Definición

El término “big data” representa la gran variedad de datos recopilados en grandes volúmenes y a una velocidad superior.

Es decir, el big data se compone de grupos de datos de gran tamaño y mayor complejidad, especialmente cuando proceden de nuevas fuentes de datos.

Estos conjuntos de datos son tan grandes que un software de procesamiento de datos convencional no es capaz de gestionarlos. Sin embargo, estos datos pueden ser utilizados para resolver problemas que antes no hubiera sido posible solucionar.

Las cinco Vs:

Características de los datos:

Volumen:

La cantidad de datos importa. Con big data, tenemos que procesar grandes volúmenes de datos estructurados, pero, sobre todo, no estructurados.

La procedencia, además, puede ser desconocida, por ejemplo, los datos de las feeds de Twitter.

Para algunas organizaciones, esto puede suponer decenas de terabytes de datos. Para otras, incluso cientos de petabytes. Esto supone un reto para almacenar y analizar los datos.

Velocidad:

Los datos cada vez se generan más rápido, haciendo necesaria una respuesta veloz. Esto requiere tiempos de recopilación, procesamiento y análisis y gestión de la información menores, en algunos casos, incluso instantáneos o prácticamente instantáneos. Estos últimos, requieren una evaluación y actuación instantánea.

Los datos de mayor velocidad se suelen transmitir directamente a la memoria, en vez de escribirse en un disco.

Variedad:

La variedad hace referencia a las diferentes fuentes, herramientas o plataformas desde las que provienen los datos y a los tipos de datos que se nos presentan.

Los datos convencionales eran estructurados y podían organizarse claramente en una base de datos relacional, en cambio, ahora los datos se presentan en varios formatos de datos no estructurados.

Los tipos de datos no estructurados y semiestructurados, como el texto, audio o vídeo, aumentan la complejidad de su almacenamiento, procesamiento y análisis, ya que, requieren de un preprocesamiento adicional para poder obtener significado y habilitar los metadatos.

Es por esto, que es habitual tratar cada tipo de dato de manera distinta mediante herramientas específicas.

Qué hacer con los datos:

Veracidad:

La confianza en los datos debe de ser alta, con resultados verificables y de calidad.

Muchos datos pueden llegar incompletos, estos deben ser limpiados y analizados y, para ello, es necesario invertir tiempo y dinero en la búsqueda de datos de calidad para poder desarrollar y aplicar soluciones y métodos capaces de eliminar la mayor cantidad de errores.

Valor:

Los datos generados, una vez procesados y convertidos en información, deben tener un valor y una utilidad real. Para tener valor, deben ayudarnos a tomar una decisión en base a ellos.

**Funcionamiento**

**Tipos de datos**

**Estructurados:**

**Datos con una longitud y formato determinado.**

**Se almacenan en tablas.**

**Por ejemplo, bases de datos relacionales.**

**No estructurados:**

**Datos con el formato de recolección, no tienen formato específico.**

**No se pueden almacenar en tablas.**

**Por ejemplo, los documentos multimedia.**

**Semiestructurados:**

**Datos de diferentes campos, pero con marcadores para separar sus**  **elementos.**

**Poco regulares, por lo tanto, no se pueden gestionar de forma**  **estándar. Pero, poseen sus propios metadatos.**

**Por ejemplo, los archivos HTML.**

**Obtención**

**Personas: correos electrónicos, mensajes por WhatsApp...**

**Transacciones: llamadas telefónicas, acceso a redes WIFI públicas...**

**Mercadotecnia: interacciones con las páginas web...**

**Interacciones M2M**(Máquina a máquina)**: sensores de temperatura, de luz...**

**Datos Biométricos: huellas dactilares, lectores de cadenas de ADN...**

**Transformación**

**Utilizamos plataformas extraer, transformar y cargar (ETL)** para extraer los datos de los diferentes campos, para después hacer transformaciones (convertir, limpiar, cambiar de formato a los datos, etc.) y cargar los datos en la base de datos especificada.

**Un ejemplo de plataforma ETL es el Pentaho Data Integration, más concretamente su aplicación Spoon.**

**Almacenamiento NoSQL**

Son sistemas que **no cumplen el modelo entidad-relación**. Sus sistemas de almacenamiento **son más flexibles y son más sencillos de consultar**, además **permiten manipular grandes cantidades de información de manera más rápida** que las bases de datos relacionales.

**Tipos:**

**Clave-valor (key-value):** los datos se almacenan de forma **similar a los mapas o diccionarios de datos**, donde se accede al dato **a partir de una clave única**. Los datos son aislados e independientes entre ellos, y no son interpretados por el sistema. Este sistema carece de una estructura de datos clara y establecida.

**Documental:** en este caso **guardamos datos semiestructurados**. Estos pasan a llamarse documentos, y pueden estar formateados en XML, JSON, Binary JSON o el que acepte la misma base de datos.

**En grafo:** se basan en la teoría de grafos, donde **se establece que la información son los nodos** y **las relaciones entre la información son las aristas**. Se utiliza en casos de **relacionar grandes cantidades de datos** que pueden ser **muy variables**.

**Orientado a columnas:** un **mapa de datos multidimensional poco denso, distribuido y persistente**. Se orienta a almacenar datos con tendencia a **escalar horizontalmente**, por lo que permite **guardar diferentes atributos y objetos bajo una misma clave**. Los atributos y los objetos no son interpretados 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.

