El ***Big Data*** o **Datos masivos** es un concepto que hace referencia a la **acumulación masiva de datos** y a los procedimientos usados para identificar patrones recurrentes dentro de esos datos. Esta cantidad de datos supera la capacidad del software habitual de capturarlos, gestionarlos y procesarlos en un tiempo razonable.

Big Data se puede dividir en tres tipos:

* **Datos estructurados (*Structured Data*):** Datos que tienen bien definidos su longitud y su formato, como las fechas, los números o las cadenas de caracteres. Se almacenan en tablas. Un ejemplo son las [bases de datos](https://es.wikipedia.org/wiki/Bases_de_datos) relacionales y las [hojas de cálculo](https://es.wikipedia.org/wiki/Hojas_de_c%C3%A1lculo).
* **Datos no estructurados (*Unstructured Data*):** Datos en el formato tal y como fueron recolectados, carecen de un formato específico. No se pueden almacenar dentro de una tabla ya que no se puede desgranar su información a tipos básicos de datos. Algunos ejemplos son los [PDF](https://es.wikipedia.org/wiki/PDF), documentos multimedia, [e-mails](https://es.wikipedia.org/wiki/E-mails) o documentos de texto.
* **Datos semiestructurados (*Semistructured Data*):** Datos que no se limitan a campos determinados, pero que contiene marcadores para separar los diferentes elementos. Es una información poco regular como para ser gestionada de una forma estándar. Estos datos poseen sus propios [metadatos](https://es.wikipedia.org/wiki/Metadatos) semiestructurados[16](https://es.wikipedia.org/wiki/Big_data#cite_note-16) que describen los objetos y las relaciones entre ellos, y pueden acabar siendo aceptados por convención. Un ejemplo es el [HTML](https://es.wikipedia.org/wiki/HTML), el [XML](https://es.wikipedia.org/wiki/XML) o el [JSON](https://es.wikipedia.org/wiki/JSON).

Estos datos obtenidos se pueden clasificar según su fuente de procedencia:

* **Generados por las personas:** Enviar correos electrónicos por e-mail o mensajes por [WhatsApp](https://es.wikipedia.org/wiki/WhatsApp), postear en [Facebook](https://es.wikipedia.org/wiki/Facebook), [tuitear](https://es.wikipedia.org/wiki/Twitter) contenidos o responder a una encuesta por la calle son algunas acciones cuotidianas que crean nuevos datos y [metadatos](https://es.wikipedia.org/wiki/Metadatos) que pueden ser analizados. Se estima que cada minuto al día se envían más de 200 millones de [e-mails](https://es.wikipedia.org/wiki/E-mail), se comparten más de 700.000 piezas de contenido en [Facebook](https://es.wikipedia.org/wiki/Facebook), se realizan dos millones de búsquedas en Google o se editan 48 horas de vídeo en YouTube.[19](https://es.wikipedia.org/wiki/Big_data#cite_note-19)Por otro lado, las trazas de utilización en un sistema [ERP](https://es.wikipedia.org/wiki/ERP), incluir registros en una [base de datos](https://es.wikipedia.org/wiki/Base_de_datos) o introducir información en una [hoja de cálculo](https://es.wikipedia.org/wiki/Hoja_de_c%C3%A1lculo) son otras formas de generar estos datos.
* **Transacciones de datos:** La facturación, las llamadas o las [transacción](https://es.wikipedia.org/wiki/Transacciones) entre cuentas generan información que tratada pueden ser datos relevantes. Un ejemplo más claro lo encontraremos en las transacciones bancarias: lo que el usuario conoce como un ingreso de X euros, la [computación](https://es.wikipedia.org/wiki/Computaci%C3%B3n) lo interpretará como una acción llevada a cabo en una fecha y momento determinado, en un lugar concreto, entre unos usuarios registrados, y más [metadatos](https://es.wikipedia.org/wiki/Metadatos).
* ***E-marketing* y web:** Generamos una gran cantidad de datos cuando navegamos por [internet](https://es.wikipedia.org/wiki/Internet). Con la [web](https://es.wikipedia.org/wiki/Web) 2.0 se ha roto el [paradigma](https://es.wikipedia.org/wiki/Paradigma) [webmaster](https://es.wikipedia.org/wiki/Webmaster" \o "Webmaster)-contenido-lector y los mismos usuarios se convierten en creadores de contenido gracias a su interacción con el sitio. Existen muchas herramientas de [tracking](https://es.wikipedia.org/w/index.php?title=Tracking&action=edit&redlink=1) utilizadas en su mayoría con fines de [marketing](https://es.wikipedia.org/wiki/Marketing) y [análisis de negocio](https://es.wikipedia.org/w/index.php?title=An%C3%A1lisis_de_negocio&action=edit&redlink=1). Los movimientos de ratón quedan grabados en [mapas de calor](https://es.wikipedia.org/w/index.php?title=Mapa_de_calor&action=edit&redlink=1) y queda registro de cuánto pasamos en cada página y cuándo las visitamos.
* ***Machine to Machine* (M2M):** Son las tecnologías que comparten datos con dispositivos: medidores, [sensores](https://es.wikipedia.org/wiki/Sensor) de temperatura, de luz, de altura, de presión, de sonido… que transforman las magnitudes físicas o químicas y las convierten en datos. Existen desde hace décadas, pero la llegada de las comunicaciones inalámbricas ([Wi-Fi](https://es.wikipedia.org/wiki/Wi-Fi" \o "Wi-Fi), [Bluetooth](https://es.wikipedia.org/wiki/Bluetooth" \o "Bluetooth),[RFID](https://es.wikipedia.org/wiki/RFID)…) ha revolucionado el mundo de los sensores. Algunos ejemplos son los [GPS](https://es.wikipedia.org/wiki/GPS) en la automoción o los sensores de signos vitales en la medicina.
* **Biométrica:** Son el conjunto de datos que provienen de la seguridad, defensa y servicios de inteligencia.[20](https://es.wikipedia.org/wiki/Big_data#cite_note-20) Son cantidades de datos generados por [lectores biométricos](https://es.wikipedia.org/w/index.php?title=Lector_biom%C3%A9trico&action=edit&redlink=1)como escáneres de retina, escáneres de [huellas digitales](https://es.wikipedia.org/wiki/Huella_digital), o lectores de cadenas de [ADN](https://es.wikipedia.org/wiki/ADN). El propósito de estos datos es proporcionar mecanismos de seguridad y suelen estar custodiadas por los ministerios de defensa y departamentos de inteligencia. Un ejemplo de aplicación es el cruce de [ADN](https://es.wikipedia.org/wiki/ADN) entre una muestra de un crimen y una muestra en nuestra [base de datos](https://es.wikipedia.org/wiki/Base_de_datos).

