Un proyecto de BigData se compone de 4 grandes componentes:

* Captar información:

Donde está la información y como hago para tomarla

* Almacenar la información:

Que dato vamos a guardar y que volumen

* Procesar información:

Utilización de software específico para cada caso

* Visualizar información o hacer análisis:

Es una parte mas humana, que se hace en equipo en pro del negocio u objetivo del proyecto

Importante la seguridad, debemos salvaguardar los datos y la privacidad de estos, con los cuales trabajamos.

Para llevar a cabo un análisis de BigData se tiene que estar formado y tener habilidades en la materia.

Hizo que surjan puestos nuevos, tales como: DATA AND ANALYTICS MANAGER, DATA SCIENTIST, DATA ENGINEER, DATA ANALYST, BUSINESS ANALYST, DATA ARCHITECT.

Video de ADEN

Al usar la tecnología enviamos información personal también. Las redes sociales como Facebook tienen nuestro permiso de recolectar nuestra información, la de nuestros conocidos, la información nuestra que publican nuestros conocidos y la interacciones con otros usuarios

Según estimaciones de la empresa IBM diariamente se generan alrededor de 2.5 millones de GB de información, lo que implica que en 2 años se genera el 90% de la data existente en la historia humana. Esta cantidad de información es llamada BigData.

Puede ser almacenada, procesada y analizada en detalle por poderosos algoritmos capaces de encontrar patrones, asociaciones y tendencias.

Aplicaciones posibles:

Detectar el estado de animo del usuario de una app (Ginger.io)

Detección de brotes virales a raíz de búsquedas en Google.

Detección de calzadas deterioradas a partir de la información que brindan los teléfonos y sensores en sistemas de suspensión y dirección con GPS.

Video de DNews español

Hablar de BigData es hablar de información generada de a millones de datos por segundo.

BigData permite:

Que las decisiones que antes eran basadas en experiencia e intuición serán hechas por el análisis de máquinas con grandes cantidades de datos

Que un auto sea capaz de tomar información del medio y reaccionar adecuadamente, mejor que un conductor humano ya que la información que toma será usada para hacer el viaje más fácil, seguro y eficiente.

El análisis de gran cantidad de datos médicos de diferentes localizaciones y demografías permitirán determinar que condiciones mejoran la eficiencia de ciertos tratamientos y cuáles no.

Revelar patrones y conexiones que van a mejorar las actividades humanas y también generar perfiles detallados de todos nosotros, incluyendo información que elegimos mantener privada

WILL BIGDATA MAKE PRIVACY OBSOLETE?

OR WILL IT BRING TRANSPARENCY, ACCOUNTABILITY AND PROGRESS

BigData hará obsoleta a la privacidad?

O traerá transparencia, responsabilidad y progreso.

La meta del BigData es generar valoraciones o revelaciones/insights

Concepto que hace referencia a grandes cantidades de información disponibles en diversos formatos y tipos de estructuras. Recopilada principalmente a través de internet mediante la interacción de los usuarios con computadoras, teléfonos celulares y dispositivos GPS, ente otros.

Los 4 problemas principales.

El 25 de mayo de 2018 comenzó el cumplimiento obligatorio del reglamento general de protección de datos o GDPR por sus siglas en ingles y este reglamento regula todo el manejo de información personal de los ciudadanos de la unión europea.

La cual establece que cualquier empresa sin importar su país de origen, que opera en la unión europea va a tener que adaptarse a la nueva normativa. Por eso muchas empresas están actualizando sus políticas de privacidad, para no infringir esta nueva normativa.

¿Porque nos afecta a nosotros también?

Porque las empresas decidieron unificar sus políticas de privacidad o los “términos y condiciones”. Para tener los mismos para todo el mundo y que todo el mundo los acepte por igual.

En que nos afecta/ beneficia esta nueva normativa?

A partir de ahora, nosotros como usuarios vamos a tener el derecho de poder pedir una copia gratuita y completa de toda la información nuestra que una empresa tenga. Algunas empresas como Facebook ya lo hacían de antes.

Además, podemos pedir que estos datos que la empresa posee nuestros, sean borrados. Además, las empresas deben notificarnos en un plazo de 72hs en caso de que nuestra información personal haya sido violada.

Además, el reglamento tiene contemplado el consentimiento del usuario para el uso de sus datos personales. Las empresas deberán pedir un consentimiento “activo” de los usuarios para poder usar sus datos personales. Estos pedidos de consentimiento estarán por separado al resto de consentimientos que el servicio/empresa requiera, como, por ejemplo, los famosos “Términos y condiciones”. Estos permisos de consentimiento deben estar expresados de forma corta y concisa para así poder leerlos rápidamente y decidir si damos nuestro consentimiento o no.

La acción “activa” y positiva de la que se hablo anteriormente es hacer que el usuario tenga que realizar una acción para poder aceptar estos permisos, por ejemplo, marcar una casilla. De la misma forma, el proceso de retirar el consentimiento del usuario debe ser rápido y sencillo

<https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX:32016R0679>

Link de la normativa

SupraPixel GDPR

“MetaData datos sobre datos”.

Pero los datos no son mas que información potencial, cruda y sin procesar.

Los datos no deben ser pensados como conceptos abstractos, sinó como objetos potencialmente informativos. MetaData puede entonces definirse como “un objeto potencialmente informativo que describe otro objeto potencialmente informativo”. Podemos definir metadata como una declaración sobre un objeto potencialmente informativo.

