TEMA 1

MÓDULO:
ARQUITECTURA E INFRAESTRUCTURA
BIG DATA

COMPRENDER EL BIG DATA COMO DATA ENGINEER

MARC PLANAGUMÀ I VALLS

Licenciado en Telecomunicaciones
por la UPC-ETSETB.
Data engineer.

Institut de Formació Contínua-IL3
UNIVERSITAT DE BARCELONA

ÍNDICE

OBJETIVOS

- Entender el concepto de *big data* desde la perspectiva de un *data engineer*: su contexto, su enfoque y su ámbito dentro de las responsabilidades de un ingeniero de datos.
- Conocer las características del big data de forma técnica y detallada para entender el origen de los retos que un data engineer debe solucionar.
- Identificar cómo, cuándo y por qué aparecen los síntomas que demandan aplicar soluciones big data en aplicaciones de analítica.
- Comprender los retos a los que nos enfrentamos cuando queremos diseñar arquitectura e infraestructuras para soluciones big data.

1. COMPRENDER EL BIG DATA COMO DATA ENGINEER

1.1. INTRODUCCIÓN

Los sistemas distribuidos como paradigma y las herramientas big data que el mercado *open source* nos ofrece dentro de este paradigma son disciplinas de trabajo comunes para un ingeniero de datos. Pero estos campos específicos de trabajo se encuentran contextualizados por un reto más general que es el concepto de *analítica big data*.

SABÍAS QUE...

Big data, aunque se trata aún de un concepto joven, es ya uno de los campos más influyentes dentro de la digitalización de las empresas.

Curva de tecnologías emergentes.
Fuente: [Digital Transformation of Business and Society](#) (blog de Frank Diana).

A continuación, vamos a definir, comprender y detallar qué aspectos del mundo big data ponen las bases y el contexto del trabajo de un data engineer.

DEFINICIÓN DE BIG DATA

Si empezamos buscando cual es la definición oficial de big data, podemos tomar la del glosario de referencia de conceptos tecnológicos de la revista *Gartner* (2012):

CITA

"Big Data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation." (IT Glossary. Gartner)

Como podemos ver, según la revista *Gartner*, big data es básicamente un concepto económico y de gestión definido como el conjunto de propiedades de **alto volumen, alta velocidad y alta variedad** de los datos, que hacen imprescindible la búsqueda de nuevas formas de procesamiento de la información, eficientes y con un coste aceptable para la mejora de la comprensión de datos, la toma de decisiones y la automatización de procesos.

Si tomamos una definición más focalizada al **ámbito de un data engineer**, básicamente puede explicarse como un gran volumen de datos que no se pueden almacenar y procesar utilizando un enfoque tradicional. Dado que estos datos pueden contener información valiosa, deben procesarse en un periodo de tiempo dentro del cual el valor esté aún vigente. En muchos casos, estas ventanas de tiempo son muy cortas tanto por motivos de competitividad como de competencia o de oportunidad. Si usáramos un enfoque tradicional, no podríamos realizar esta tarea dentro de un corto espacio de tiempo, ya que la capacidad de almacenamiento y procesamiento no sería suficiente.

IMPORTANTE

Una vez conocida la definición teórica, podemos extraer la conclusión de que tenemos un problema big data cuando el volumen, la velocidad y/o la variedad de nuestros datos hace imposible tratarlos con las tecnologías IT (*software* y *hardware*) comunes.

1.2. COMPRENDER EL BIG DATA

Más allá de la definición, necesitamos comprender cómo la problemática big data define el ámbito de trabajo de un ingeniero de datos.

RECUERDA

Big data es un campo dedicado al análisis, procesamiento y almacenamiento de grandes colecciones de datos que, con frecuencia, proceden de fuentes dispares.

1.2.1. ¿CUÁL ES SU CONTEXTO?

Como hemos visto, la digitalización extendida y creciente de las actividades productivas y organizativas nos sitúan en el contexto de trabajo big data.

Aunque el big data puede aparecer como una nueva disciplina, su cometido se ha desarrollado durante años. La gestión y el análisis de grandes conjuntos de información suponen un problema desde hace mucho tiempo. La necesidad de reinventar, reenfocar y replantear cómo obtener valor de los datos ha estado presente varias veces a lo largo de la historia. Las soluciones de referencia de gestión documental (no digital) han evolucionado enormemente desde la aparición de la escritura hasta llegar a complejos sistemas de gestión documental e incluso a una disciplina propia: la biblioteconomía.

El término *big data* no aparece hasta que los sistemas de gestión de datos digitales sufren su primera gran crisis, que provoca que todas las tecnologías y soluciones conocidas sean superadas por el volumen de datos y se abra la búsqueda e investigación de todo un abanico nuevo de tecnologías, paradigmas y soluciones.

Como hemos comentado, las soluciones y prácticas big data se requieren cuando las tecnologías y técnicas de análisis, procesamiento y almacenamiento de datos tradicionales son insuficientes. En concreto, big data aborda retos distintos, como la combinación de varios conjuntos de datos (*data sets*) no relacionados, el procesamiento de grandes cantidades de datos no estructurados y la recopilación de información oculta de forma eficiente en el tiempo.

1.2.2. ¿CUÁL ES SU ENFOQUE?

Además de los enfoques analíticos tradicionales basados en estadísticas, el big data agrega nuevas técnicas que aprovechan los **recursos y enfoques computacionales** para ejecutar algoritmos analíticos. Este cambio permite que las soluciones sigan siendo óptimas a medida que los data sets van siendo más grandes, más diversos y más complejos.

IMPORTANTE

Mientras que, desde tiempos bíblicos, los enfoques estadísticos se han utilizado para aproximar las medidas de una población a través del muestreo, los avances en la ciencia computacional han permitido el procesamiento de conjuntos de datos enteros, lo que hace innecesario ese muestreo.

1.2.3. ¿CUÁL ES SU ÁMBITO?

En el análisis de los data sets, el big data es un esfuerzo **interdisciplinario** que combina matemáticas, estadística, informática y la experiencia en la materia o su dominio.

Esta mezcla de habilidades y perspectivas ha provocado cierta confusión en cuanto a lo que comprende el campo big data y su análisis, de modo que la respuesta a esta pregunta dependerá del punto de vista de quien la responda.

Los límites de lo que constituye un problema big data también se están viendo modificados como consecuencia del panorama siempre cambiante y avanzado de la tecnología del software y el hardware. Esto se debe al hecho de que la definición de big data tiene en cuenta el impacto de las características de los datos en el diseño del propio entorno de la solución.

EJEMPLO

Hace treinta años, un *gigabyte* de datos podía equivaler a un problema big data y requería recursos informáticos de propósito especial. En la actualidad, los gigabytes de datos son comunes y se pueden transmitir, procesar y almacenar fácilmente en dispositivos orientados al consumidor.

1.2.4. ¿CUÁL ES SU APLICACIÓN?

Los datos procesados por una solución big data pueden ser utilizados directamente por las **aplicaciones** o por **sistemas de datos** para enriquecer los datos existentes.

Los resultados obtenidos a través del procesamiento big data pueden llevar a una amplia gama de **perspectivas y beneficios**, tales como:

- Optimización operativa.
- Inteligencia accionable.
- Identificación de patrones.
- Predicciones precisas.
- Detección de fallos y fraudes.
- Registros más detallados.
- Mejor toma de decisiones.
- Descubrimientos científicos.

Sin embargo, también es preciso considerar la aparición de numerosos problemas que deberán tenerse en cuenta al adoptar enfoques de análisis big data. Estas cuestiones deben entenderse y ponderarse contra los beneficios previstos para poder tomar decisiones y llevar a cabo planes informados. Más adelante veremos algunos de estos problemas y retos por superar.

