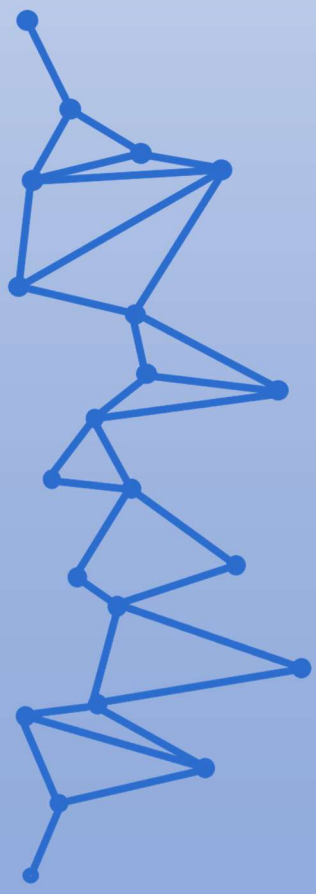
Curso de Especialización de
Inteligencia Artificial y Big Data (IABD)



Sistemas de Big Data

UD01. Introducción a Big Data.
Resumen.

JUAN ANTONIO GARCIA MUELAS

Las metodologías y tecnologías para Big Data (macrodatos en castellano) aparecen como respuesta la necesidad de tratar cantidades de datos tan grandes que desbordan los sistemas convencionales monomáquina.

¿Qué problema de base origina la aparición de las metodologías y tecnologías Big Data? El tener grandes cantidades de datos que desbordan los recursos de máquinas individuales.

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

https://en.wikipedia.org/wiki/Big_data

5 características de Big Data para discernir si el procesamiento que realizamos es o no big data (las 5 Vs).

- ✓ Volumen (cantidad de bytes).
- ✓ Velocidad.
- ✓ Variedad.
- ✓ Veracidad.
- ✓ Valor.

VOLUMEN

¿A partir de qué cantidad de datos es Big Data?

No existe ninguna entidad u organismo que regule de algún modo cuál es el tamaño de datos concreto a partir de la cual se considera que estamos en un ambiente Big Data.

Unidades de cantidad de información digital	
Unidades	**Significado**
Bit	Unidad mínima de información en un sistema de computación (almacena un "0" o un "1").
byte (B)	8 bits
kilobyte (kB)	1000 bytes (10^3 bytes)
megabyte (MB)	1000 kilobytes (10^6 bytes)
gigabyte (GB)	1000 megabytes (10^9 bytes)
terabyte (TB)	1000 gigabytes (10^{12} bytes)
petabyte (PB)	1000 terabytes (10^{15} bytes)
exabyte (EB)	1000 petabytes (10^{18} bytes)
zettabyte (ZB)	1000 exabytes (10^{21} bytes)
yottabyte (YB)	1000 zettabytes (10^{24} bytes)

Hay que tener en cuenta que si bien el significado de un kilobyte (kB) es 1000 bytes, dado que en ambientes de computacionales se emplea constantemente la numeración en base 2 también existe el kibibyte (KiB), el cual corresponde a 1024 bytes (2^{10}). De igual modo, también existe el

mebibyte (MiB = 2^{20} bytes), el gibibyte (GiB = 2^{30} bytes), y así toda la progresión hasta llegar al yobibyte (YiB = 2^{80} bytes).

Simplemente nos quedaremos con que los sistemas para Big Data hoy en día trabajan con volúmenes del orden de los petabytes (PB) e incluso de los exabytes (EB).

Son considerados datos abiertos todos aquellos datos accesibles y reutilizables, sin exigencia de permisos específicos. Han de ser fiables, estructurados, documentados y fácilmente accesibles.

Si en algún atributo de la especificación de un dispositivo hardware vemos un valor de 1 kB, ¿a cuántos bytes corresponde? Dependiendo de la situación, quizás se refiera a 1 kB (que corresponde a 1000 bytes), o a 1KiB (que corresponde a 1024 bytes).

VELOCIDAD.