Dentro de Big Data encontramos NoSQL, nombre que hace referencia a Not Only SQL o No Solo SQL, el cual es un sistema que almacenamiento de datos que no cumple con el esquema entidad-relación que encontramos en las bases de datos relacionales.

Los objetivos de una base de datos NoSQL son:

* Presentar simpleza en el diseño de la Base de datos.
* Permitir escalabilidad horizontal. Escalabilidad horizontal se refiere a la posibilidad de aumentar el rendimiento del sistema simplemente añadiendo más nodos, sin necesidad en muchos casos de realizar ninguna otra operación más que indicar al sistema cuáles son los nodos disponibles
* Permitir controlar la disponibilidad de datos.

A diferencia de las bases de datos relacionales, NoSQL provee un sistema de almacenamiento mucho más flexible y concurrente, y permite manipular grandes cantidades de información de manera mucho más rápida. Son de esquema libre, fáciles de replicar, y cuentan con APIs para acceder de manera simple a los datos. Sin embargo, también presenta ciertas desventajas en relación a las bases de datos relacionales, tales como, no soportan operaciones JOIN, y no garantizan ACID, lo que hace referencia a Atomicidad, Consistencia, Aislamiento y Durabilidad.

Según la manera de almacenar los datos, podemos calificar a las bases de datos NoSQL en cuatro tipos:

* **Almacenamiento Clave-Valor (Key-Value):** Los datos se almacenan de forma similar a los maps o [diccionarios de datos](https://es.wikipedia.org/wiki/Diccionario_de_datos), donde se accede al dato a partir de una clave única.[23](https://es.wikipedia.org/wiki/Big_data#cite_note-survey-23) Los valores (datos) son aislados e independientes entre ellos, y no son interpretados por el sistema. Pueden ser [variables](https://es.wikipedia.org/wiki/Variable) simples como [enteros](https://es.wikipedia.org/wiki/Enteros) o [caracteres](https://es.wikipedia.org/wiki/Car%C3%A1cter), u[objetos](https://es.wikipedia.org/wiki/OOP). Por otro lado, este sistema de almacenamiento carece de una estructura de datos clara y establecida, por lo que no requiere un formateo de los datos muy estricto.[24](https://es.wikipedia.org/wiki/Big_data#cite_note-24)

