Definición de Big Data

Módulo 1 – Lección 2

ibigdata.es

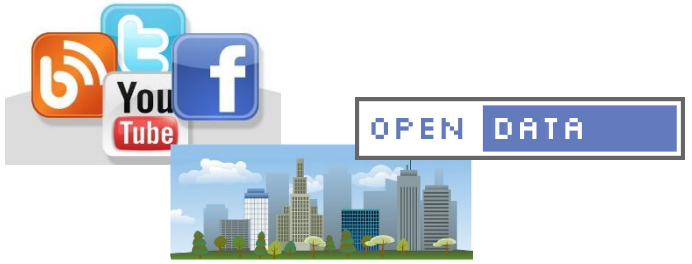
Universitat d'Alacant
Universidad de Alicante

Agenda

- Volumen, Velocidad, Variedad, Veracidad, Valor
- Definición de Big Data
- Problemática

Introducción

- Principales dispositivos y aplicaciones que están dando lugar a un nuevo escenario de datos...



- ...que supone importantes cambios en las características de los datos disponibles para el análisis

Volumen

Velocidad

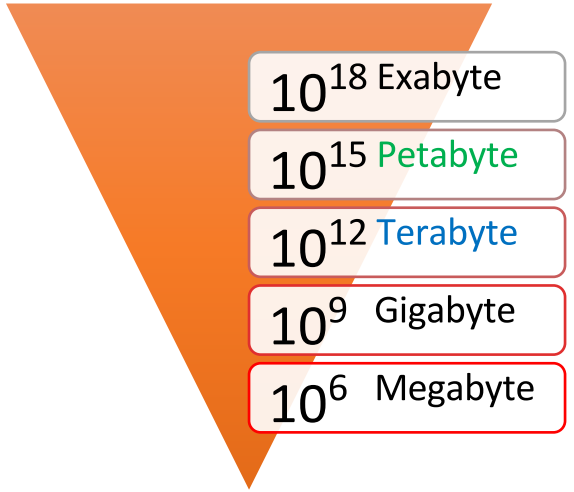
Variedad

Veracidad

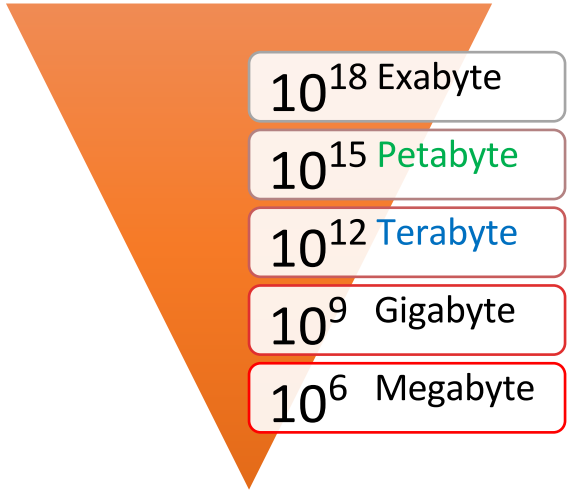
Valor

Volumen

- Aumento del **Volumen** de los datos disponibles para el análisis

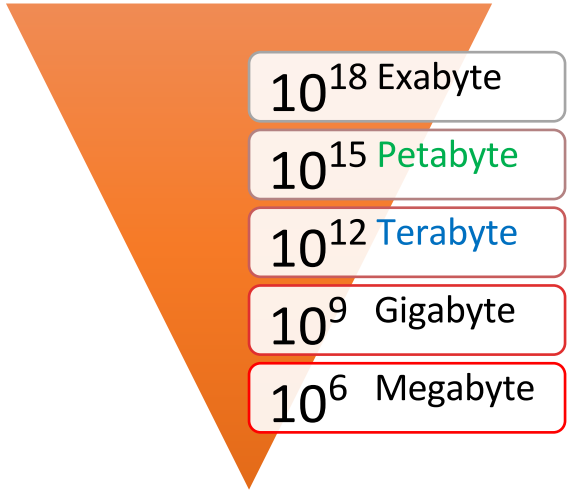


Volumen



- Aumento del **Volumen** de los datos disponibles para el análisis
- Las empresas que han adoptado Big Data gestionan desde unos pocos Terabytes (10^3 Gb) hasta Petabytes (10^6)

Volumen



- Aumento del **Volumen** de los datos disponibles para el análisis
- Las empresas que han adoptado Big Data gestionan desde unos pocos Terabytes (10^3 Gb) hasta Petabytes (10^6)
- Muchas de ellas ya superan los **10 Terabytes**
  - En tres años lo habitual será >= 100 Terabytes