El problema con respecto a la velocidad no es únicamente el hecho de que el volumen de datos continúe creciendo sin parar, sino lo rápido que es necesario obtenerlos y ser capaces de integrarlos junto con los que ya tenemos.

Si quisiésemos almacenar diariamente un valor de 4 bytes con el peso de los aproximadamente 7870 millones de personas, ¿cuánto nos ocuparía tal información a largo de un año?

$4 * (7{,}890 * 10^9) * 365 = 11{,}5194 * 10^{12}$ bytes = 11,5194 TB

De la gran velocidad a la que llegan datos nuevos nacen las estrategias de procesamiento tipo streaming, las cuales estudiaremos más adelante.

VARIEDAD.

- ✓ **Datos estructurados:** existentes den registros como bases de datos, con esquema definido y tipo de dato.
- ✓ **Datos no estructurados:** sin esquema, como audios, videos… y que son en proporción el 80% del total por su espacio.
- ✓ **Datos semiestructurados**: Definidos según cierta estructura pero sin naturaleza relacional (no pertenecen a una tabla)…CSV, JSON, XML…
- ✓ **Metadatos**: datos extra muchas veces automáticos para favorecer una interpretación posterior.

VERACIDAD.

Un problema extra con el que tenemos que tratar es el hecho de que los datos no siempre cuentan con la calidad deseada o no son totalmente fieles a la realidad.

Este término está muy relacionado con el concepto de **relación seña/ruido** en cualquier flujo de información.

- ✓ El ruido son datos que no pueden ser convertidos en información (ya sea porque no la contienen o porque ésta está corrupta y es irrecuperable).
- ✓ La señal está constituida por datos que sí pueden ser convertidos en información con sentido.

Por ello, es necesario conocer en qué condiciones se adquirieron los datos (para estimar la veracidad) y procesarlos en su caso para resolver problemas y eliminar información inválida.

Hay ruido en los datos cuando parte de los datos no contienen información usable o de la que se pueda obtener algún tipo de valor.

VALOR.

El concepto de valor en relación con los datos tiene que ver con cómo de útiles son estos para una institución, empresa o persona.

Es importante su veracidad, el tiempo transcurrido desde que se produjeron o la propia interpretación, para determinar el valor.

El orden de etapas desde que ocurren los eventos hasta que de ellos se genera valor:

Eventos -> Datos -> Información -> Conocimiento -> Sabiduría -> Valor

Si los datos van perdiendo valor con el tiempo y tenemos muchos datos antiguos, ¿merece la pena utilizarlos siempre junto con los más nuevos?

Retroalimentación

No siempre. Dependerá del caso.

Sobre todo para el análisis descriptivo suele ser útil tener en cuenta datos antíguos para comprobar cuáles son las tendencias (es decir, tener algo con lo que comparar los datos nuevos).

En otros casos, como por ejemplo en la generación de modelos predictivos para intentar acertar con una oferta, lo que querremos es tener cuantos más datos nuevos mejor para así no necesitar usar los más antiguos (ya que los gustos de los usuarios son cambiantes).

Aportes generales de Big Data

Las metodologías y tecnologías para Big Data nos permiten realizar diversas operaciones con grandes cantidades de datos, entre las cuales se encuentran:

- ✓ Capturarlos desde sus orígenes.
- ✓ Integrarlos para poderlos almacenar de un modo unificado.
- ✓ Almacenarlos de un modo distribuido y replicado, gracias lo cual conseguimos altos valores de disponibilidad.
- ✓ Tratarlos de forma distribuida, empleando para ello un alto número de máquinas que los procesan en paralelo.
  En relación a procesamiento paralelo de tareas, **algunas tareas no pueden paralelizarse** para que se ejecuten más rápido usando varias máquinas.
- ✓ Aplicar técnicas de minería de datos (también llamado ciencia de datos cuando esa minería de datos se realiza en ambientes Big Data) para crear modelos predictivos.
- ✓ Usar esos modelos para realizar predicciones a utilizar en sistemas automáticos.
- ✓ Crear visualizaciones y cuadros de mando usando tanto los propios datos como los modelos creados para así dar soporte a la toma de decisiones.