La metadata es la base de la bigdata.

Metdata (Por Jeffrey Pomerantz) -The MIT press essential knowledge series- -MIT Press-

Macrodatos

Big data, macrodatos,[1](https://es.wikipedia.org/wiki/Macrodatos#cite_note-1)​ datos masivos, inteligencia de datos o datos a gran escala es un concepto que hace referencia a un [conjuntos de datos](https://es.wikipedia.org/wiki/Conjuntos_de_datos) tan grandes que [aplicaciones informáticas](https://es.wikipedia.org/wiki/Aplicaci%C3%B3n_inform%C3%A1tica) tradicionales de [procesamiento de datos](https://es.wikipedia.org/wiki/Procesamiento_de_datos) no son suficientes para tratar con ellos y los procedimientos usados para encontrar patrones repetitivos dentro de esos datos. Los textos científicos en español con frecuencia se usa directamente el término en inglés big data, tal como aparece en el ensayo de Viktor Schönberger: La revolución de los datos masivos.[2](https://es.wikipedia.org/wiki/Macrodatos#cite_note-2)​

La disciplina dedicada a los datos masivos se enmarca en el sector de las [tecnologías de la información y la comunicación](https://es.wikipedia.org/wiki/Tecnolog%C3%ADas_de_la_informaci%C3%B3n_y_la_comunicaci%C3%B3n). Esta disciplina se ocupa de todas las actividades relacionadas con los sistemas que manipulan grandes [conjuntos de datos](https://es.wikipedia.org/wiki/Conjunto_de_datos). Las dificultades más habituales vinculadas a la gestión de estas cantidades de datos se centran en la recolección y el almacenamiento,[3](https://es.wikipedia.org/wiki/Macrodatos#cite_note-3)​ búsqueda, compartición, análisis,[4](https://es.wikipedia.org/wiki/Macrodatos#cite_note-4)​ y visualización. La tendencia a manipular enormes cantidades de datos se debe a la necesidad en muchos casos de incluir dicha información para la creación de informes estadísticos y modelos predictivos utilizados en diversas materias, como los análisis de negocio, publicitarios, los datos de enfermedades infecciosas, el espionaje y seguimiento a la población o la lucha contra el crimen organizado.[5](https://es.wikipedia.org/wiki/Macrodatos#cite_note-5)​

El límite superior de procesamiento ha ido creciendo a lo largo de los años. Se estima que el mundo almacenó unos 5 [zettabytes](https://es.wikipedia.org/wiki/Zettabyte" \o "Zettabyte) en 2014. Si se pone esta información en libros, convirtiendo las imágenes y todo eso a su equivalente en letras, se podría hacer 4500 pilas de libros que lleguen hasta el sol.[6](https://es.wikipedia.org/wiki/Macrodatos#cite_note-theclinic.cl-6)​ Los científicos con cierta regularidad encuentran límites en el análisis debido a la gran cantidad de datos en ciertas áreas, tales como la [meteorología](https://es.wikipedia.org/wiki/Meteorolog%C3%ADa), la [genómica](https://es.wikipedia.org/wiki/Gen%C3%B3mica),[7](https://es.wikipedia.org/wiki/Macrodatos#cite_note-7)​ la [conectómica](https://es.wikipedia.org/w/index.php?title=Conect%C3%B3mica&action=edit&redlink=1" \o "Conectómica (aún no redactado)), las complejas simulaciones de procesos físicos[8](https://es.wikipedia.org/wiki/Macrodatos#cite_note-8)​ y las investigaciones relacionadas con los procesos biológicos y ambientales,[9](https://es.wikipedia.org/wiki/Macrodatos#cite_note-9)​ Las limitaciones también afectan a los [motores de búsqueda en internet](https://es.wikipedia.org/wiki/Motor_de_b%C3%BAsqueda), a los sistemas [finanzas](https://es.wikipedia.org/wiki/Finanzas) y a la [informática de negocios](https://es.wikipedia.org/wiki/Inform%C3%A1tica_de_negocios). Los data sets crecen en volumen debido en parte a la recolección masiva de información procedente de los [sensores inalámbricos](https://es.wikipedia.org/w/index.php?title=Sensor_inal%C3%A1mbrico&action=edit&redlink=1) y los dispositivos móviles (por ejemplo las [VANET](https://es.wikipedia.org/wiki/VANET)), el constante crecimiento de los históricos de aplicaciones (por ejemplo de los [registros](https://es.wikipedia.org/wiki/Log_(inform%C3%A1tica))), cámaras (sistemas de [teledetección](https://es.wikipedia.org/wiki/Teledetecci%C3%B3n)), micrófonos, lectores de [identificación por radiofrecuencia](https://es.wikipedia.org/wiki/RFID).[10](https://es.wikipedia.org/wiki/Macrodatos#cite_note-10)​[11](https://es.wikipedia.org/wiki/Macrodatos#cite_note-11)​ La capacidad tecnológica per cápita a nivel mundial para almacenar datos se dobla aproximadamente cada cuarenta meses desde los [años 1980](https://es.wikipedia.org/wiki/A%C3%B1os_1980).[12](https://es.wikipedia.org/wiki/Macrodatos#cite_note-HilbertLopez2011-12)​ Se estima que en 2012 cada día fueron creados cerca de 2.5 trillones de bytes de datos.[13](https://es.wikipedia.org/wiki/Macrodatos#cite_note-13)​