La gestión de datos es cada vez más compleja. El big data está en todas partes, en la mente de todos, y en muchas formas diferentes: publicidad, gráficos sociales, noticias, recomendaciones, *marketing*, salud, seguridad, gobierno, etc. En los últimos cinco años, han surgido miles de tecnologías que tienen que ver con la adquisición, administración y análisis de big data.

IMPORTANTE

Los equipos de IT han llevado a cabo la ardua tarea de deber elegir entre una enorme y creciente oferta tecnológica, sin tener referencias suficientes ni metodologías estandarizadas para manejar tal elección. La tarea del diseño de soluciones big data es una disciplina más cercana a la investigación que a la implantación de soluciones maduras.

PIENSA UN MINUTO

Cuando te encuentras ante tal decisión de diseño, te puedes preguntar:

- ¿Cuándo debo pensar en emplear big data para mi sistema de IT?
- ¿Estoy listo para emplearlo?
- ¿Con qué empiezo?
- ¿Debería realmente ir a por ello a pesar de la sensación de que big data es solo una tendencia de marketing?

Todas estas preguntas están en la mente de la mayoría de los directores de información (CIO), directores de estrategia de datos (CDO) y jefes de tecnología (CTO) que cubren de forma global la toma de decisiones por las que una compañía decide usar o no soluciones big data. Para responder a tales preguntas, es necesario el análisis y el veredicto de equipos de ingenieros de datos.

1.3. CARACTERÍSTICAS DEL BIG DATA

Para identificar un problema big data y resolverlo implementando soluciones como data engineer, primero debemos conocer las distintas características del big data.

IMPORTANTE

Para que un conjunto de datos se considere big data, debe poseer una o más características que requieran el diseño de una solución y/o arquitectura avanzada que permita que estos datos puedan analizarse.

La mayoría de estas características fueron identificadas inicialmente por Doug Laney a principios de 2001, cuando publicó un artículo que describía el impacto del **volumen**, la **velocidad** y la **variedad** de los datos del comercio electrónico en los almacenes de datos empresariales.

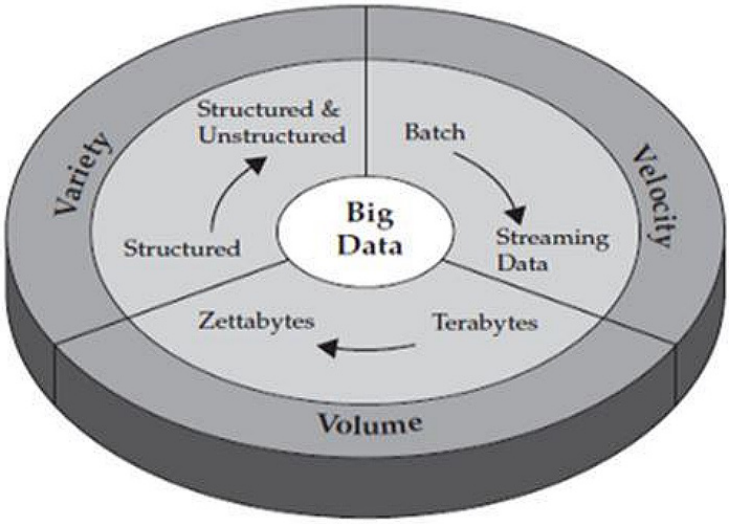
Estas características se conocen como las **3 *V***. Seguidamente explicaremos cómo se pueden utilizar para diferenciar los datos clasificados como big data de otras formas de datos.

PARA SABER MÁS

Puedes leer el artículo [3D Data Management: Controlling Data Volume, Velocity, and Variety] de Doug Leany.

A esta lista, se han agregado otras características que empiezan por *V*, como veracidad, viabilidad, valencia o valor, que también veremos que aportan matices a la definición original.

1.3.1. LAS 3 V DEL BIG DATA



Interrelación de las 3 V del big data.
Fuente: [Big Data: A Foundational Explanation].

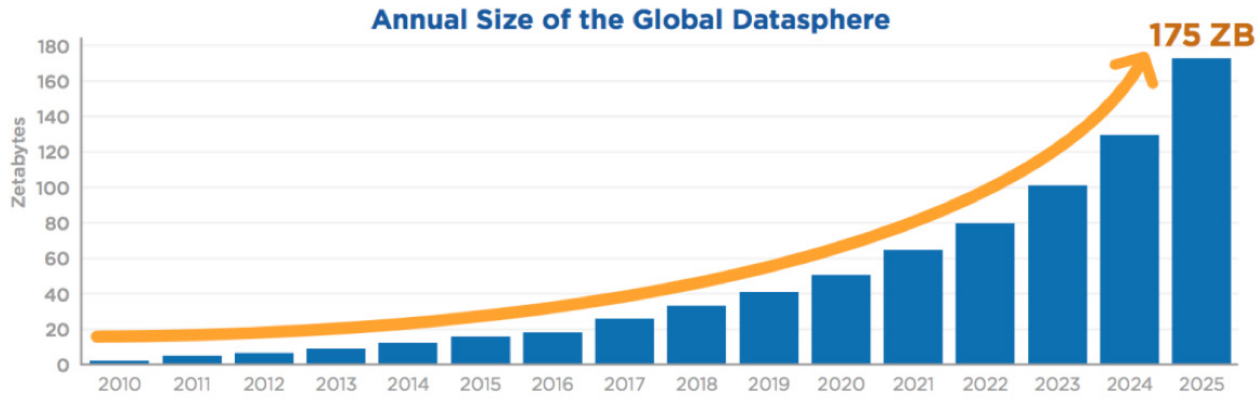
VOLUMEN

El volumen se refiere al **tamaño del conjunto de datos** que hay que manejar. Dicho tamaño adquiere un papel muy relevante en la determinación del valor de los datos, y también es un factor clave que define si podemos juzgar el fragmento como grande. Por lo tanto, el volumen justifica uno de los atributos más importantes del big data.

Se trata, probablemente, de la principal característica con la que todo el mundo asocia al big data.

SABÍAS QUE...

Según un *white paper* de la consultora IDC y la empresa de fabricación de discos duros Seagate, en 2018, la cantidad global de datos digitales en el mundo se estimó en 33 *zettabytes*. La predicción para el 2025 es de 175 zettabytes, y que casi el 30 % de los datos del mundo necesitarán procesamiento en tiempo real.



Comparativa del crecimiento de los datos desde 2010 y perspectivas de futuro.
Fuente: Reinsel, Gantz y Rydning (2018).

El beneficio obtenido de la capacidad de procesar grandes cantidades de información es la principal atracción del análisis big data. Tener **más datos** significa tener **mejores modelos**: las soluciones matemáticas sobre los datos pueden ser irrazonablemente más efectivas en caso de grandes cantidades de datos.

EJEMPLO

Si podemos ejecutar una previsión teniendo en cuenta 300 factores en lugar de seis, la calidad de esa previsión puede mejorar órdenes de magnitud.

El mundo de la estadística ha proporcionado históricamente estudios y conclusiones a partir de **muestras** de datos del mundo real donde, a través de distintas técnicas matemáticas, extrapola un resultado estimado a escala global. El **muestreo**, pues, ha sido una herramienta muy poderosa para hacer estudios de comportamiento real que permite reducir mucho los datos empleados y, de este modo, facilitar aún más el trabajo de cálculo. Pero todo muestreo, así como otras técnicas de reducción de datos, **añade un margen de error** al resultado final. Un ejemplo de ello serían las encuestas electorales comparadas con los resultados finales de la votación, que se basan en la totalidad del universo de los datos.