El ser capaces de realizar tales operaciones con los datos, nos permiten obtener los siguientes aportes y beneficios (entre otros):

- ✓ Generar registros más detallados mediante la integración desde diversas fuentes.
- ✓ Optimizar las operaciones de instituciones y empresas.
- ✓ Poder actuar de modo inteligente basándonos en la evidencia de los datos.
- ✓ Identificar nuevos mercados.
- ✓ Realizar predicciones basándonos en modelos creados a partir de los datos.
- ✓ Detectar casos de fraude e impagos.
- ✓ Dar soporte a la toma de decisiones.
- ✓ Realizar descubrimientos científicos.

- ✓ Ayudar a los médicos a detectar enfermedades en función del historial de los pacientes y las pruebas que se les realizan.
- ✓ Crear nuevos fármacos más efectivos y con menos efectos secundarios.

A pesar de que dentro de las tecnologías de Big Data se suele englobar lo relacionado con obtener valor del dato, en la práctica son la minería de datos o la ciencia de datos las disciplinas que terminan de obtener el valor (haciendo uso de esas tecnologías).

CLÚSTERES DE COMPUTADORAS

Son un **conjunto de computadoras** (también referenciados como servidores o como nodos) **conectados entre sí mediante red para trabajar como una única unidad** resolviendo cargas de trabajo de forma conjunta.

Han ido evolucionando desde caras computadoras a frameworks y plataformas de computación distribuida con computadoras de uso común (commodity harware).

Ventajas del clúster:

- ✓ **Alto rendimiento:** Dado que cada componente del clúster es una computadora completa, con sus propios recursos (procesador, memoria y almacenamiento), las cargas de trabajo susceptibles de paralelización pueden acelerarse en gran medida dividiéndolas en subtareas y distribuyéndolas para que sean ejecutadas en los distintos nodos.
  Gracias a esto se pueden resolver problemas muy complejos que no sería posible resolver en un tiempo razonable en una máquina individual por muy potente que ésta sea.
- ✓ **Alta disponibilidad:** Mediante una continua monitorización entre los propios nodos del clúster, se puede detectar la no disponibilidad de un subconjunto de los mismos (ya sea por fallo eléctrico, por avería o por corte de las comunicaciones) y se pueden tomar medidas para que los servicios o datos que hay (o había) en esas máquinas sigan estando disponibles.
  - ➢ Rearrancando un nodo caído o arrancando un nuevo nodo para suplirlo.
  - ➢ Respondiendo las peticiones desde otro nodo del clúster que también contenga una réplica de esos datos.
- ✓ **Equilibrado de carga:** El equilibrado de carga (o también balance o balanceo) se consigue mediante algoritmos destinados a distribuir las cargas de trabajo entre los diversos nodos del clúster para así evitar cuellos de botella. Tales cuellos de botella se producen cuando el envío de trabajos a nodos sobrecargados aumenta la latencia media con la que tales trabajos son finalizados.
  Para ello, se realiza una monitorización del estado de carga de cada nodo y se decide para cada paquete de trabajo a qué nodo enviarlo, atendiendo a:
  - ➢ El tamaño del trabajo.
  - ➢ El estado de carga de cada nodo.
  - ➢ La potencia de procesamiento de cada nodo.
- ✓ **Escalabilidad:** Gracias a que el clúster está formado por un número indeterminado de nodos, no sólo conseguimos una mayor potencia de cálculo al utilizarlos para una misma tarea, sino que podemos hacer crecer dicha potencia de cálculo añadiendo nuevos nodos. En otras palabras, la potencia de cálculo del clúster es ampliable.

Esta característica es muy deseable para sistemas Big Data, ya que desaparece la necesidad de realizar una estimación de potencia necesaria a priori, lo cual en por lo general siempre lleva a una sobreestimación para guardar un margen de seguridad. Con un clúster escalable podemos comenzar con un número determinado de nodos e ir añadiendo más según sea necesario.