Son útiles para operaciones simples basadas en las claves. Un ejemplo es el aumento de velocidad de carga de un sitio [web](https://es.wikipedia.org/wiki/Web) que pueden utilizar diferentes perfiles de [usuario](https://es.wikipedia.org/wiki/Usuario), teniendo mapeados los archivos que hay que incluir según el id de usuario y que han sido calculados con anterioridad. [Cassandra](https://es.wikipedia.org/wiki/Cassandra" \o "Cassandra) es la tecnología de almacenamiento clave-valor más reconocida por los usuarios.[25](https://es.wikipedia.org/wiki/Big_data#cite_note-Bianchi-25)

* **Almacenamiento Documental:** Las [bases de datos documentales](https://es.wikipedia.org/wiki/Base_de_datos_documental) guardan un gran parecido con las bases de datos Clave-Valor, diferenciándose en el dato que guardan. Si en la anterior no requería una [estructura de datos](https://es.wikipedia.org/wiki/Estructura_de_datos) concreta, en este caso guardamos datos semiestructurados.[25](https://es.wikipedia.org/wiki/Big_data#cite_note-Bianchi-25) Estos datos pasan a llamarse documentos, y pueden estar formateados en [XML](https://es.wikipedia.org/wiki/XML), [JSON](https://es.wikipedia.org/wiki/JSON), [Binary JSON](https://es.wikipedia.org/wiki/BSON" \o "BSON) o el que acepte la misma [base de datos](https://es.wikipedia.org/wiki/Base_de_datos).

Todos los documentos tienen una clave única con la que puede ser accedido e identificado explícitamente. Estos documentos no son opacos al sistema, por lo que pueden ser interpretados y lanzar queries sobre ellos.[23](https://es.wikipedia.org/wiki/Big_data#cite_note-survey-23) Un ejemplo que aclare cómo se usa lo encontramos en un [blog](https://es.wikipedia.org/wiki/Blog): se almacena el autor, la fecha, el título, el resumen y el contenido del post.

[CouchDB](https://es.wikipedia.org/wiki/CouchDB) o [MongoDB](https://es.wikipedia.org/wiki/MongoDB)[25](https://es.wikipedia.org/wiki/Big_data#cite_note-Bianchi-25) son quizá las más conocidas. Hay que hacer mención especial a [MapReduce](https://es.wikipedia.org/wiki/MapReduce" \o "MapReduce), una tecnología de [Google](https://es.wikipedia.org/wiki/Google) inicialmente diseñada para su algoritmo[PageRank](https://es.wikipedia.org/wiki/PageRank), que permite seleccionar un subconjunto de datos, agruparlos o reducirlos y cargarlos en otra colección, y a [Hadoop](https://es.wikipedia.org/wiki/Hadoop" \o "Hadoop) que es una tecnología de [Apache](https://es.wikipedia.org/wiki/Servidor_HTTP_Apache) diseñada para almacenar y procesar grandes cantidades de datos.