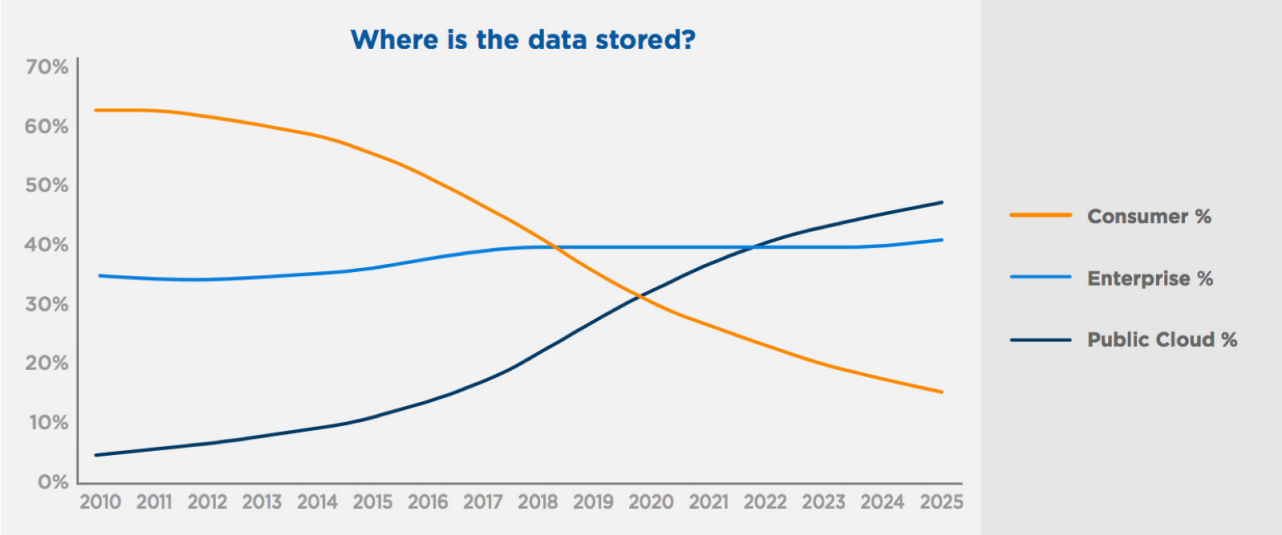
Así pues, de todos es sabido que, cuanto más grande sea la muestra de datos reales, menor error tendremos en el cálculo, hasta el punto de que si somos capaces de trabajar con la totalidad de los datos, ese error será cero. La ambición de querer utilizar la máxima cantidad de datos posibles para reducir el error y obtener más valor en el análisis es el principal motivo por el que nos enfrentamos a estos crecientes volúmenes de datos.

RECUERDA

El volumen presenta el desafío más inmediato a las estructuras de IT convencionales. Un alto volumen requiere un almacenamiento escalable y un enfoque distribuido para su consulta y proceso.

Muchas empresas ya tienen grandes cantidades de datos archivados, tal vez en forma de registros, pero no la capacidad de procesarlos.

En la imagen siguiente podemos observar dónde se encuentran los datos digitales almacenados. Vemos que la tendencia es el crecimiento de los **servicios de *cloud* públicos** frente al almacenamiento por parte del consumidor final. Mientras que el volumen de dato digital en *data centers* privados de empresas y organizaciones se mantiene constante en proporción.



Localización de los datos digitales.
Fuente: Reinsel, Gantz y Rydning (2018).

VELOCIDAD

La velocidad se refiere, no solo a la alta frecuencia con la que se generarán nuevos datos, sino también a la necesidad de **responder** a la información **en tiempo real**.

EJEMPLO

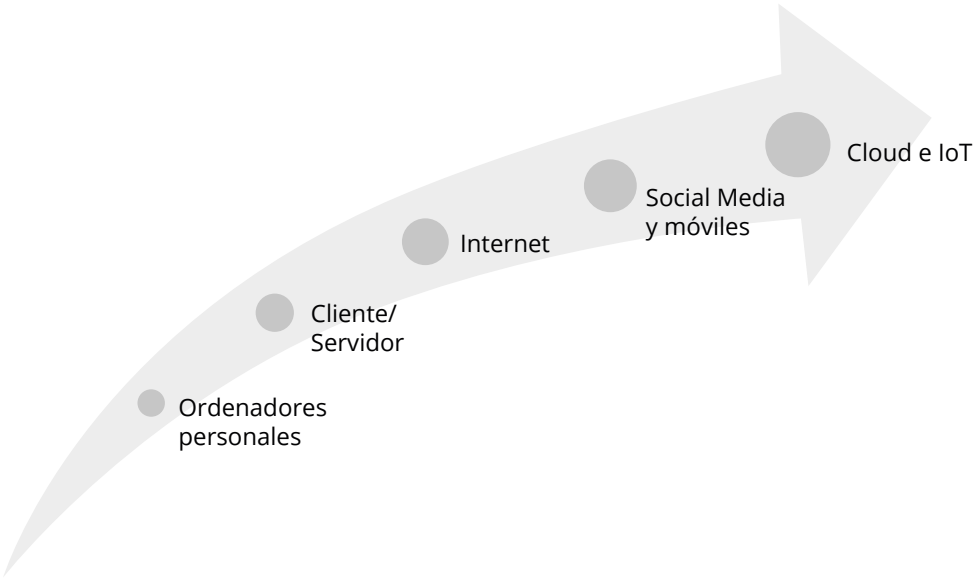
Veamos unos ejemplos relacionados con la velocidad:

- La bolsa de valores de Nueva York captura 1 TB de datos durante cada sesión de *trading*.
- En 2018 se generaron, cada minuto, unos 2000 TB en:
  - 120 horas de vídeos en YouTube.
  - 200 millones de correos electrónicos.
  - 700.000 publicaciones de redes sociales.
  - 11 millones de mensajes.
- El colisionador de partículas del CERN genera un *petabyte* por segundo.

La importancia de la velocidad de los datos (el aumento de la velocidad a la que fluyen los datos en una organización) ha seguido un patrón similar al del volumen.

Los problemas, previamente restringidos a los segmentos mayoristas de la industria, ahora están presentes en un entorno mucho más amplio. ¿Por qué es así? Internet y la era móvil implican que la manera en que entregamos y consumimos productos y servicios se instrumenta y digitaliza cada vez más de forma nativa, generando un flujo de datos que puede ser capturado y utilizado de forma sencilla por el proveedor. También la aparición de los clouds públicos, como veremos más adelante, ha reducido mucho los costes de transporte y almacenaje de datos, y ha generado un amplio mercado de datos disponibles para comprar y vender.

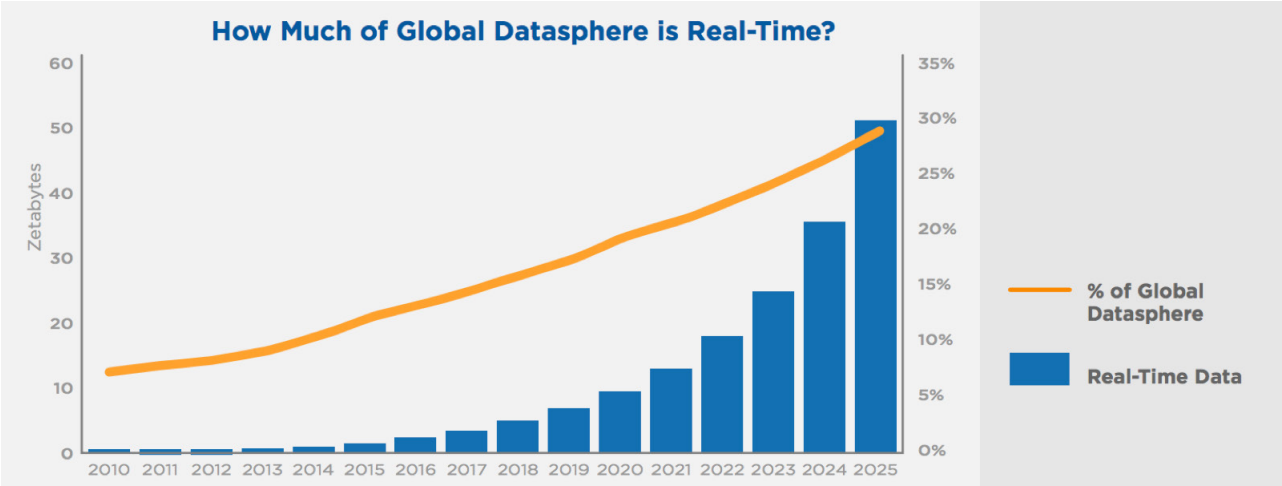
Asimismo, las empresas minoristas son capaces de recopilar grandes cantidades de datos y utilizar rápidamente esa información para obtener una ventaja competitiva.