Definición[[editar](https://es.wikipedia.org/w/index.php?title=Macrodatos&action=edit&section=1)]

Big data o macrodatos es un término que hace referencia a una cantidad de datos tal que supera la capacidad del software convencional para ser capturados, administrados y procesados en un tiempo razonable. El volumen de los datos masivos crece constantemente. En 2012 se estimaba su tamaño de entre una docena de [terabytes](https://es.wikipedia.org/wiki/Terabyte) hasta varios [petabytes](https://es.wikipedia.org/wiki/Petabyte" \o "Petabyte) de datos en un único conjunto de datos. En la metodología [MIKE2.0](https://es.wikipedia.org/w/index.php?title=MIKE2.0_methodology&action=edit&redlink=1) dedicada a investigar temas relacionados con la [gestión de información](https://es.wikipedia.org/w/index.php?title=Gesti%C3%B3n_de_informaci%C3%B3n&action=edit&redlink=1), definen big data[14](https://es.wikipedia.org/wiki/Macrodatos#cite_note-14)​ en términos de permutaciones útiles, complejidad y dificultad para borrar registros individuales.

Se ha definido también como datos lo suficientemente masivos como para poner de relieve cuestiones y preocupaciones en torno a la efectividad del anonimato desde una perspectiva más práctica que teórica.[15](https://es.wikipedia.org/wiki/Macrodatos#cite_note-15)​

En 2001, en un informe de investigación que se fundamentaba en congresos y presentaciones relacionadas,[16](https://es.wikipedia.org/wiki/Macrodatos#cite_note-16)​ la META Group (ahora [Gartner](https://es.wikipedia.org/wiki/Gartner_(empresa))) definía el crecimiento constante de datos como una oportunidad y un reto para investigar en el volumen, la velocidad y la variedad. Gartner' continúa usando datos masivos como referencia.[17](https://es.wikipedia.org/wiki/Macrodatos#cite_note-17)​ Además, grandes proveedores del mercado de datos masivos están desarrollando soluciones para atender las demandas más críticas sobre cómo procesar tal cantidad de datos, como MapR y Cloudera.

Características[[editar](https://es.wikipedia.org/w/index.php?title=Macrodatos&action=edit&section=2)]

Los macrodatos se pueden describir por las siguientes características:[18](https://es.wikipedia.org/wiki/Macrodatos#cite_note-18)​

Volumen: la cantidad de datos generados y guardado. El tamaño de los datos determina el valor y entendimiento potencial, y si los puede considerar como auténticos macrodatos.

Variedad: el tipo y naturaleza de los datos para ayudar a las personas a analizar los datos y usar los resultados de forma eficaz. Los macrodatos usan textos imágenes, audio y vídeo. También completan pedazos pedidos a través de la fusión de datos.

Velocidad: en este contexto, la velocidad a la cual se generan y procesan los datos para cumplir las exigencias y desafíos de su análisis.

Veracidad: la [calidad de los datos](https://es.wikipedia.org/wiki/Calidad_de_datos) capturados puede variar mucho y así afectar a los resultados del análisis.

Tecnología[[editar](https://es.wikipedia.org/w/index.php?title=Macrodatos&action=edit&section=3)]

Existen muchísimas herramientas para tratar con big data. Algunos ejemplos incluyen Hadoop, [NoSQL](https://es.wikipedia.org/wiki/NoSQL), Cassandra, [inteligencia empresarial](https://es.wikipedia.org/wiki/Inteligencia_empresarial), [aprendizaje automático](https://es.wikipedia.org/wiki/Aprendizaje_autom%C3%A1tico) y [MapReduce](https://es.wikipedia.org/wiki/MapReduce). Estas herramientas tratan con algunos de los tres tipos de big data:[19](https://es.wikipedia.org/wiki/Macrodatos#cite_note-19)​

Datos estructurados: 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 los [almacenes de datos](https://es.wikipedia.org/wiki/Almac%C3%A9n_de_datos).

Datos no estructurados: 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, [correos electrónicos](https://es.wikipedia.org/wiki/E-mails) o documentos de texto.

Datos semiestructurados: 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[20](https://es.wikipedia.org/wiki/Macrodatos#cite_note-20)​ que describen los objetos y las relaciones entre ellos, y pueden acabar siendo aceptados por convención. Como ejemplos tenemos los archivos tipo [hojas de cálculo](https://es.wikipedia.org/wiki/Hojas_de_c%C3%A1lculo), [HTML](https://es.wikipedia.org/wiki/HTML), [XML](https://es.wikipedia.org/wiki/XML) o [JSON](https://es.wikipedia.org/wiki/JSON).