Es interesante conocer la diferencia entre escalado horizontal y vertical:

- ✓ **Escalado vertical (scale-in):**
  - ➢ Es el que se consigue mejorando las características hardware de la computadora (individual) en el que se están ejecutando las cargas de trabajo (procesador, memoria o almacenamiento). Por lo tanto, está limitado por la mejor especificación de hardware que sea posible encontrar en el mercado.
  - ➢ Por ello, aunque reciba el nombre de "escalado" en la práctica no sirve para conseguir la característica de escalabilidad.
- ✓ **Escalado horizontal (scale-out):**
  - ➢ Es el que se consigue añadiendo más nodos a un clúster.
  - ➢ Por ello es el tipo de escalado que realmente nos permite conseguir la característica de escalabilidad.

Conceptos relacionados con almacenamiento que es importante conocer si vamos a trabajar en entornos Big Data.

Veremos aquí un esquema/resumen para que puedas tener una vista general:

- ✓ **Base de Datos Relacional:** El tipo de bases de datos más utilizado en el mundo (pero no escalable para Big Data). Las bases de datos relacionales escalan en vertical (esto es, para tener más potencia o capacidad de almacenamiento ponemos un servidor mejor). Esto supone un cuello de botella para si nos encontramos en un entorno Big Data, ya que no siempre podremos poner un servidor con CPU más potente, con más memoria o con más capacidad de almacenamiento.
- ✓ **Dataset:** Un conjunto o colección de datos (quizás enorme) que guardan cierta relación.
- ✓ **Almacén de Datos:** Una sistema especial para almacenar datos (típicamente para analítica).
- ✓ **ACID:** Una serie de propiedades (**A**tomicidad, **C**onsistencia, A**i**slamiento/**I**solate y **D**urabilidad) que deben cumplir las bases de datos que vayan a ser usadas para realizar transacciones.
- ✓ **Teorema CAP:** Un teorema acerca de las propiedades que podemos conseguir en una base de datos distribuida.
  El teorema CAP (también conocido como conjetura de Brewer) establece que una base de datos distribuida sólo puede cumplir como máximo con 2 de las siguientes 3 propiedades:
  - ➢ Consistencia (consistency).
  - ➢ Disponibilidad (availability).
  - ➢ Tolerancia a particionamiento (partition tolerance).

  En otras palabras, según el teorema, nunca puede cumplirse C+A+P, sino que habrá que escoger siempre entre C+A, C+P o A+P a la hora de diseñar la base de datos distribuida.
- ✓ **BASE:** Un principio de diseño de base de datos distribuidas.
  BASE es un principio de diseño de bases de datos basado en las restricciones impuestas por el teorema CAP, y típicamente empleado por muchas implementaciones de bases de datos distribuidas.
  El significado del acrónimo es:
  - ➢ Básicamente disponible (básically available).
  - ➢ Estado blando (soft state).
  - ➢ Consistencia eventual (eventual consistency).

Una base de datos que conforme a la filosofía BASE prefiere la disponibilidad antes que la consistencia (es decir, desde el punto de vista del teorema CAP es A+P).

Conceptos relacionados con **procesamiento de datos** de que es importante conocer si vamos a trabajar en entornos Big Data.

Veremos aquí un esquema/resumen para que puedas tener una vista general:

- ✓ **Procesamiento en paralelo:** Distintos procesos dentro del mismo procesador. Tiene que ver con la capacidad de realizar varias tareas al mismo tiempo
- ✓ **Procesamiento en distribuido:** Distintos procesos para un mismo trabajo ejecutándose en distintas máquinas. Como el clúster.
- ✓ **Estrategias de procesamiento de datos:** Cómo trabajamos con datos según el tipo de actividad que vayamos a realizar.
  Por lotes, transaccional, en tiempo real, Streaming…
  Por lotes implica tener mucho tiempo disponible para procesar los datos.
  Tiempo real implica que los resultados se producen en poco tiempo, mientras que en streaming implica que es capaz de tener en cuenta datos que van entrando constantemente.
- ✓ **OLTP:** Procesamiento transaccional (sistema que está orientado a transacciones).
- ✓ **OLAP:** Procesamiento para analítica en tiempo real. Para inteligencia de negocio (BI) y minería de datos.
- ✓ **Principio SCV:** Un principio que nos dice qué propiedades podemos conseguir en un sistema de procesamiento distribuido.
  De modo similar a lo que ocurría con el teorema CAP, el principio SCV establece que un sistema de procesamiento distribuido sólo puede soportar como máximo 2 de las siguientes 3 características.
  - ✓ Velocidad (speed).
  - ✓ Consistencia (consistency).
  - ✓ Volumen (volume).

Se emplea una **arquitectura** según la cual el flujo de datos va pasando por una serie de **capas**.

- ✓ **Capa de ingestión:** Como primer paso, los datos se obtienen desde múltiples fuentes con las cuales es necesario conectarse de algún modo. Hay que tener en cuenta que la gran mayoría de las fuentes son preexistentes a la creación del sistema Big Data que se esté desarrollando, por lo que es el sistema el que tiene que adaptarse a las fuentes y no a la inversa (empleando el protocolo correspondiente a las mismas y siendo capaz de interpretar los datos que de ellas se obtienen).
- ✓ **Capa de colección:** Una vez que se obtienen e interpretan los datos viene el trabajo relacionado con integrarlos para darles una estructura propia. Hay que tener en cuenta que las fuentes de datos pueden ser muchas y de naturaleza muy variada, cada una emitiendo información en un formato distinto, de modo que hay que unificarlo todo para representarlo como un único conjunto de datos con sentido y ya prácticamente listos para ser utilizados.
  La capa de colección en una arquitectura para Big Data es aquella en la que se pasa de los datos que vienen de diversas fuentes a un conjunto de datos unificado y ya casi listo para ser usado.

- ✓ **Capa de almacenamiento:** Como no podía ser de otro modo, esa gran cantidad de datos que da origen al concepto Big Data debe ser almacenada, para lo cual se emplean sistemas de almacenamiento distribuido especialmente diseñados para ello.
- ✓ **Capa de procesamiento:** Es la capa que provee de infraestructura a la siguiente capa (la de consulta y analítica) para poder tratar con gran cantidad de datos. Es decir, facilita el procesamiento (por lotes, en tiempo real, streaming o híbrido) pero únicamente hace lo que le está pidiendo la capa superior, no obteniendo valor del dato de por sí.
- ✓ **Capa de consulta y analítica:** Es la capa en la que se comienza a obtener valor al dato, realizando la estadística, algoritmia o análisis que se considere oportuno, para ello siempre basándose en la capa previa de procesamiento.
- ✓ **Capa de visualización:** Es la capa con la que interacciona el usuario final, el cual puede consultar reportes estáticos o acceder a cuadros de mando interactivos con diversas visualizaciones y controles desde los cuales puede decidir qué información ver y cómo quiere verla representada. Desde esta capa es desde por lo general se toman las decisiones de negocio.
- ✓ **Capa de seguridad:** Capa transversal que da soporte a todo lo relacionado con asegurar la seguridad en los datos empleando métodos tanto físicos como de software. Incluye protección ante el ataque o uso malintencionado tanto desde dentro como desde fuera de la empresa o institución.
- ✓ **Capa de monitorización:** Capa transversal que da soporte a todo lo relacionado con la monitorización tanto de los datos como del propio sistema. La monitorización de datos incluye auditoría, testeo, gestión y control, de modo que los datos a emplear para obtener valor sean correctos y frescos. Tal monitorización es una parte importante de los mecanismos de gobernanza de datos.

Paisaje Big Data (más usado en inglés, como "The Big Data Landscape") para referirnos al panorama de las diversas herramientas y utilidades que se pueden emplear para desarrollar proyectos Big Data.