Incremento de la velocidad de los datos respecto a la evolución tecnológica digital.

IMPORTANTE

El problema no es solo la velocidad de los datos entrantes, sino que también cobra mucha importancia la velocidad desde la entrada de los datos hasta su salida (con la acción del consumidor). Cuanto más rápido sea el bucle de retroalimentación, mayor será la ventaja competitiva.



Proporción de datos en tiempo real.
Fuente: Reinsel, Gantz y Rydning (2018).

La **velocidad en los datos** se mide con el concepto de **latencia**, que establece la diferencia de tiempo entre la entrada y la salida a un sistema. Cuanto más baja es la latencia, menos tiempo tenemos entre la recepción del dato y la toma de la decisión o acción. Cuando hablamos de aplicaciones en tiempo real, lo estamos haciendo de latencias muy bajas en las que la percepción para el usuario es de servicio instantáneo.

SABÍAS QUE...

Según el estudio de Miller (1968) Response time in man-computer conversational transactions:

- Una respuesta de 100 ms es percibida como instantánea.
- Una respuesta menor a 1 segundo es percibida como interactiva.
- Las respuestas de más de 10 segundos pierden la atención del usuario.

Según la latencia permitida, se establecen principalmente las **tipologías de análisis** que se pueden utilizar:

- **Para alta latencia se usa el procesamiento por lotes (*batch*)**: es una forma de procesamiento pensado para el tratamiento de volúmenes de datos históricos almacenados. Su resultado se refiere a los datos que han entrado en esa ejecución y es conocido como *ventana de procesamiento*. Si se quieren procesar nuevos datos, es necesario ejecutar otro procesamiento batch con la nueva entrada. Ese tipo de procesamiento se usa habitualmente en ejecuciones periódicas para ir actualizando los resultados de forma cíclica. Por ejemplo, la generación automática de un informe teniendo en cuenta los datos acumulados durante el día.

- **Para latencia media se usa el procesamiento interactivo**: se refiere al procesamiento derivado de la intervención humana en los datos con finalidad consultiva o exploratoria. Mayoritariamente, esta interacción se consigue con *queries* que definen la consulta u operación del usuario en una instrucción en texto para que el sistema de información pueda ejecutar el procesamiento deseado y devolver el resultado. El más conocido es el lenguaje SQL, aunque existen muchos más. La interacción puede ser directa o indirecta, y realizarse a través de distintas interfaces de interacción con los sistemas de información, como línea de comandos, interfaz gráfica de usuario o API.

- **Para baja latencia se usa el procesamiento en tiempo real (*streaming*)**: se trata del procesamiento continuo de datos de forma reactiva a su llegada con una garantía de respuesta dentro de unos estrictos límites de tiempo. En el mundo big data, la baja latencia es de vital importancia y hace referencia al hecho de que el retardo se reduce a un mínimo para lograr un servicio instantáneo, también llamado *en tiempo real*. *Streaming* es el término usado en la industria para nombrar este flujo de datos continuo, aunque también es conocido como *complex event processing*.

SABÍAS QUE...

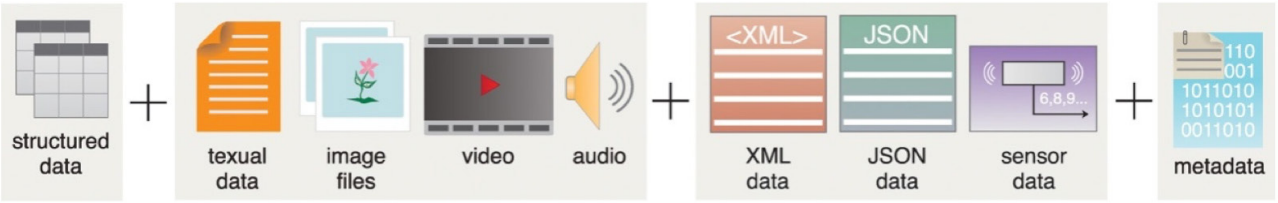
Los coches actuales superan ya los cien sensores para monitorear en tiempo real distintos elementos como la presión de combustible y neumáticos, el control de tracción, velocidad y frenado, o la detección de luz, lluvia y obstáculos circundantes para la toma de decisiones en tiempo real. En el caso de los coches con conducción autónoma, el número de sensores se dispara a más de 10.000.

PIENSA UN MINUTO

Otra dimensión de la velocidad es el periodo de tiempo durante el cual los datos tendrán sentido y serán valiosos. Los datos de entrada que utilizas, ¿envejecen y pierden valor con el tiempo, o son permanentemente valiosos? Este análisis también es muy importante porque si los datos envejecen y pierden valor con el tiempo, entonces, es posible que, debido a una entrada de datos tardía, tus conclusiones sean erróneas.

VARIEDAD

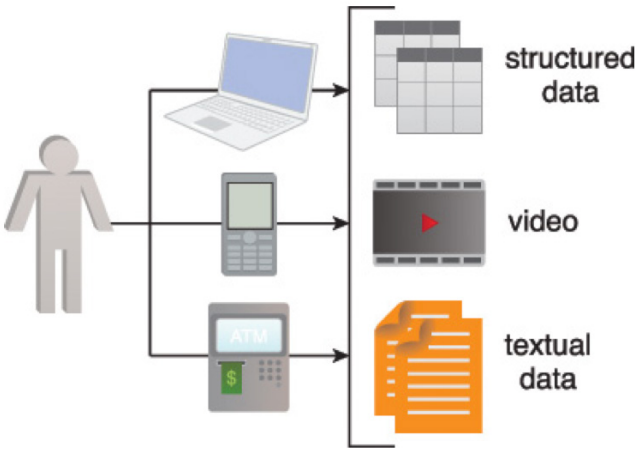
La variedad se refiere a la diversa naturaleza de la información que se tiene que manejar. Rara vez los datos se presentan en una forma perfectamente ordenada y lista para su procesamiento. Así pues, un aspecto común en los sistemas big bata es que los datos sean de orígenes diversos y no cumplan con una estructura ordenada, conocida y preparada para la integración en una aplicación.



Tipologías de datos digitales.
Fuente: Buhler, Khattak y Erl (2016).

Los datos digitales pueden ser generados por el ser humano o por máquinas.

- Los **datos generados por los seres humanos** son el resultado de la interacción humana con sistemas, como los servicios en línea o los dispositivos digitales. Ejemplos de estos datos incluyen: redes sociales, publicaciones de blogs, correos electrónicos, intercambio de fotos y mensajería.