* **Almacenamiento en Grafo:** Las [bases de datos en grafo](https://es.wikipedia.org/wiki/Base_de_datos_orientada_a_grafos) rompen con la idea de tablas y se basan en la [teoría de grafos](https://es.wikipedia.org/wiki/Teor%C3%ADa_de_grafos), donde se establece que la información son los nodos y las relaciones entre la información son las aristas,[25](https://es.wikipedia.org/wiki/Big_data#cite_note-Bianchi-25) algo similar en el [modelo relacional](https://es.wikipedia.org/wiki/Modelo_relacional). Su mayor uso se contempla en casos de relacionar grandes cantidades de datos que pueden ser muy variables. Por ejemplo, los [nodos](https://es.wikipedia.org/wiki/Nodo_(inform%C3%A1tica)) pueden contener [objetos](https://es.wikipedia.org/wiki/Objeto_(programaci%C3%B3n)), [variables](https://es.wikipedia.org/wiki/Variable_(programaci%C3%B3n)) y [atributos](https://es.wikipedia.org/wiki/Atributo_(inform%C3%A1tica)) diferentes en unos y los otros. Las operaciones de [join](https://es.wikipedia.org/wiki/Join" \o "Join) se sustituyen por recorridos a través del grafo, y se guarda una [lista de adyacencias](https://es.wikipedia.org/wiki/Lista_de_adyacencia) entre los nodos.[23](https://es.wikipedia.org/wiki/Big_data#cite_note-survey-23) Encontramos un ejemplo en las redes sociales: en [Facebook](https://es.wikipedia.org/wiki/Facebook) cada nodo se considera un [usuario](https://es.wikipedia.org/wiki/Usuario), que puede tener [aristasde](https://es.wikipedia.org/wiki/Arista_(teor%C3%ADa_de_grafos)" \o "Arista (teoría de grafos)) amistad con otros usuarios, o [aristas](https://es.wikipedia.org/wiki/Arista_(teor%C3%ADa_de_grafos)) de publicación con [nodosde](https://es.wikipedia.org/wiki/Nodo_(inform%C3%A1tica)" \o "Nodo (informática)) contenidos. Soluciones como [Neo4J](https://es.wikipedia.org/w/index.php?title=Neo4J&action=edit&redlink=1) y GraphDB[25](https://es.wikipedia.org/wiki/Big_data#cite_note-Bianchi-25) son las más conocidas dentro de las [bases de datos en grafo](https://es.wikipedia.org/wiki/Base_de_datos_orientada_a_grafos).
* **Almacenamiento Orientado a Columnas:** Por último, el almacenamiento Column-Oriented es parecido al [Documental](https://es.wikipedia.org/wiki/Base_de_datos_documental). Su modelo de datos es definido como “un mapa de datos [multidimensional](https://es.wikipedia.org/w/index.php?title=Multidimensional&action=edit&redlink=1) poco denso, distribuido y persistente”.[23](https://es.wikipedia.org/wiki/Big_data#cite_note-survey-23) Se orienta a almacenar datos con tendencia a escalar horizontalmente, por lo que permite guardar diferentes[atributos](https://es.wikipedia.org/wiki/Atributo_(inform%C3%A1tica)) y [objetos](https://es.wikipedia.org/wiki/Objeto_(programaci%C3%B3n)) bajo una misma Clave. A diferencia del [Documental](https://es.wikipedia.org/wiki/Base_de_datos_documental) y el Key-Value, en este caso podremos almacenar varios [atributos](https://es.wikipedia.org/wiki/Atributo_(inform%C3%A1tica)) y [objetos](https://es.wikipedia.org/wiki/Objeto_(programaci%C3%B3n)), pero no serán interpretables directamente por el sistema. Permite agrupar columnas en famílias y guardar la información cronológicamente, mejorando el rendimiento. Esta tecnología se acostuma a usar en casos con 100 o más atributos por clave.[25](https://es.wikipedia.org/wiki/Big_data#cite_note-Bianchi-25) Su precursor es [BigTable](https://es.wikipedia.org/wiki/BigTable" \o "BigTable) de Google, pero han aparecido nuevas soluciones como HBase o HyperTable.

CoachDB

Es un gestor de bases de datos de código abierto, de tipo de almacenamiento documental. La misma emplea JSON para almacenar los datos, Javascript como lenguaje de consulta y el protocolo HTTP como API para acceder a los datos.

CouchDB in the wild

CouchDB se utiliza en distintas aplicaciones, como son el Horóscopo y las tarjetas de felicitaciones de cumpleaños en Facebook.

Dentro del sector empresarial, podemos encontrar CouchDB en:

* Ubuntu: Lo utilizaba para el servicio de sincronización “Ubuntu One”.
* BBC: Para su plataforma de contenidos dinámicos
* Credit Suisse: Es una empresa de servicios financieros. Ofrece asesoría financiera sobre productos bancarios. Y utiliza CoachDB como base de datos de su intranet y para obtener datos sobre los mercados actuales.
* Meebo: Era una plataforma web, que permitia usar todos los tipos de mensajería (Google Talk, Facebook Messenger, Windows Messenger, Yahoo Messenger), hasta que fue comprada por Google y fue descontinuada.

Por que elegir CoachDB?

* Semántica ACID: CouchDB provee una semántica de atomicidad, consistencia, aislamiento y durabilidad. Esto lo hace implementando una forma de control de concurrencia multiversión, lo que significa que puede manejar un gran número de lectores y escritores en paralelo, sin que surjan conflictos
* Arquitectura distribuida con replicación: se diseñó con teniendo en mente la replicación bidireccional (o sincronización) y la operación off-line. Eso significa que múltiples réplicas pueden tener cada una sus propias copias de los mismos datos, modificarlas y luego sincronizar esos cambios en un momento posterior
* Interfaz REST: Todos los ítems tienen una URI única que queda expuesta vía HTTP. REST usa los métodos HTTP POST, GET, PUT y DELETE para las cuatro operaciones básicas CRUD (Create, Read, Update, Delete).
* Gran tolerancia a fallos: CouchDB garantiza consistencia eventual para poder ofrecer tanto disponibilidad como tolerancia a las particiones.
* Operación offline: Está creada para funcionar de manera offline sin ningún tipo de problemas, pudiendo replicar datos a dispositivos (como smartphones) que pueden quedar offline y manejar automáticamente la sincronización de los datos cuando el dispositivo vuelve a estar en línea.