Big data

¿La hoja de cálculo con las notas de tu clase es big data? Claro que no. ¿Y cada validación de cada tarjeta de transporte en cada autobús de Mallorca durante todo un año? ¿O cada búsqueda de vuelo y hotel que recibe una web de viajes de Palma en agosto? En algún punto entre esos ejemplos hay una línea. ¿Dónde está?

El big data es una de las tecnologías habilitadoras que conociste en la Unidad 2, y en la Unidad 4 viste que la IA aprende de los datos. Esta lección explica qué hace que unos datos sean «big», por qué «muchos datos» no es la respuesta correcta y cómo encajan el big data, el análisis de datos, el machine learning, el deep learning y la IA, para que dejes de usar los cinco términos como si fueran sinónimos.

Qué aprenderás

- Describir las cinco V que definen el **big data**: volumen, velocidad, variedad, veracidad y valor
- Distinguir los datos **estructurados**, **semiestructurados** y **no estructurados**
- Decidir si un caso es un problema de big data y qué V pesa más
- Explicar cómo se relacionan el big data, el **análisis de datos**, el **machine learning**, el **deep learning** y la IA
- Nombrar los cuatro tipos de análisis por la pregunta que responden
- Evitar la trampa: big data no significa simplemente «muchos datos»

Qué hace que unos datos sean «big»

> El **big data** son conjuntos de datos que, por su volumen, por la velocidad a la que llegan o por su variedad, no se pueden almacenar, procesar ni analizar de forma eficiente con herramientas tradicionales (una hoja de cálculo, una única base de datos en un servidor) y necesitan una arquitectura **escalable**.

En el juego de construir ciudades, un pueblo pequeño tiene unas pocas decenas de sensores y podrías leerlos uno a uno. Una ciudad de un millón de habitantes tiene contadores de tráfico, contadores de agua, GPS en los autobuses y ciudadanos quejándose en las redes sociales, todo a la vez, cada segundo. Las capas siguen funcionando, pero detrás ya no hay un ordenador: hay todo un centro de datos.

Las tres primeras V las propuso el analista Doug Laney en una nota de investigación de 2001 para META Group (hoy parte de Gartner). La mayoría de los manuales actuales añaden dos más.

V	Qué significa	Ejemplo
Volumen	La cantidad de datos es muy grande y no para de crecer	Cada búsqueda de vuelo y hotel que recibe una web de viajes cada día
Velocidad	Los datos llegan rápido y hay que tratarlos casi en tiempo real	La posición GPS de cada autobús de una red, enviada cada pocos segundos

V	Qué significa	Ejemplo
Variedad	Muchos formatos y fuentes mezclados	Reservas (tablas), opiniones (texto), fotos, registros del servidor (JSON), sensores
Veracidad	Calidad y fiabilidad: datos incompletos, duplicados, erróneos o sesgados	Opiniones falsas, sensores mal calibrados, el mismo cliente escrito dos veces
Valor	Solo vale la pena guardar los datos si aportan un beneficio	Prever la ocupación para ajustar precios y personal

Algunas fuentes añaden más V (variabilidad, validez, visualización...). El Instituto Nacional de Estándares y Tecnología de EE. UU. (NIST), por ejemplo, añade la **variabilidad**. No hace falta que las memorices: volumen, velocidad y variedad están en casi todas las definiciones, y veracidad y valor completan las cinco habituales.

Estructurados, semiestructurados, no estructurados

La variedad se entiende mejor si clasificas los datos según cuánta estructura tienen.