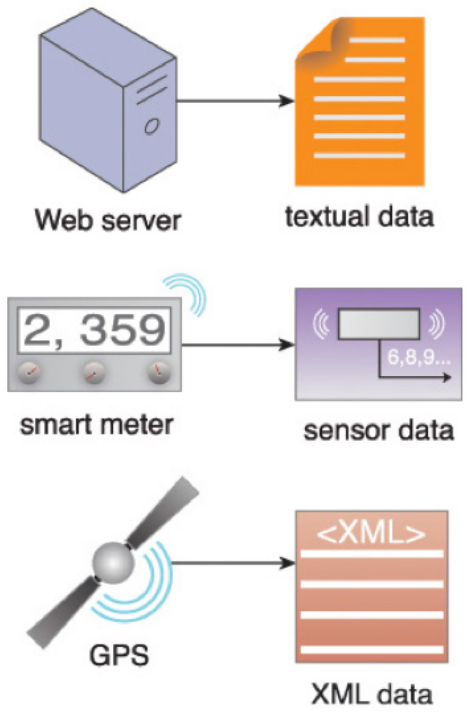


Tipologías de datos generados por humanos.
Fuente: Buhler, Khattak y Erl (2016).

- Los **datos generados por máquinas** provienen de programas de software y dispositivos de hardware en respuesta a eventos del mundo real. Ejemplos de este tipo de datos incluyen: registros web, datos de sensores, datos de telemetría y trazas de uso de dispositivos (o *logs*).

EJEMPLO

Un archivo de registro que captura una decisión de autorización realizada por un servicio de seguridad, un sistema de punto de venta que genera una transacción contra el inventario para reflejar los artículos comprados por un cliente o los datos que se transmiten desde los sensores de un teléfono móvil reportando información, incluidas la posición, la torre y la intensidad de la señal.



Tipologías de datos generados por máquinas. Fuente: Buhler, Khattak y Erl (2016).

Podemos ver que, los datos generados tanto por el hombre como por la máquina pueden provenir de una gran variedad de fuentes y estar representados en diversos formatos o clases. Sin embargo, podemos clasificarlos en **3 tipos**:

- Datos **estructurados**.
- Datos **no estructurados**.
- Datos **semiestructurados**.

Estas tipologías se refieren a la organización interna de los datos y comúnmente se las denomina *formatos de los datos*. Debemos también recordar otro tipo importante y especial de datos en entornos big data: los **metadatos**.

Datos estructurados

Los datos estructurados se ajustan a un modelo o esquema y a menudo se almacenan en **forma tabular**. Se utilizan para capturar relaciones entre distintas entidades y, por lo tanto, se almacenan con mayor frecuencia en una base de datos relacional. Son datos que han definido de forma conocida su longitud y su estructura detallando el significado, el tipo y la relación entre ellos.

Los datos estructurados son generados, con frecuencia, por aplicaciones empresariales y sistemas de información que los almacenan en bases de datos relacionales. Estos sistemas se conocen como ***relational data base management systems*** (**RDBMS**). Se trata de soluciones diseñadas en los años 70 y que, desde los años 80, han sido la opción más empleada para la persistencia, organización y consulta de datos para aplicaciones digitales. Se componen de **tablas y relaciones entre ellas** para organizar los datos de forma **relacional** utilizando claves e índices.

También se caracterizan por su **lenguaje estándar** de uso, el **SQL**, que permite realizar todas las operaciones de datos sobre el sistema, desde la ingestión y la modificación hasta la lectura y el borrado.

Esta tecnología clásica tiene como requerimiento la estructura y organización de los datos. Dado que la estructuración comporta una reducción en la redundancia de los datos, el tamaño de los que hay que almacenar no llega a los volúmenes característicos de big data pero, aun así, esta tecnología se emplea mucho en proyectos big data, siempre combinada con otras tecnologías no relacionales.

Debido a la abundancia de herramientas y bases de datos que apoyan de forma nativa la información estructurada, disponemos de una gran cantidad de fuentes de este tipo que rara vez requieren una consideración especial en lo que respecta a procesamiento o almacenamiento.

Ejemplos de este tipo de datos incluyen:

- Datos de todo tipo de transacciones: financieras, *retail*, logística, *booking*, etc.
- Datos generados en sistemas de gestión de información en organizaciones, como CRM y ERP.
- Datos de gestión gubernamental, como censos, catastros, etc.
- Datos generados en sistemas de gestión de contenido, como CMS.
- Otros datos de aplicaciones que usen bases de datos relacionales como motor de persistencia de datos.

Datos no estructurados

Los datos que no se ajustan a un modelo o esquema de datos se conocen como *datos no estructurados*.

SABÍAS QUE...

Se estima que los datos no estructurados conforman el 80 % de los existentes en una empresa determinada y tienen una tasa de crecimiento más rápida que los datos estructurados.

Esta forma de datos puede ser **textual o binaria** y, a menudo, se transmite a través de archivos que son autónomos y **no relacionales**.

Un archivo de texto puede contener cualquier contenido y ser estructurado o no. En el caso de los archivos de texto no estructurados, estos se componen de un texto cuya estructura no es comprensible de forma directa por un sistema de información, como una tabla o una lista, sino que se compone de texto natural comprensible solo por los humanos.

SABÍAS QUE...

Para extraer la información de texto natural existen disciplinas como el *natural language processing* (NLP) que extraen del lenguaje humano conceptos semánticos interpretables por sistemas digitales para conseguir interacción hombre-máquina con lenguaje humano.

Los archivos binarios suelen ser archivos multimedia que contienen datos de imagen, audio o vídeo. Esos datos, aunque pueden ser reproducidos por sistemas digitales, incorporan información que no es comprensible para los sistemas de procesamiento de información.

SABÍAS QUE...

Las máquinas no saben interpretar los contenidos de ficheros de audio y vídeo. Mayoritariamente, los servicios digitales de reproducción de multimedia se basan en metadatos para conocer el contenido de un documento, ya que no son capaces de interpretar ni las imágenes ni el vídeo para entender los conceptos que aparecen. Existen disciplinas de análisis de imagen y audio para extraer la información que contienen, como es la basada en el campo *deep learning* de la inteligencia artificial.

Técnicamente, tanto los archivos de texto como los binarios tienen una estructura definida por el propio formato de archivo, pero esta estructura, en muchos casos, es desconocida por el consumidor por falta de estandarización. La noción de ser no estructurado se debe al desconocimiento del formato de los datos contenidos en el propio archivo.

Los archivos de **vídeo, imagen y audio** son todos tipos de datos no estructurados.

Normalmente, se requiere una lógica de propósito especial para procesar y almacenar datos no estructurados. Por ejemplo, para reproducir un archivo de vídeo es esencial que el códec correcto (codificador-decodificador) esté disponible.

Los datos no estructurados no se pueden procesar ni consultar directamente mediante SQL, como los datos estructurados. Si se requiere que se almacenen dentro de una base de datos relacional, se almacenan en una tabla como un **objeto binario grande** (**BLOB**).

Para la extracción de información de estos datos son necesarias técnicas avanzadas de analítica en distintos campos que, dependiendo de la naturaleza del dato, se basan principalmente en la **inteligencia artificial**, puesto que deben ser interpretados mayormente como si de un humano se tratase.

Datos semiestructurados

Los datos semiestructurados tienen un nivel definido de estructura y consistencia, pero no son relacionales en su naturaleza. Son jerárquicos o basados en gráficos. Este tipo de datos se almacenan comúnmente en archivos que contienen texto y disponen de estructura.

IMPORTANTE

Su diferencia principal con respecto a los datos estructurados es que su estructura puede variar y que no disponen de relaciones entre otras estructuras.

Unas de las formas más comunes de datos semiestructurados son **XML y JSON**, usadas sobre todo para la transmisión de datos entre distintos sistemas. Debido a la naturaleza textual de estos datos y a su conformidad con algún nivel de estructura, se procesan más fácilmente que los datos no estructurados.