Captura[[editar](https://es.wikipedia.org/w/index.php?title=Macrodatos&action=edit&section=4)]

¿De dónde provienen todos estos datos? Los fabricamos directa e indirectamente segundo tras segundo. Un [iPhone](https://es.wikipedia.org/wiki/IPhone) hoy en día tiene más capacidad de cómputo que la [NASA](https://es.wikipedia.org/wiki/NASA) cuando el hombre llegó a la luna[21](https://es.wikipedia.org/wiki/Macrodatos#cite_note-21)​ por lo que la cantidad de datos generados por persona y en unidad de tiempo es muy grande. Catalogamos la procedencia de los datos según las siguientes categorías:[22](https://es.wikipedia.org/wiki/Macrodatos#cite_note-22)​

Generados por las personas: el hecho de enviar correos electrónicos o mensajes por [WhatsApp](https://es.wikipedia.org/wiki/WhatsApp), publicar un estado 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 cosas que hacemos a diario y 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 [correos electrónicos](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.[23](https://es.wikipedia.org/wiki/Macrodatos#cite_note-23)​ Por otro lado, las trazas de utilización en un sistema [ERP](https://es.wikipedia.org/wiki/Sistema_de_planificaci%C3%B3n_de_recursos_empresariales), 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 [transacciones](https://es.wikipedia.org/wiki/Transacciones) entre cuentas generan información que tratada puede 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).

Marketing electrónico y web: se generan una gran cantidad de datos cuando se navega 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 seguimiento utilizadas en su mayoría con fines de [mercadotecnia](https://es.wikipedia.org/wiki/Mercadotecnia) 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.

Máquina a máquina (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), [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.[24](https://es.wikipedia.org/wiki/Macrodatos#cite_note-24)​ 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 custodiados 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).

Transformación[[editar](https://es.wikipedia.org/w/index.php?title=Macrodatos&action=edit&section=5)]

Una vez encontradas las fuentes de los datos necesarios, muy posiblemente dispongamos de un sinfín de tablas de origen que no estarán relacionadas. El siguiente objetivo es hacer que los datos se recojan en un mismo lugar y darles un formato.

Aquí entran en juego las plataformas [extraer, transformar y cargar](https://es.wikipedia.org/wiki/Extract,_transform_and_load) (ETL). Su propósito es extraer los datos de las diferentes fuentes y sistemas, para después hacer transformaciones (conversiones de datos, limpieza de [datos sucios](https://es.wikipedia.org/w/index.php?title=Dirty_data&action=edit&redlink=1), cambios de formato…) y finalmente cargar los datos en la base de datos o [almacén de datos](https://es.wikipedia.org/wiki/Data_Warehouse) especificada.[25](https://es.wikipedia.org/wiki/Macrodatos#cite_note-25)​ Un ejemplo de plataforma ETL es el [Pentaho](https://es.wikipedia.org/wiki/Pentaho) Data Integration, más concretamente su aplicación Spoon.

Almacenamiento NoSQL[[editar](https://es.wikipedia.org/w/index.php?title=Macrodatos&action=edit&section=6)]

El término [NoSQL](https://es.wikipedia.org/wiki/NoSQL) se refiere a Not Only SQL (no solo SQL) y son sistemas de almacenamiento que no cumplen con el esquema entidad-relación.[26](https://es.wikipedia.org/wiki/Macrodatos#cite_note-26)​ Proveen un sistema de almacenamiento mucho más flexible y concurrente y permiten manipular grandes cantidades de información de manera mucho más rápida que las [bases de datos relacionales](https://es.wikipedia.org/wiki/Base_de_datos).

Distinguimos cuatro grandes grupos de bases de datos [NoSQL](https://es.wikipedia.org/wiki/NoSQL):

Almacenamiento clave-valor (key-value): los datos se almacenan de forma similar a los mapas o [diccionarios de datos](https://es.wikipedia.org/wiki/Diccionario_de_datos), donde se accede al dato a partir de una clave única.[27](https://es.wikipedia.org/wiki/Macrodatos#cite_note-survey-27)​ 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_(programaci%C3%B3n)) simples como [enteros](https://es.wikipedia.org/wiki/Enteros) o caracteres, 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.[28](https://es.wikipedia.org/wiki/Macrodatos#cite_note-28)​

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 puede 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. [Apache Cassandra](https://es.wikipedia.org/wiki/Apache_Cassandra) es la tecnología de almacenamiento clave-valor más reconocida por los usuarios.[29](https://es.wikipedia.org/wiki/Macrodatos#cite_note-Bianchi-29)​

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 el anterior no se requería una [estructura de datos](https://es.wikipedia.org/wiki/Estructura_de_datos) concreta, en este caso guardamos datos semiestructurados.[29](https://es.wikipedia.org/wiki/Macrodatos#cite_note-Bianchi-29)​ 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 pueden ser accedidos e identificados explícitamente. Estos documentos no son opacos al sistema, por lo que pueden ser interpretados y lanzar queries sobre ellos.[27](https://es.wikipedia.org/wiki/Macrodatos#cite_note-survey-27)​ 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)[29](https://es.wikipedia.org/wiki/Macrodatos#cite_note-Bianchi-29)​ son quizá las más conocidas. Hay que hacer mención especial a [MapReduce](https://es.wikipedia.org/wiki/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) 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,[29](https://es.wikipedia.org/wiki/Macrodatos#cite_note-Bianchi-29)​ algo similar al [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 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.[27](https://es.wikipedia.org/wiki/Macrodatos#cite_note-survey-27)​ 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 [aristas](https://es.wikipedia.org/wiki/Arista_(teor%C3%ADa_de_grafos)) de amistad con otros usuarios, o [aristas](https://es.wikipedia.org/wiki/Arista_(teor%C3%ADa_de_grafos)) de publicación con [nodos](https://es.wikipedia.org/wiki/Nodo_(inform%C3%A1tica)) de contenidos. Soluciones como [Neo4J](https://es.wikipedia.org/w/index.php?title=Neo4J&action=edit&redlink=1) y GraphDB[29](https://es.wikipedia.org/wiki/Macrodatos#cite_note-Bianchi-29)​ 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, este almacenamiento 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».[27](https://es.wikipedia.org/wiki/Macrodatos#cite_note-survey-27)​ 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 clave-valor, en este caso se pueden 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 familias y guardar la información cronológicamente, mejorando el rendimiento. Esta tecnología se acostumbra a usar en casos con 100 o más atributos por clave.[29](https://es.wikipedia.org/wiki/Macrodatos#cite_note-Bianchi-29)​ Su precursor es [BigTable](https://es.wikipedia.org/wiki/BigTable" \o "BigTable) de Google, pero han aparecido nuevas soluciones como HBase o HyperTable.

