Contexto. Se obtiene al organizar, procesar y dar contexto a los datos.
- **Conocimiento:** El Descubrimiento. Se adquiere al analizar la información para encontrar patrones, tendencias y relaciones. Es el "porqué" o el "qué pasará después".

Objetivo → **TOMA DE DECISIONES**

El Viaje de los Datos: De la Adquisición al Valor de Negocio



El éxito en la ciencia de datos y la inteligencia artificial depende de una base sólida en la adquisición, ingeniería y gestión de datos. No se puede construir la cima sin una pirámide estable.

Tipos de datos

Datos estructurados

- Se organizan en un formato definido
- Son fáciles de procesar y analizar debido a su estructura predefinida.
- Ejemplo: registros de clientes en una base de datos de ventas.

Datos no estructurados

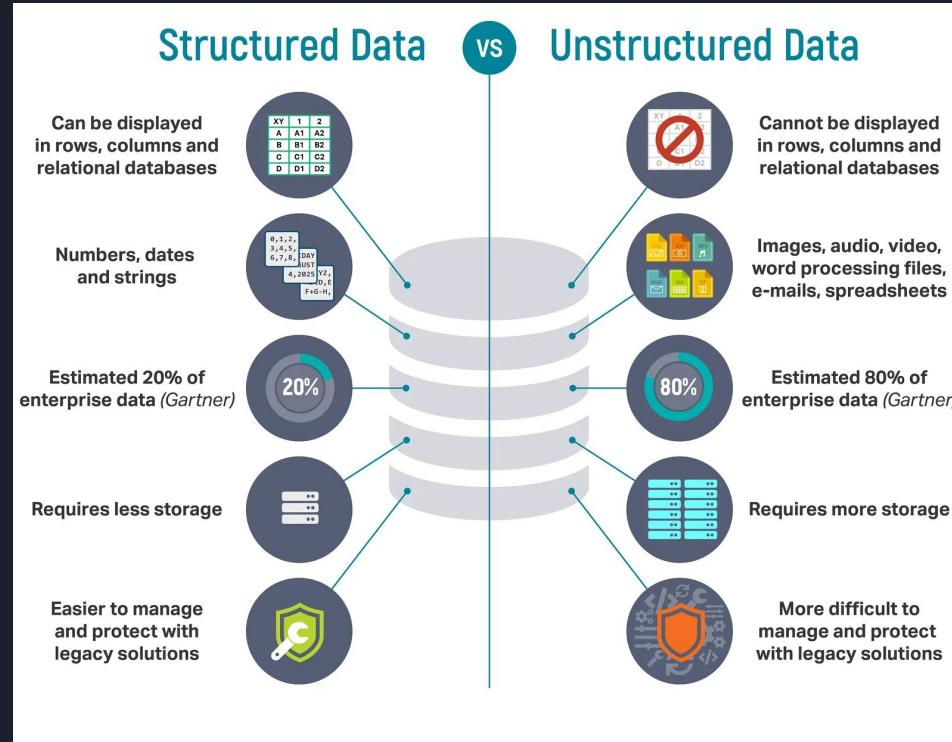
- No tienen un formato predefinido y son difíciles de organizar y analizar.
- Incluyen texto, imágenes, audio y video.
- Ejemplo: publicaciones en redes sociales, correos electrónicos, videos de YouTube.

Transaction table		
Transaction ID	Customer ID	Product ID
53666	24221	389
50333	24222	789
54673	24223	879
58930	24224	975

Product table		
Product ID	Product name	Price per kg
389	Banana	4
789	Apple	5
879	Watermelon	5
975	Mango	7

Customer table		
Customer ID	Last name	First name
24221	Smith	James
24222	Jones	Sam
24223	Taylor	Ann
24224	Burton	Sue

Tipos de datos





Tipos de modelos

Modelos predictivos:

- Función Principal: Pronosticar. Responden a preguntas como: "¿Qué va a pasar?" o "¿En qué categoría caerá este nuevo caso?".
- Ejemplo: un modelo de regresión que predice las ventas futuras de un producto.

Modelos descriptivos:

- Función Principal: Comprender. Responden a preguntas como: "¿Qué ha pasado?" o "¿Cómo se relacionan estas cosas?".
- Ejemplo: un modelo de agrupamiento que identifica grupos de clientes con comportamientos similares.



El proceso de descubrimiento de conocimiento en bases de datos

Un término muy utilizado, y el más relacionado con la minería de datos, es la extracción o “descubrimiento de conocimiento en bases de datos” (*Knowledge Discovery in Databases, KDD*)

Propiedades deseables del conocimiento extraído:

- Válidos: precisos para datos nuevos.
- Novedosos: desconocidos para el sistema y el usuario.
- Potencialmente útiles: conducen a beneficios.
- Comprensibles: fáciles de interpretar y usar.