Los datos XML, JSON y de sensores son semiestructurados.

Los datos semiestructurados a menudo llevan un preprocesamiento especial y requisitos de almacenamiento distintos a los relacionales, especialmente si el formato subyacente no está basado en texto.

Como alternativa a las bases de datos SQL, se tratan mayormente usando **bases de datos NoSQL**, que son soluciones específicas para datos con estructuras variables en el tiempo. Las bases de datos de NoSQL semiestructuradas satisfacen esta necesidad de flexibilidad: proporcionan una estructura suficiente para organizar la información, pero no requieren el esquema exacto de los datos antes de almacenarla.

Metadatos

Los metadatos proporcionan información sobre las características y la estructura de un **data set**. Este tipo de datos se genera principalmente en la máquina y se puede anexar a los datos.

IMPORTANTE

El seguimiento de los metadatos es crucial para el procesamiento, almacenamiento y análisis big data, ya que proporciona información sobre el contenido de los datos y su procedencia.

Las **soluciones big data** se basan en metadatos, especialmente cuando se procesan datos semiestructurados y no estructurados.

Ahora que ya conocemos las distintas variedades de datos, podemos concluir que un uso común del procesamiento big data es tomar datos no estructurados y semiestructurados para extraer su significado ordenado, que será consumido por los seres humanos o servirá de entrada estructurada a una aplicación. Una vez que estos datos hayan sido estructurados, podrán relacionarse también con entradas nativamente estructuradas.

El proceso de pasar de los datos de origen a los datos de aplicación procesados implica siempre una pérdida de información.

RECUERDA

Cuando ordenas, siempre terminas tirando cosas. Esto pone el énfasis en uno de los principios del big data: siempre que puedas, mantén todos los datos posibles. Puede haber información útil en los bits que tiras. Si pierdes los datos de origen, no hay vuelta atrás.

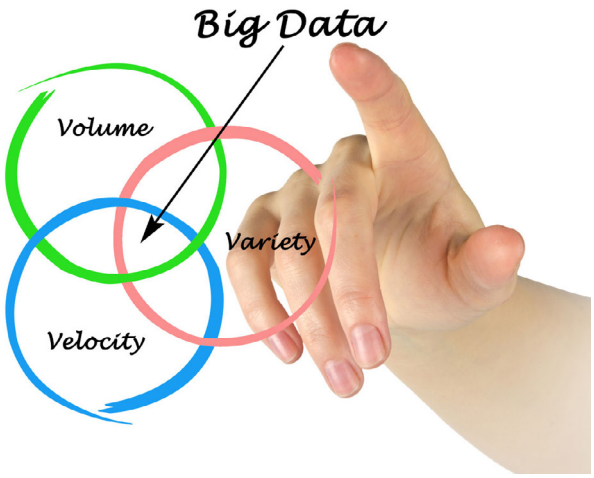
VALOR Y OTRAS V

Quizás por efecto del marketing, con el tiempo han ido apareciendo otras características que empiezan por *V* para describir el big data en un intento de darle una definición más completa a este concepto complejo.

Estas *V* adicionales son, entre otras: valor, validez, veracidad, viabilidad, valencia, viralidad, viscosidad, vocabulario y visualización.

PARA SABER MÁS

Te recomendamos consultar este artículo en el que se llegan a explicar 42 *V* del big data: The 42 V's of Big Data and Data Science.



En lo que se refiere a las características que convierten los datos en big data, solo las tres especificadas anteriormente (volumen, variedad y velocidad) tienen esa capacidad. Las otras *V* detallan otras características deseadas en los datos que ayudan en distintos ámbitos al procesamiento o a la analítica, pero su aparición o incremento no provoca que los sistemas comunes de procesamiento de datos sean inviables, como sí ocurre en escenarios big data.

En el caso del **valor**, sí que encontramos esta característica en la definición del concepto big data, aunque no referida a los datos que deben tratarse sino al resultado que, en cualquier caso, debe aportar valor para ser válido.

RECUERDA

Big data es el conjunto de propiedades de alto volumen, alta velocidad y/o alta variedad de los datos que hacen imprescindible la búsqueda de nuevas formas de procesamiento de la información, eficientes en coste y tiempo para poder extraer valor de ellos.

1.4. IDENTIFICACIÓN DE SÍNTOMAS BIG DATA

Una vez conocidas las características de los datos big data, podemos optar por iniciar un proyecto big data en función de las diferentes necesidades y debido a:

- El volumen de datos que se maneja.
- La variedad de estructuras de datos que tiene el sistema.
- Los problemas de escalabilidad que se experimentan para cumplir con los tiempos requeridos.
- Se desea reducir el coste del procesamiento de los datos.

A continuación, veremos qué síntomas pueden hacer que un equipo advierta que necesita iniciar un proyecto de big data.

1.4.1. EL TAMAÑO IMPORTA

El primer síntoma que hace que surja la idea inicial de pensar en big data es empezar a tener problemas relacionados con el **tamaño y volumen** de los datos. Aunque, en la mayoría de las ocasiones, esto presenta verdaderas y legítimas razones para pensar en big data, hoy en día, no es el único motivo para ir por este camino.

Los problemas por volumen son fáciles de detectar pero, contrariamente a lo que se piensa, no son una razón evidente para optar por estrategias big data, sino que son solo un indicio.

Las soluciones clásicas pueden procesar sin problemas volúmenes muy grandes de datos organizados y ordenados. Por ejemplo: los censos, los registros de clientes o los catálogos de productos de tamaños inmensos son procesados por sistemas de información clásicos gracias a la naturaleza relacional de los datos.

IMPORTANTE

Por lo general, el desorden del dato más el volumen es justo la combinación letal que nos empuja al diseño de soluciones big data.

1.4.2. EL ORDEN IMPORTA

La **variedad** es la característica que agrava más los problemas en los sistemas de información clásicos, especialmente los basados en soluciones de bases de datos relacionales basadas en SQL.

PIENSA UN MINUTO

Los RDBMS no están preparados para afrontar la diversidad de tipos de datos. ¿Cómo se las puede arreglar uno para integrar distintos tipos de datos cuando los almacenes de datos tradicionales (SQL Database) esperan que se realice una estructuración en forma de tablas y relaciones?

Esto no es factible sin agregar una **tecnología flexible** y de **estructura dinámica** (*schemaless*) que maneje nuevas estructuras de datos a medida que llegan. Cuando hablamos de distintos tipos de datos, debemos imaginar datos no estructurados o semiestructurados, como XML, JSON, texto, gráficos, imágenes, vídeos, voces, etc.

PARA SABER MÁS

Sobre NoSQL, te recomendamos la consulta de la web NOSQL Databases o el libro *Next Generation Databases: NoSQL, NewSQL, and Big Data* (Harrison, 2016).

1.4.3. EL VALOR IMPORTA

Sí, es bueno almacenar datos no estructurados, pero es mejor si podemos sacar algo de ellos. Otro síntoma se puede deducir de esta premisa.

IMPORTANTE

El big data trata de extraer información de valor añadido de un gran volumen de datos, y este valor viene determinado sobre todo por dos factores:

- La **rentabilidad** entre el coste del proceso y el beneficio del resultado.
- El **tiempo de validez del resultado** que, en los casos más estratégicos, tiende a ser muy corto.

¿Cómo podemos extraer valor añadido de los datos de forma rápida y eficiente?

Para responder a esta pregunta, pensemos de nuevo en una base de datos tradicional en la que se crean índices en diferentes columnas para acelerar la consulta de las búsquedas.

¿Qué ocurre si queremos indexar las cien columnas porque necesitamos poder ejecutar consultas complejas que implican un número no determinista de columnas clave?