Análisis de datos[[editar](https://es.wikipedia.org/w/index.php?title=Macrodatos&action=edit&section=7)]

Teniendo los datos necesarios almacenados según diferentes tecnologías de [almacenamiento](https://es.wikipedia.org/wiki/Dispositivo_de_almacenamiento_de_datos), nos daremos cuenta que necesitaremos diferentes técnicas de [análisis de datos](https://es.wikipedia.org/wiki/An%C3%A1lisis_de_datos) como las siguientes:

Asociación: permite encontrar relaciones entre diferentes variables.[30](https://es.wikipedia.org/wiki/Macrodatos#cite_note-30)​ Bajo la premisa de causalidad, se pretende encontrar una predicción en el comportamiento de otras variables. Estas relaciones pueden ser los sistemas de [ventas cruzadas](https://es.wikipedia.org/wiki/Venta_cruzada) en los [comercios electrónicos](https://es.wikipedia.org/wiki/E-commerce).

Minería de datos ([data mining](https://es.wikipedia.org/wiki/Data_Mining)): tiene como objetivo encontrar comportamientos predictivos. Engloba el conjunto de técnicas que combina métodos estadísticos y de [aprendizaje automático](https://es.wikipedia.org/wiki/Aprendizaje_autom%C3%A1tico) con almacenamiento en bases de datos.[31](https://es.wikipedia.org/wiki/Macrodatos#cite_note-Manyika-31)​ Está estrechamente relacionada con los [modelos](https://es.wikipedia.org/wiki/Modelo_estad%C3%ADstico) utilizados para descubrir patrones en grandes cantidades de datos.

Agrupación ([clustering](https://es.wikipedia.org/wiki/Clustering" \o "Clustering)): el análisis de clústeres es un tipo de [minería de datos](https://es.wikipedia.org/wiki/Data_Mining) que divide grandes grupos de individuos en grupos más pequeños de los cuales no conocíamos su parecido antes del análisis.[31](https://es.wikipedia.org/wiki/Macrodatos#cite_note-Manyika-31)​ El propósito es encontrar similitudes entre estos grupos, y el descubrimiento de nuevos, conociendo cuáles son las cualidades que lo definen. Es una metodología apropiada para encontrar relaciones entre resultados y hacer una evaluación preliminar de la estructura de los datos analizados. Existen diferentes técnicas y algoritmos de clusterización.[32](https://es.wikipedia.org/wiki/Macrodatos#cite_note-32)​

Análisis de texto ([text analytics](https://es.wikipedia.org/wiki/Miner%C3%ADa_de_textos" \o "Minería de textos)): gran parte de los datos generados por las personas son textos, como [correos](https://es.wikipedia.org/wiki/Correo_electr%C3%B3nico), búsquedas [web](https://es.wikipedia.org/wiki/Web) o contenidos. Esta metodología permite extraer información de estos datos y así modelar temas y asuntos o predecir palabras.[33](https://es.wikipedia.org/wiki/Macrodatos#cite_note-33)​

Visualización de datos[[editar](https://es.wikipedia.org/w/index.php?title=Macrodatos&action=edit&section=8)]

Tal y como el Instituto Nacional de Estadística dice en sus tutoriales, «una imagen vale más que mil palabras, o que mil datos». La mente agradece mucho más una presentación bien estructurada de resultados estadísticos en gráficos o mapas en vez de en tablas con números y conclusiones. En los macrodatos se llega un paso más allá: parafraseando a [Edward Tufte](https://es.wikipedia.org/wiki/Edward_Tufte), uno de los expertos en visualización de datos más reconocidos a nivel mundial «el mundo es complejo, dinámico, multidimensional, el papel es estático y plano. ¿Cómo vamos a representar la rica experiencia visual del mundo en la mera planicie?». [Mondrian](https://es.wikipedia.org/wiki/Mondrian_(inform%C3%A1tica))[34](https://es.wikipedia.org/wiki/Macrodatos#cite_note-34)​ es una plataforma que permite visualizar la información a través de los análisis llevados a cabo sobre los datos que disponemos. Con esta plataforma se intenta llegar a un público más concreto, y una utilidad más acotada como un cuadro de mando integral de una organización. Por otro lado, las [infografías](https://es.wikipedia.org/wiki/Infograf%C3%ADa) se han vuelto un fenómeno viral, donde se recogen los resultados de los diferentes análisis sobre nuestros datos, y son un material atractivo, entretenido y simplificado para audiencias masivas.[35](https://es.wikipedia.org/wiki/Macrodatos#cite_note-35)​

Utilidad[[editar](https://es.wikipedia.org/w/index.php?title=Macrodatos&action=edit&section=9)]

Este conjunto de tecnologías se puede usar en una gran variedad de ámbitos, como los siguientes.