**Análisis**

**Permite mirar los datos y explicar lo que está pasando.**

**Tipos:**

**Asociación:** encuentra **relaciones entre diferentes variables** pretendiendo **encontrar una predicción en el comportamiento de otras variables**.

**Minería de datos (data mining): su objetivo es encontrar comportamientos predictivos.** Engloba el conjunto de técnicas que **combina métodos estadísticos y de aprendizaje automático con almacenamiento en bases de datos.** Está estrechamente relacionada con los modelos utilizados para descubrir patrones en grandes cantidades de datos.

**Agrupación (clustering):** un **tipo de minería de datos** que **divide grandes grupos de individuos en grupos más pequeños** de los cuales no conocíamos su parecido antes del análisis. El **propósito es encontrar similitudes entre estos grupos, y descubrir nuevos, conociendo cuáles son las cualidades que los 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.

**Análisis de texto (text analytics):** gran parte de los datos generados por las personas son textos, como correos, búsquedas web o contenidos. **Esta metodología permite extraer información de los textos generados por personas y así modelar temas y asuntos o predecir palabras.**

**Aplicaciones**

**Gobierno**

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.

**Desarrollo internacional**

Los avances en el análisis de big data ofrecen oportunidades rentables para mejorar la toma de decisiones en áreas de desarrollo críticas como la atención médica, el empleo... Sin embargo, problemas como la infraestructura tecnológica inadecuada y la escasez de recursos económicos y humanos aumentan las preocupaciones existentes con los grandes datos, como la privacidad, la metodología imperfecta y los problemas de interoperabilidad.

**Industria**

El big data proporciona una infraestructura para la transparencia en la industria manufacturera. Una gran cantidad de datos sensoriales, tales como acústica, vibración, presión, corriente, voltaje y datos de controlador, además de los datos históricos, construyen los grandes datos en la fabricación. Los big data generados actúan como la entrada en herramientas predictivas y estrategias preventivas como Pronósticos y Gestión de Salud.

**Medios**

Esta industria aprovecha a los consumidores utilizando tecnologías que hacen que la información llegue a las personas objetivo en momentos óptimos en ubicaciones óptimas. El objetivo final es servir o transmitir, un mensaje o contenido que (estadísticamente hablando) esté en línea con la mentalidad del consumidor.

**Seguros**

Los proveedores de seguros médicos recopilan datos sobre "determinantes sociales", como el consumo de alimentos y televisión, el estado civil, el tamaño de la vestimenta y los hábitos de compra, desde los cuales hacen predicciones sobre los costos de salud para detectar problemas de salud en sus clientes. Es controvertido si estas predicciones se están utilizando actualmente para fijar precios.

**Deportes**

Sistemas de cámaras y diversos ordenadores instalados en 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. 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.

**Finanzas**

El crecimiento de datos en el mundo financiero obliga al uso del big data para el procesamiento rápido de datos, gestión de la omnicanalidad, segmentación avanzada de clientes, creación de estrategias de precios dinámicos, gestión de riesgos, prevención de fraudes, apoyo en la toma de decisiones, detectar tendencias de consumo, definir nuevas formas de hacer mejor las cosas, detectar alertas y otro tipo de eventos complejos, hacer un seguimiento avanzado de la competencia.

**Mercadotecnia y ventas**

Los macrodatos cada vez se utilizan más para segmentación avanzada de los consumidores, automatizar la personalización de los productos, adaptar las comunicaciones al momento del ciclo de venta, captar nuevas oportunidades de venta, apoyo en la toma de decisiones a tiempo real, gestión de crisis.

**Desafíos**

**El almacenamiento**, aun desarrollándose nuevas tecnologías para el almacenamiento de los datos, el volumen de datos se duplica cada dos años aproximadamente. Las organizaciones continúan intentando mantener el rimo encontrando nuevas formas de almacenar los datos eficazmente.

**El coste** es uno de los principales problemas que supone la manipulación de cantidades tan grandes de datos, muchas veces incluso se recurre a terceros para manipular la información, dado que la organización no se puede hacer cargo sola de todos los datos recopilados.

**La redundancia de la información**, si no se organiza correctamente, los mismos datos se podrían estar manipulando y transformando múltiples veces sin que nos demos cuenta haciendo que perdamos eficiencia y tiempo.

**La selección de los datos**, para que estos sean de algún valor, deben poder utilizarse, y eso depende de su organización y su relevancia. Esto hace que sean significativos en el momento y lugar necesario.

**Conclusión**

**Big Data se refiere a la gran cantidad de datos que recopilan las diferentes** **organizaciones** **día tras** **día.** Pero, en el Big Data lo importante no es únicamente la cantidad de los datos a recoger, sino, también, cómo se organizan esos datos, cómo se analizan y cómo y cuándo se utilizan. Sin estos procesos, entre estas inmensas cantidades de información, no sabríamos distinguir los datos útiles de los datos innecesarios. Y sin esta distinción, no sería posible utilizar los datos adecuadamente.

Webgrafia

<https://es.wikipedia.org/wiki/Macrodatos>

<https://www.computerworld.com/article/2472667/the-government-and-big-data--use--problems-and-potential.html>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1182327/>

<https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2205145>

<https://web.archive.org/web/20181016203319/https://www.datahack.es/cinco-v-big-data/>

<https://www.eldiario.es/tecnologia/diario-turing/big-data_1_5767121.html>

<https://www.oracle.com/es/big-data/what-is-big-data/>