La rigidez de las soluciones SQL, que en su día aportaron estandarización y orden a los sistemas de información, en la actualidad son el principal problema para la búsqueda de soluciones ágiles y flexibles.

1.4.4. EL TIEMPO IMPORTA

Tradicionalmente, en los sistemas de información, había más transacciones de lectura que de escritura en las que técnicas de cachés comunes y bases de datos eran suficientes para satisfacer la alta demanda de lectura. Respecto a la escritura, con herramientas y técnicas ETL (*exctract*, *transform*, *load*) era suficiente para diseñar y ejecutar trabajos de llenado periódico. Esta es una estrategia clásicamente usada y aún válida para muchos escenarios, pero ya no es la tendencia para obtener valor diferencial.

Ahora se necesita una solución que sea capaz de controlar los datos a medida que llegan, un procesamiento continuo que dé **resultados** casi en **tiempo real**, porque en la velocidad está actualmente el valor deferencial respecto a la competencia. Ya no es suficiente con llegar a un resultado, sino que debes hacerlo lo antes posible para maximizar el valor.

1.4.5. EL COSTE IMPORTA

Además de las consideraciones técnicas y de arquitectura, es posible que te enfrentes a casos de uso en los que podemos definir big data de forma nativa. La mayoría de ellos están vinculados a la industria digital y los servicios en Internet. En estas industrias se suele seguir un modelo de negocio exclusivamente basado en la **explotación y monetización** de los datos. Por ello, persiguen analizar todos los datos que sea factible, de todas las fuentes disponibles y de la forma más rápida posible. Este escenario, desde el primer momento, contiene las 3 *V* combinadas en su máximo exponente.

Todo ello, junto a la ambición de sacar el máximo beneficio de los datos, ha llevado a las industrias a impulsar, diseñar y desarrollar herramientas y estrategias big data, puesto que el coste del procesamiento mediante tecnologías clásicas para cubrir sus necesidades era inasumible.

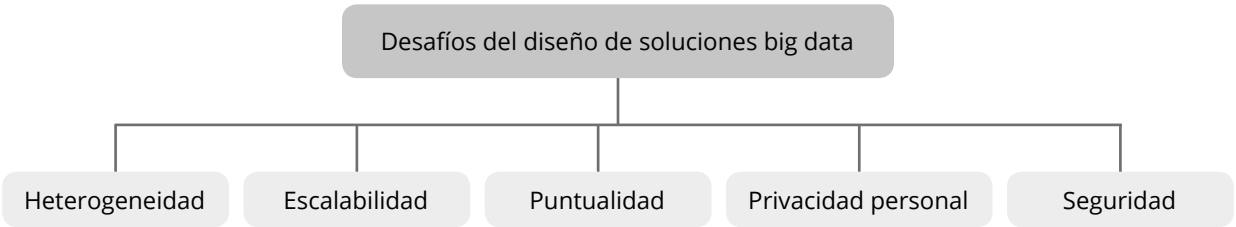
IMPORTANTE

El coste de los sistemas de información clásicos ha sido la razón más importante para que el mundo big data se transforme en una industria que ya no utiliza la gestión de datos como soporte de un negocio, sino que el propio procesamiento del dato es la razón de ser de la industria.

1.5. DESAFÍOS DEL DISEÑO DE SOLUCIONES BIG DATA

Una vez detalladas las características e identificados los síntomas del big data, podemos concretar cuáles son los **desafíos** que tendremos como **data engineers** en el diseño de soluciones big data.

Existen ciertos aspectos clave que hacen que los casos big data sean muy desafiantes. Vamos a detallar algunos de ellos.



HETEROGENEIDAD

Los seres humanos somos capaces de consumir una gran diversidad de formatos de información. El lenguaje humano, por ejemplo, está repleto de matices y connotaciones, como las figuras retóricas, que añaden una valiosa profundidad al mensaje. Somos capaces de adaptarnos a nuevas estructuras complejas de datos y aprender nuevas formas de comunicación. Sin embargo, los sistemas digitales de análisis esperan un conocimiento **estandarizado** y **coherente**, y no tienen la capacidad de entender matices, contextos ni segundas intenciones. Como consecuencia, el conocimiento debe estructurarse cuidadosamente como un primer paso para su análisis.

Los sistemas informáticos funcionan de manera más eficiente si pueden homogenizar los datos en formato y estructura.

IMPORTANTE

El **desafío** como data engineer reside en la búsqueda de soluciones para la extracción de conocimiento entre la compleja diversidad de los datos, aunque ello requiera un mayor trabajo.

ESCALABILIDAD

Como su nombre indica, big data pide un trabajo masivo. Cuando tiene lugar un aumento de tamaño, variedad o velocidad, aparecen problemas subyacentes en términos de almacenamiento, búsqueda, procesamiento, transformación y análisis.

Como ya se ha comentado, las características big data (las 3 *V*) se incrementan mucho más rápido que las capacidades hardware que pueden ofrecer los equipos informáticos, como las velocidades de CPU o el espacio de memoria en RAM. Por ese motivo, el concepto de **escalabilidad vertical**, entendido como la capacidad de los sistemas hardware de crecer en recursos para aumentar el rendimiento de las soluciones analíticas, ya no es una opción válida ni rentable.

Es entonces cuando aparece el concepto de **escalabilidad horizontal** gracias a otro concepto clave, el de distribución, que, recordemos, nos permite crecer modularmente y de forma cuasi ilimitada aumentando de forma lineal el rendimiento sin un incremento exponencial del coste.

IMPORTANTE

El **desafío** como data engineer lo encontramos en la complejidad del diseño, el desarrollo y el mantenimiento de estos sistemas distribuidos.

PUNTUALIDAD

Con esto nos referimos a la velocidad para llegar a un resultado de valor, ya que cuanto mayor sea el tamaño de los datos que se procesen, más se tardará en analizarlos.

Existen muchos escenarios en los que los resultados del análisis se requieren en tiempo real o en una breve ventana de tiempo. Por ejemplo, los coches autónomos deben tomar decisiones de conducción a partir de los datos recogidos por los sensores lo más rápido posible para poder reaccionar ante los peligros, y los resultados de los análisis médicos deben llegar antes de que sea demasiado tarde aplicar la cura.

IMPORTANTE

Esto crea uno de los **desafíos** más novedosos e importantes: crear un sistema big data que pueda procesar los datos a tiempo y de manera oportuna.

PRIVACIDAD PERSONAL

Mucha de la **información personal** que se captura, almacena, analiza y procesa en sistemas big data proviene de redes móviles y sensores, operadores de red, servicios de internet, supermercados, transportes, instituciones educativas, médicas o legales, organizaciones gubernamentales, servicios financieros, incluidos bancos, compañías de seguros y agencias de tarjetas de crédito. Asimismo, se está almacenando una gran cantidad de información tanto en redes sociales como en otros sistemas de rastreo y captura de datos.

Esto implica que la privacidad es un problema cuya importancia, en particular para el cliente, crece a medida que el valor del big data se hace más evidente. Los **algoritmos** utilizan estos datos personales para personalizar, perfilar, etiquetar, detectar, inferir y, en definitiva, extraer valor de los datos para un **uso lucrativo**.

IMPORTANTE

Para el data engineer se plantea claramente un **desafío** ético y legal de uso de los datos para no incurrir en una violación de la privacidad personal.

SEGURIDAD

La seguridad también es una gran preocupación, tanto para las empresas como para las personas. Los grandes almacenes de datos pueden ser un objetivo atractivo para *hackers* o amenazas complejas y persistentes.

IMPORTANTE

La seguridad es un **desafío** para el data engineer y un atributo esencial en la arquitectura big data que revela formas de almacenar y proporcionar acceso a la información de manera segura.