Democracia[[editar](https://es.wikipedia.org/w/index.php?title=Macrodatos&action=edit&section=10)]

Los datos masivos se usan habitualmente para influenciar el proceso democrático. Los representantes del pueblo pueden ver todo lo que hacen los ciudadanos, y los ciudadanos pueden dictar la vida pública de los representantes mediante tuits y otros métodos de extender ideas en la sociedad. Las campañas presidenciales de Obama y Trump los usaron de manera generalizada[6](https://es.wikipedia.org/wiki/Macrodatos#cite_note-theclinic.cl-6)​ y hay expertos que advierten de que hay que «reinventar la democracia representativa. Si no, es posible que se convierta en una dictadura de la información».[36](https://es.wikipedia.org/wiki/Macrodatos#cite_note-36)​

Empresas[[editar](https://es.wikipedia.org/w/index.php?title=Macrodatos&action=edit&section=11)]

[Redes sociales](https://es.wikipedia.org/wiki/Red_Social)

Cada vez más los internautas tienden a subir a las redes sociales toda su actividad y la de sus conocidos. Las empresas utilizan esta información para cruzar los datos de los candidatos a un trabajo. [Oracle](https://es.wikipedia.org/wiki/Oracle) ha desarrollado una herramienta llamada [Taleo Social Sourcing](https://es.wikipedia.org/w/index.php?title=Taleo_Social_Sourcing&action=edit&redlink=1" \o "Taleo Social Sourcing (aún no redactado)),[37](https://es.wikipedia.org/wiki/Macrodatos#cite_note-37)​ la cual está integrada con las [API](https://es.wikipedia.org/wiki/API) de [Facebook](https://es.wikipedia.org/wiki/Facebook), [Twitter](https://es.wikipedia.org/wiki/Twitter) y [LinkedIn](https://es.wikipedia.org/wiki/LinkedIn). Gracias a su uso, los departamentos de recursos humanos pueden ver, entrando la identidad del candidato, su perfil social y profesional en cuestión de segundos. Por otro lado, les permite crear una lista de posibles candidatos según el perfil profesional necesario, y así pasar a ofrecer el puesto de trabajo a un público mucho más objetivo.

Por otro lado, [Gate Gourmet](https://es.wikipedia.org/wiki/Gate_Gourmet) —una compañía de cáterin para aerolíneas— experimentaba una tasa de abandono del 50 % de sus trabajadores asignados al aeropuerto de Chicago, y sospecharon que el problema existía en el tiempo dedicado al viaje. Para demostrarlo, hicieron análisis juntando varios data sets de sus sistemas internos y de otros externos como datos demográficos, datos de tráfico y datos de uso de redes sociales. Los resultados que obtuvieron fueron patrones que relacionaban muy consistentemente la alta tasa de abandono con la distancia del lugar de trabajo a casa y la accesibilidad al transporte público.[38](https://es.wikipedia.org/wiki/Macrodatos#cite_note-38)​  
Consumo

[Amazon](https://es.wikipedia.org/wiki/Amazon) es líder en ventas cruzadas. El éxito se basa en la [minería de datos](https://es.wikipedia.org/wiki/Miner%C3%ADa_de_datos) masiva basando los patrones de compra de un usuario cruzados con los datos de compra de otro, creando así anuncios personalizados y [boletines electrónicos](https://es.wikipedia.org/wiki/Newsletter) que incluyen justo aquello que el usuario quiere en ese instante.[39](https://es.wikipedia.org/wiki/Macrodatos#cite_note-39)​ También hay casos sin conexión de aplicaciones de macrodatos. Los teléfonos móviles envían peticiones de escucha [wifi](https://es.wikipedia.org/wiki/Wifi) a todos los puntos de acceso que se cruzan. Algunas compañías han decidido hacer un trazo de estas peticiones con su localización y [dirección MAC](https://es.wikipedia.org/wiki/Direcci%C3%B3n_MAC) para saber qué dispositivo hace cuál ruta dentro de un recinto. No hay que asustarse ya que con la dirección MAC no pueden invadir la intimidad.[40](https://es.wikipedia.org/wiki/Macrodatos#cite_note-40)​ Las grandes superficies aprovechan estos datos para sacar información como contabilizar cuánto tiempo pasan los clientes en su interior, qué rutas siguen, dónde permanecen más tiempo (ya sea escogiendo un producto o el tiempo de espera en caja) o cuál es la frecuencia de visita.

Macrodatos e intimidad

La cantidad de datos creados anualmente es de 2.8 [zettabytes](https://es.wikipedia.org/wiki/Zettabyte" \o "Zettabyte) en 2012, de los cuales el 75 % son generados por los individuos según su uso de la red ya sea bajarse un archivo, conectar el GPS o enviar un correo electrónico. Se calcula que un oficinista medio genera 1.8 [terabytes](https://es.wikipedia.org/wiki/Terabyte) al año por lo que son unos 5 GB al día de información.[41](https://es.wikipedia.org/wiki/Macrodatos#cite_note-41)​