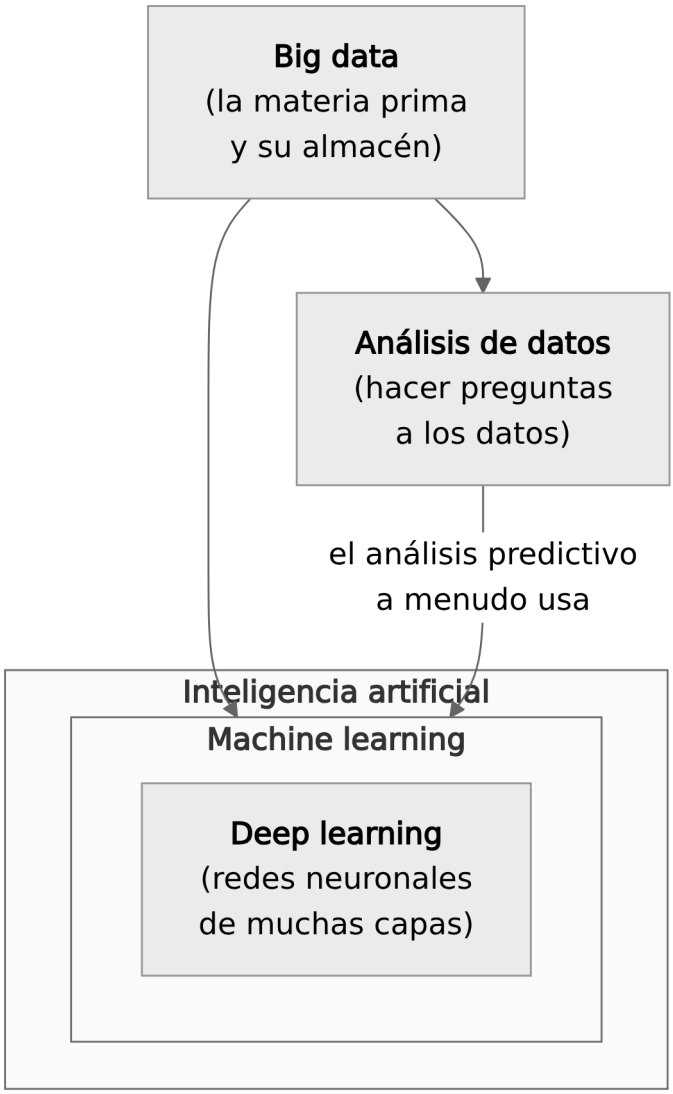
Tipo	Qué aspecto tiene	Ejemplo
Estructurados	Filas y columnas con un esquema fijo	Una tabla SQL de reservas
Semiestructurados	Etiquetas o claves, pero sin un esquema rígido	JSON de una API, XML, un registro del servidor
No estructurados	Contenido libre sin campos predefinidos	Texto de opiniones, fotos, audio, vídeo

> **Cuidado: big data no es «muchos datos»**
> Un archivo enorme pero uniforme que nunca cambia no es el mismo problema que un flujo continuo de datos mezclados. Y al revés: **no todas las empresas necesitan big data**. Muchas pymes están perfectamente servidas con una buena base de datos y alguien que sepa hacerle preguntas. La prueba no es «¿cuántos gigabytes?», sino «¿aquí deja de funcionar la herramienta tradicional, y por qué?».

Quién es quién: big data, análisis, ML, DL e IA

Término	Qué es
Inteligencia artificial (IA)	Sistemas que hacen tareas que asociamos a la inteligencia humana (la viste en la Unidad 4)
Machine learning (ML)	La parte de la IA en la que el sistema **aprende de los datos** en lugar de seguir reglas escritas a mano
Deep learning (DL)	La parte del ML basada en redes neuronales de muchas capas; suele necesitar muchos datos y mucha potencia de cálculo
Big data	No es IA. Es la **materia prima** y la infraestructura para guardarla y moverla
Análisis de datos	El proceso de examinar datos para sacar información de ellos; puede usar estadística sencilla o ML

Big data sin análisis es un almacén lleno que nadie mira. Análisis sin buenos datos da malas respuestas. Y buena parte de la IA que las empresas usan de verdad es análisis predictivo hecho a gran escala. Por eso los tres suelen ir juntos.



Cuatro preguntas, cuatro tipos de análisis

Tipo	Pregunta	Ejemplo del hotel	Herramienta típica
Descriptivo	¿Qué pasó?	Ocupación media en julio	Informes, gráficos
Diagnóstico	¿Por qué pasó?	La ocupación bajó la semana en que se canceló un evento	Cruzar fuentes de datos
Predictivo	¿Qué pasará?	Reservas esperadas para la semana que viene	ML
Prescriptivo	¿Qué deberíamos hacer?	Precio recomendado por habitación y día	ML más optimización

En clase: big data en el sector de tu empresa

Trabaja **en parejas**, sobre el sector de la empresa del proyecto de tu grupo. Es una búsqueda guiada en la web: cada afirmación necesita una fuente que funcione. **Sin herramientas de IA.**

Paso	Haz esto	Deberías ver
1	Busca un caso **real y documentado** de big data o de analítica de datos en el sector de tu empresa, a ser posible en las Illes Balears o en España (una nota de prensa de una empresa, una noticia, un proyecto público, un caso de estudio). Prueba búsquedas como `big data turismo Baleares`, `analítica de datos agua contadores inteligentes`, `big data puerto`	Dos o tres artículos candidatos
2	Elige el más fiable: quién lo publica, cuándo, ¿nombra a la empresa y los datos? Copia la URL y escribe la cita completa.	Una fuente de confianza
3	Rellena una tabla: qué **V** están presentes y por qué (al menos dos), qué tipos de datos (estructurados, semiestructurados, no estructurados) y qué **tipo de análisis** es (descriptivo, diagnóstico, predictivo, prescriptivo).	Una tabla de cinco filas
4	Di si el caso usa **ML** o **DL**, o solo estadística, y cómo lo sabes a partir de la fuente. Si la fuente no lo dice, escribe «no consta».	Una frase
5	Ahora tu propia empresa: ¿tiene, o podría tener, un problema de big data? Justifícalo con al menos dos V. Escribe una **pregunta predictiva** que le gustaría responder con sus datos.	Dos párrafos cortos
6	Entrega el documento en Classroom (uno por pareja, con los dos nombres).	