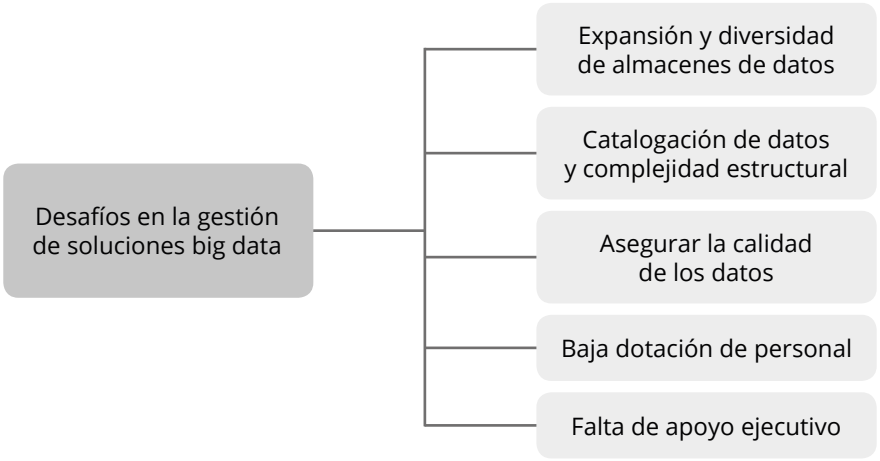
1.6. DESAFÍOS EN LA GESTIÓN DE SOLUCIONES BIG DATA

Con el auge de las soluciones big data en las organizaciones, las empresas tienen un gran interés en explorar estos sistemas para que proporcionen oportunidades y perspectivas de aumentar los beneficios en el negocio.

IMPORTANTE

El mantenimiento, la administración y el control de las soluciones big data conllevan todavía una alta dificultad debido al poco grado de madurez de estas tecnologías.

A continuación, se indican algunos de los principales desafíos que presenta el proceso de gestión de big data.



EXPANSIÓN Y DIVERSIDAD DE ALMACENES DE DATOS

Tener un enorme volumen de datos involucrados, y el hecho de que crecen continuamente con el tiempo, hace que la gestión de la **persistencia**, del **acceso** y de la **operación** en los datos sea muy compleja y desafiante.

También es crucial la capacidad de realizar cualquier tipo de operación entre distintos sistemas y tecnologías de almacenaje de datos **sin perder rendimiento ni calidad** en el análisis, ya que en soluciones big data, tendremos fácilmente conviviendo múltiples soluciones de almacenaje dentro de una misma solución.

IMPORTANTE

El desafío de gestión más complejo es conseguir interconectar todas las soluciones de almacenamiento de datos para obtener un acceso global y romper los silos aislados de datos.

CATALOGACIÓN DE DATOS Y COMPLEJIDAD ESTRUCTURAL

Las empresas suelen contar con datos estructurados y datos no estructurados, y todos en una amplia gama de formatos (JSON, XML, bases de datos, archivos de texto, datos binarios, etc.). Una empresa, generalmente, tiene varios miles de aplicaciones en sus sistemas, y cada una de ellas puede consultar y escribir en varias bases de datos distintas. Como resultado, el simple hecho de catalogar qué tipo de datos tiene una organización en sus sistemas de almacenamiento es, a menudo, extraordinariamente difícil.

IMPORTANTE

Catalogar los datos, como fondo de conocimiento del universo de datos que tiene una compañía, es una de las necesidades y desafíos más vitales para la viabilidad de una empresa que dependa de ellos.

ASEGURAR LA CALIDAD DE LOS DATOS

La calidad de los datos es una de las claves para que las empresas garanticen la confiabilidad y exactitud de dichos datos.

Como ya se ha mencionado anteriormente, el **déficit de sincronización** entre los silos de datos puede complicar la confiabilidad de que parte de los datos no sean precisos y/o completos. Del mismo modo, la **falta de validación y trazabilidad** de los procesamientos puede producir errores no detectados que afecten a la calidad del dato.

RECUERDA

Si en un sistema de datos se introducen datos incorrectos, la salida generada también será incorrecta.

IMPORTANTE

El desafío se halla en la protección del dato original, la validación de la información por el cruce de múltiples fuentes y el control exhaustivo de los procesos para protegerlos de errores, incluidos los humanos.

BAJA DOTACIÓN DE PERSONAL

Es difícil y desafiante encontrar **personal cualificado** con conocimiento sobre el dominio del problema big data. Debido a la falta de ingenieros de datos, científicos de datos, analistas de datos y diferentes profesionales especializados en big data, la creación de equipos y el trabajo de gestión resulta muy complicado.

FALTA DE APOYO EJECUTIVO

Desgraciadamente, en muchas empresas, los altos directivos aún no aprecian la importancia y el valor de una buena gestión de los datos. Es muy difícil convencerlos y demostrarles cómo estas técnicas de gestión pueden ser beneficiosas para su organización. En otras palabras, están centrados en la toma de decisiones a partir de casos de uso del dominio del negocio y no basan parte de su decisión estratégica en la información que pueden extraer de sus datos.

Por suerte, algunas organizaciones ya han realizado este cambio y han adoptado un nuevo paradigma de toma de decisiones basadas en datos denominado ***data driven***.

1.7. CONCLUSIONES

Como data engineer, cuando te enfrentes a los casos y desafíos que hemos citado, es posible que desees considerar una arquitectura big data distribuida para poder escalar horizontalmente a medida que crece tu negocio de modo eficiente en coste y tiempo, así como proveer soluciones flexibles a los cambios en los datos y las tecnologías.

IMPORTANTE

El objetivo global de un data engineer es diseñar, implementar y mantener soluciones para cubrir las necesidades del análisis big data.

En el siguiente tema veremos las distintas opciones de arquitectura de las que disponemos para ofrecer estas soluciones, pero ya podemos intuir que se basarán en arquitecturas distribuidas para no depender del rígido, clásico y costoso *mainframe* de alto rendimiento. En su lugar, deberemos basarnos en tecnologías más disponibles, impulsadas por el rendimiento y más baratas para aportar más flexibilidad y eficiencia.

IDEAS CLAVE

- Es preciso entender la problemática big data desde un punto de vista ingenieril para comprender el cometido de un ingeniero de datos.
- Tras la crisis generada por el fenómeno big data en los sistemas de información clásicos, se hace relevante y aumenta la demanda de perfiles como el de ingeniero de datos.
- La diagnosis de los problemas big data es una parte esencial del futuro diseño de soluciones para dichos problemas.

BIBLIOGRAFÍA

BUHLER, P.; KHATTAK, K.; ERL, T. (2016) *Big Data Fundamentals: Concepts, Drivers & Techniques*. Nueva Jersey: Prentice Hall. Disponible en: https://learning.oreilly.com/library/view/big-data-fundamentals/9780134291185/.

FAHAD AKHTAR, S. M. (2018) *Big Data Architect's Handbook*. Birmingham: Packt Publishing. Disponible en: https://learning.oreilly.com/library/view/big-data-architects/9781788835824/.

HARRISON, G. (2016) *Next Generation Databases: NoSQL, NewSQL, and Big Data*. Nueva York: Apress. Disponible en: https://learning.oreilly.com/library/view/next-generation-databases/9781484213292/.

REINSEL, D.; GANTZ, J.; RYDNING, J. (2018) *The Digitalization of the World. From Edge to Core*. Framingham: IDC & Seagate. Disponible en: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf.

SALÉ, M. J.; BARILLA, R. J.; LINDO, S. (2015) "Big Data: A Foundational Explanation". Proceedings of Student-Faculty Research Day, CSIS. Disponible en: http://csis.pace.edu/~ctappert/srd2015/2015PDF/d6.pdf.