Aquí entran en juego las empresas llamadas corredores de datos. Acxiom es una de ellas, y posee unas 1500 trazas de datos de más de 500 millones de usuarios de [internet](https://es.wikipedia.org/wiki/Internet). Todos estos datos son transformados y cruzados para incluir al usuario analizado en uno de los 70 segmentos de usuarios, llamado PersonicX.[42](https://es.wikipedia.org/wiki/Macrodatos#cite_note-42)​ Descrito como un «resumen de indicadores de estilo de vida, intereses y actividades», esta correduría de datos basa su [clusterización](https://es.wikipedia.org/wiki/Clustering" \o "Clustering) en los acontecimientos vitales y es capaz de predecir más de 3000 reacciones ante estímulos de estos clientes. En un primer momento captaba la información de los hechos reales y no virtuales de los usuarios.[43](https://es.wikipedia.org/wiki/Macrodatos#cite_note-43)​ En febrero de 2013, [Facebook](https://es.wikipedia.org/wiki/Facebook) acordó la cesión de la información personal de sus usuarios con Acxiom y otros corredores de datos, cruzando las actividades en línea como las de fuera de internet.

A nivel usuario, nos encontramos con [Google](https://es.wikipedia.org/wiki/Google) Location History: un servicio de [Google](https://es.wikipedia.org/wiki/Google) que registra las ubicaciones en las cuales ha estado un usuario que lleva el [móvil](https://es.wikipedia.org/wiki/Telefon%C3%ADa_m%C3%B3vil) encima, y con el [servicio de localización](https://es.wikipedia.org/wiki/Servicio_basado_en_localizaci%C3%B3n) activado (que por defecto lo está en los terminales con [Android](https://es.wikipedia.org/wiki/Android)). Al acceder a él muestra un mapa con las rutas que ha seguido el usuario, con la hora de llegada y salida de cada ubicación.[44](https://es.wikipedia.org/wiki/Macrodatos#cite_note-44)​ Gracias a esta utilidad se crean [algoritmos](https://es.wikipedia.org/wiki/Algoritmo) de recomendaciones de amigos en [redes sociales](https://es.wikipedia.org/wiki/Red_social) y ubicaciones a visitar basados en el historial de ubicaciones del usuario.[45](https://es.wikipedia.org/wiki/Macrodatos#cite_note-45)​

Por último, [Google](https://es.wikipedia.org/wiki/Google) pagó 3200 millones de dólares por Nest, una empresa de detectores de humo.[46](https://es.wikipedia.org/wiki/Macrodatos#cite_note-46)​ Ahora Google ha abierto Nest Labs, donde se pretende desarrollar [sensores](https://es.wikipedia.org/wiki/Sensor) para convertir la casa en una vivienda inteligente. Incluir estos sensores permitirá saber por ejemplo cuándo hay alguien o no en casa gracias a su interacción con los [wearables](https://es.wikipedia.org/wiki/Computadora_corporal), a qué temperatura está el ambiente o detectar si hay algún peligro dentro como fuego.[47](https://es.wikipedia.org/wiki/Macrodatos#cite_note-47)​

Deportes[[editar](https://es.wikipedia.org/w/index.php?title=Macrodatos&action=edit&section=12)]

Profesional

En un ámbito donde se mueve tanto dinero, suelen utilizar las nuevas tecnologías antes que los usuarios de base. Nos encontramos por ejemplo que el análisis de los partidos constituye una parte fundamental en el entrenamiento de los profesionales, y la toma de decisiones de los entrenadores.

[Amisco](https://es.wikipedia.org/w/index.php?title=Amisco&action=edit&redlink=1)[48](https://es.wikipedia.org/wiki/Macrodatos#cite_note-Futbol-48)​ es un sistema aplicado por los más importantes equipos de las ligas [Española](https://es.wikipedia.org/wiki/Ligas_de_f%C3%BAtbol_de_Espa%C3%B1a), [Francesa](https://es.wikipedia.org/wiki/Ligue_de_Football_Professionnel), [Alemana](https://es.wikipedia.org/wiki/Bundesliga_(Alemania)) e [Inglesa](https://es.wikipedia.org/wiki/Premier_League) desde el 2001. Consta de 8 cámaras y diversos ordenadores instalados en los estadios, que registran los movimientos de los jugadores a razón de 25 registros por segundo, y luego envían los datos a una central donde hacen un [análisis masivo de los datos](https://es.wikipedia.org/wiki/Miner%C3%ADa_de_datos). La información que se devuelve como resultado incluye una reproducción del partido en dos dimensiones, los datos técnicos y estadísticas, y un resumen de los datos físicos de cada jugador, permitiendo seleccionar varias dimensiones y visualizaciones diferentes de datos.[48](https://es.wikipedia.org/wiki/Macrodatos#cite_note-Futbol-48)​  
Aficionado

Aplicaciones como [Runtastic](https://es.wikipedia.org/wiki/Runtastic" \o "Runtastic), [Garmin](https://es.wikipedia.org/wiki/Garmin) o [Nike](https://es.wikipedia.org/wiki/Nike)+ proveen de resultados big data al usuario. Este último —Nike+— va un paso más allá a nivel de organización, ya que fabrican un producto básico para sus usuarios: las zapatillas. Los 7 millones de usuarios generan una gran cantidad de datos para medir el rendimiento y su mejora, por lo que la empresa genera unos [clústeres](https://es.wikipedia.org/wiki/Cl%C3%BAster_(inform%C3%A1tica)) con los patrones de [comportamiento](https://es.wikipedia.org/wiki/Comportamiento) de sus usuarios. Uno de sus objetivos pues, es controlar el tiempo de vida de sus zapatillas encontrando fórmulas para mejorar la calidad. Por último, aumenta la competitividad entre sus usuarios con el uso de la [ludificación](https://es.wikipedia.org/wiki/Ludificaci%C3%B3n): establece que comunidades de usuarios lleguen a metas y consigan objetivos conjuntamente con el uso de la aplicación, motivando e inspirando a los corredores para usar su aplicación y a más largo plazo, sus productos deportivos.