Para tu proyecto

Los pasos 3 a 5 son materia prima para la **sección B (miniproceso de ciencia de datos)** de la Entrega 5: la sección B parte de una pregunta, y la que escribes hoy es una buena candidata. El caso real que has encontrado también puede ir a la parte de grupo como contexto del sector. Cítalo.

Ideas clave

- El **big data** son datos que las herramientas tradicionales no pueden tratar de forma eficiente por su **volumen**, su **velocidad** o su **variedad**; la **veracidad** y el **valor** completan las cinco V habituales.
- Los datos pueden ser **estructurados**, **semiestructurados** o **no estructurados**; la variedad consiste en mezclarlos.
- Big data no es «muchos datos», y no todas las empresas lo necesitan.
- El **deep learning** está dentro del **machine learning**, que está dentro de la **IA**; el big data es la materia prima, y el **análisis de datos** es hacer preguntas a los datos.
- Los cuatro tipos de análisis responden cuatro preguntas: qué pasó, por qué, qué pasará, qué deberíamos hacer.

Comprobación rápida

1 ¿Qué tres V aparecen en casi todas las definiciones de big data?

A Valor, veracidad y visualización
B Volumen, velocidad y variedad
C Volumen, valor y validez
D Velocidad, veracidad y variabilidad

2 Texto de opiniones, fotos, lecturas de sensores y tablas de reservas, todo mezclado en un mismo sistema. ¿Qué V es esa?

- **A** Volumen
- **B** Velocidad
- **C** Variedad
- **D** Valor

3 Una gestoría guarda sus 800 clientes en una hoja de cálculo. ¿Es big data?

- **A** Sí, porque contiene datos personales
- **B** No: pocos datos, estructurados y estables, que una base de datos normal maneja bien
- **C** Sí, porque cualquier lista de clientes es big data
- **D** Depende solo del tamaño del archivo en megabytes

4 ¿Qué afirmación sobre la IA, el ML y el DL es correcta?

- **A** El machine learning está dentro del deep learning, que está dentro de la IA
- **B** El deep learning está dentro del machine learning, que está dentro de la IA
- **C** El big data es un tipo de IA
- **D** IA, ML y DL son tres nombres para lo mismo

5 «¿Cuántas reservas tendremos la semana que viene?» ¿Qué tipo de análisis lo responde?

- **A** Descriptivo
- **B** Diagnóstico
- **C** Predictivo
- **D** Prescriptivo

6 ¿Qué V trata de si nos podemos fiar de los datos?

- **A** Velocidad
- **B** Veracidad
- **C** Volumen
- **D** Variedad

7 Una empresa de agua de Mallorca lee miles de contadores inteligentes cada hora y además recibe quejas telefónicas por fugas. Elige las dos afirmaciones verdaderas.

- **A** La velocidad y la variedad pesan mucho: lecturas frecuentes más quejas de voz o de texto
- **B** No puede ser big data porque el agua es un sistema OT
- **C** Un modelo que avise de una fuga probable antes de que alguien la comunique sería análisis predictivo
- **D** Solo es big data si usa deep learning

Ya sabes qué es el big data y cómo alimenta el análisis y la IA. Ahora harás tú el análisis, paso a paso, sobre un conjunto de datos real, y decidirás dónde deben vivir los datos: Ciencia de datos y nube.

Soluciones

1. B · Las tres clásicas, propuestas por Doug Laney en 2001, son **volumen**, **velocidad** y **variedad**. La veracidad y el valor se añadieron después para formar las cinco habituales.

2. C · Muchos **formatos y fuentes** a la vez, estructurados y no estructurados: eso es **variedad**.

3. B · Aquí no deja de funcionar ninguna herramienta tradicional. Los datos son pocos, **estructurados** y estables. El big data lo define el problema, no la palabra «datos».

4. B · El **DL** (redes neuronales de muchas capas) es una parte del **ML** (aprender de los datos), que es una parte de la **IA**. El big data es la materia prima, no IA.

5. C · «¿Qué pasará?» es la pregunta **predictiva**. «¿Qué deberíamos hacer al respecto?» sería prescriptiva.

6. B · La **veracidad** trata de la calidad y la fiabilidad: las opiniones falsas, los clientes duplicados y los sensores mal calibrados la hacen bajar.

7. A, C · Sector nuevo, mismas herramientas. Lecturas frecuentes de contadores significan **velocidad**, lecturas más quejas significan **variedad**, y avisar antes de que pase es **análisis predictivo**. El big data no exige deep learning.