Proceso habitual de recopilación, almacenamiento y análisis

-  Integración de datos: comprender el negocio, determinar fuentes útiles y obtenerlas.
-  Transformación a formato común: datawarehouse para unificar y resolver inconsistencias.
-  Selección, limpieza y transformación: correcciones, tratar nulos y atributos relevantes.
-  Minería de datos: elegir tarea (clasificación, agrupamiento, etc.) y método.
-  Evaluación e interpretación: evaluar e iteración si es necesario.
-  Difusión: usar y compartir el nuevo conocimiento.



Relación con otras disciplinas



- **Bases de datos:** almacenes de datos, OLAP, indexación y acceso eficiente a datos
- **Recuperación de información (IR):** obtención de información de textos, búsqueda por palabras clave, medidas de similitud
- **Estadística:** conceptos, algoritmos y técnicas (media, varianza, regresión, etc.)
- **Aprendizaje automático:** algoritmos de aprendizaje, modelos a partir de ejemplos
- **Sistemas de toma de decisión:** herramientas para decisiones efectivas (análisis ROC, árboles de decisión)
- **Computación paralela y distribuida:** procesamiento eficiente de grandes volúmenes de datos
- **Otras disciplinas:** lenguaje natural, análisis de imágenes, procesamiento de señales, etc.
- **Visualización de datos:** técnicas para descubrir patrones visualmente (gráficas, diagramas, etc.)



Aplicaciones

Marketing

- Segmentación de clientes para campañas personalizadas.
- Análisis de la cesta de la compra para recomendaciones de productos.
- Detección de fraude en transacciones online.

Salud

- Diagnóstico temprano de enfermedades a partir de datos clínicos.
- Descubrimiento de nuevos fármacos y tratamientos.
- Análisis de datos genómicos para la medicina personalizada.



Aplicaciones

Finanzas

- Detección de fraude en tarjetas de crédito.
- Predicción del riesgo crediticio.
- Análisis de tendencias del mercado bursátil.

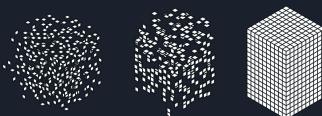
Redes sociales

- Análisis de sentimiento de publicaciones.
- Detección de tendencias y temas populares.
- Identificación de comunidades y grupos de interés.

El papel de la minería de datos en la era del Big Data

La minería de datos es una disciplina clave en la era del Big Data, ya que proporciona las herramientas y técnicas necesarias para extraer conocimiento valioso de los volúmenes masivos de datos que se generan hoy en día. Sin la minería de datos, el Big Data sería solo un océano de información sin sentido.

Ejemplos:



- Escalabilidad: Permite analizar conjuntos de datos de diversos tamaños. 
- Descubrimiento de patrones: No se limita a verificar hipótesis, sino que encuentra relaciones y tendencias ocultas. 
- Automatización: Utiliza algoritmos inteligentes para automatizar el proceso de búsqueda de conocimiento. 
- Toma de decisiones: Transforma los datos en información estratégica. 

Sistemas y herramientas

The image is a composite graphic. On the left, large white and cyan text reads "FREE DATA MINING TOOLS". Below it, the website "www.oragetechnologies.com" is listed. On the right, a central white circle contains icons representing data mining concepts: a magnifying glass over a bar chart, a stack of coins, a gear, and a cloud. Surrounding this center are ten colored hexagonal icons, each containing a logo and name for a different tool: SAS (red), WEKA (orange), rapidminer (yellow), KNIME (green), Rattle (teal), Qlik Sense (blue), orange (purple), TERADATA (dark purple), and Qlik Sense (light purple).

FREE
DATA
MINING
TOOLS

www.oragetechnologies.com

sas

WEKA

rapidminer

KNIME

Rattle

Qlik Sense

orange

TERADATA



Cultura Data-Driven

Los datos son el activo más valioso de cualquier empresa. La capacidad de recopilar, analizar y utilizar datos de manera efectiva puede marcar la diferencia entre el éxito y el fracaso.

¿Qué es una Cultura Data-Driven?

Una cultura data-driven se define por una mentalidad y un enfoque que **prioriza la toma de decisiones basadas en el análisis y la interpretación de los datos** almacenados a partir de fuentes digitales. En lugar de depender únicamente de la intuición o la experiencia, las organizaciones data-driven confían en datos verificables para guiar sus acciones y estrategias.



Cultura Data-Driven

Importancia de una cultura Data-Driven

- Decisiones más informadas
- Identificación de oportunidades
- Optimización de procesos
- Personalización y segmentación
- Competitividad



Cultura Data-Driven

Fomentar una cultura Data-Driven

- Liderazgo comprometido
- Inversión en tecnología y capacitación
- Fomento de la colaboración interdepartamental
- Énfasis en la ética y la seguridad de los datos
- Iteración continua





Unidad 1 - Resumen

En resumen, la minería de datos es una disciplina multidisciplinaria que se basa en la estadística, el aprendizaje automático, la inteligencia artificial y las bases de datos para extraer conocimiento útil a partir de grandes conjuntos de datos. La visualización de datos juega un papel crucial en la exploración, comunicación y toma de decisiones.