Investigación[[editar](https://es.wikipedia.org/w/index.php?title=Macrodatos&action=edit&section=13)]

Salud y medicina

Hacia mediados 2009, el mundo experimentó una pandemia de [gripe A](https://es.wikipedia.org/wiki/Gripe_A), llamada gripe porcina o H1N1. La web [Google Flu Trends](https://www.google.org/flutrends/es) fue capaz de predecirla gracias a los resultados de las búsquedas. Flu Trends usa los datos de las búsquedas de los usuarios que contienen síntomas parecidos a la enfermedad de la gripe y los agrupa según ubicación y fecha, y es capaz de predecir la actividad de la gripe hasta con dos semanas de antelación más que los sistemas tradicionales.

Más concretamente en [Nueva Zelanda](https://es.wikipedia.org/wiki/Nueva_Zelanda)[49](https://es.wikipedia.org/wiki/Macrodatos#cite_note-49)​ cruzaron los datos de las tendencias de gripe de Google con datos existentes de los sistemas de salud nacionales, y comprobaron que estaban alineados. Los gráficos mostraron una [correlación](https://es.wikipedia.org/wiki/Correlaci%C3%B3n) con las búsquedas de síntomas relacionados con la gripe y la extensión de la [pandemia](https://es.wikipedia.org/wiki/Pandemia) en el país. Los países con sistemas de predicción poco desarrollados pueden beneficiarse de una predicción fiable y pública para abastecer a su población de las medidas de seguridad oportunas.

Defensa y seguridad

Para incrementar la seguridad frente a los ataques de las propias organizaciones, ya sean empresas en el entorno económico o los propios ministerios de defensa en el entorno de [ciberataques](https://es.wikipedia.org/wiki/Ciberataques), se contempla la utilidad de las tecnologías de big data en escenarios como la vigilancia y seguridad de fronteras, lucha contra el terrorismo y crimen organizado, contra el fraude, planes de seguridad ciudadana o planeamiento táctico de misiones e [inteligencia militar](https://es.wikipedia.org/wiki/Inteligencia_militar).[50](https://es.wikipedia.org/wiki/Macrodatos#cite_note-50)​

Caso específico del proyecto Aloja

El proyecto Aloja[51](https://es.wikipedia.org/wiki/Macrodatos#cite_note-51)​ ha sido iniciado por una apuesta en común del [Barcelona Supercomputing Center](https://es.wikipedia.org/wiki/Barcelona_Supercomputing_Center) (BSC) y [Microsoft](https://es.wikipedia.org/wiki/Microsoft) Research. El objetivo de este proyecto de big data quiere «conseguir una optimización automática en despliegues de [Hadoop](https://es.wikipedia.org/wiki/Hadoop) en diferentes infraestructuras». [40]

Caso específico de sostenibilidad

Conservation International es una organización con el propósito de concienciar a la sociedad de cuidar el entorno de una manera responsable y sostenible. Con la ayuda de la plataforma Vertica Analytics de [HP](https://es.wikipedia.org/wiki/HP), han situado 1000 cámaras a lo largo de dieciséis bosques en cuatro continentes. Estas cámaras incorporan unos [sensores](https://es.wikipedia.org/wiki/Sensor), y a modo de cámara oculta graban el comportamiento de la fauna. Con estas imágenes y los datos de los [sensores](https://es.wikipedia.org/wiki/Sensor) (precipitaciones, temperatura, humedad, solar…) consiguen información sobre cómo el [cambio climático](https://es.wikipedia.org/wiki/Cambio_clim%C3%A1tico) o el desgaste de la tierra afecta en su comportamiento y desarrollo.[52](https://es.wikipedia.org/wiki/Macrodatos#cite_note-52)​

Virtualización de big data[[editar](https://es.wikipedia.org/w/index.php?title=Macrodatos&action=edit&section=14)]

La virtualización de big data es una forma de recopilar información de múltiples fuentes en el mismo lugar. El ensamblaje es virtual: a diferencia de otros métodos, la mayoría de los datos permanecen en su lugar y se toman bajo demanda directamente desde el sistema de origen.[53](https://es.wikipedia.org/wiki/Macrodatos#cite_note-53)​

Wikipedia BigData(EN) Macrodatos(ES)

Metadata is the information that describes other data – ‘data about data’

Yet metadata specifically identifies the attributes, properties and tags that will describe and classify information

It would be more appropriately defined as ‘information about data’. It is represented in the form of any number of characteristics associated with the data information asset such as type of asset, author, date originated, workflow state, and usage within the Enterprise, among numerous others. Though once defined, metadata provides the value and purpose of the data content, and thus becomes an effective tool for quickly locating information – a must for Big Data analytics and business user reporting. But metadata can also identify ‘Little Data’ that ultimately provides structure to what becomes Big Data

Metadata ensures a more accurate picture of data across your enterprise and further ensures this level of data consistency for Big Data analytics and business applications.

Data science central