Volumen

10^{18} Exabyte

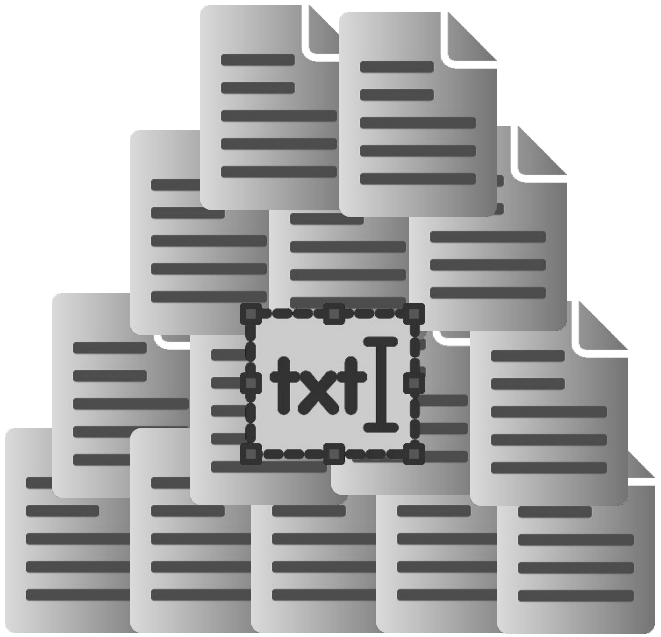
10^{15} Petabyte

10^{12} Terabyte

10^{9} Gigabyte

10^{6} Megabyte

Sin embargo, el Volumen no es lo más importante...



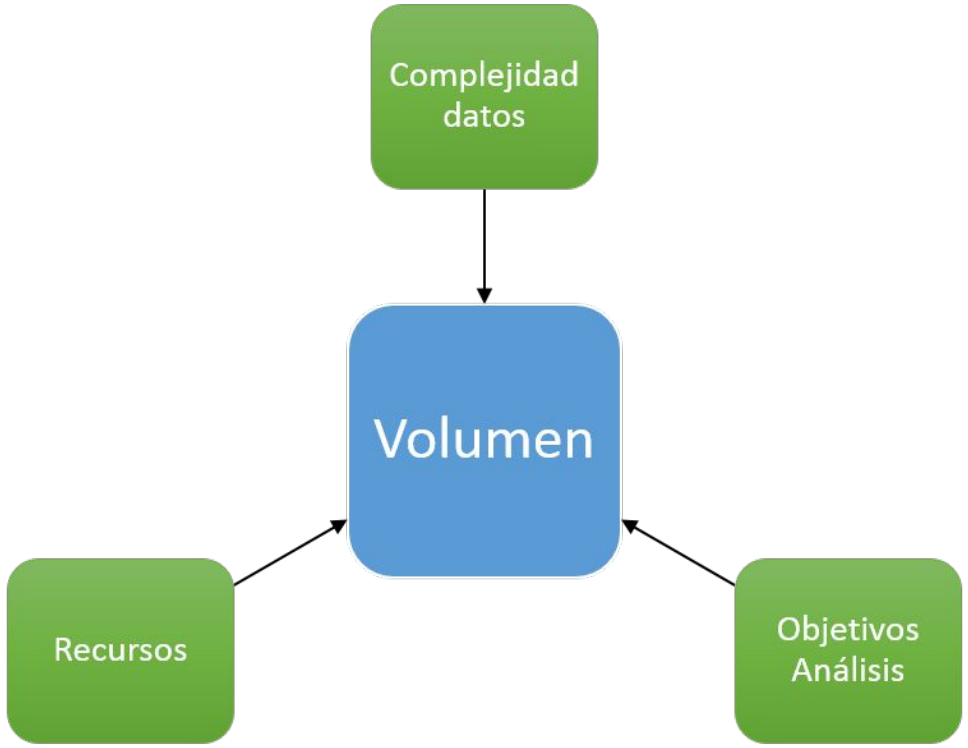
≠

1 Terabyte de texto

1 Gigabyte de imágenes médicas

Volumen

A la hora de valorar el Volumen de los datos hemos de tener en cuenta otros factores...



Velocidad

- Incremento de la **Velocidad** a la que se generan y se distribuyen los datos en las fuentes
  - Es una de las razones del incremento del volumen de datos

Telescopio SKA
10 Petabytes / hora

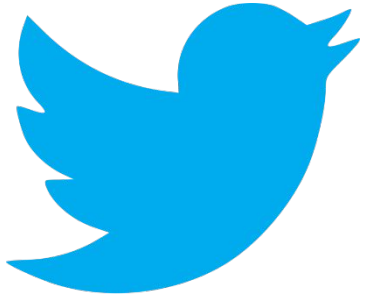
Twitter
100.000 tweets / min

Velocidad

- Incremento de la **Velocidad** a la que se generan y se distribuyen los datos en las fuentes
  - Es una de las razones del incremento del volumen de datos

- **Streaming**: datos que se generan y distribuyen en tiempo real
  - Generados por sensores, servidores web, redes sociales...

Telescopio SKA
10 Petabytes / hora



Twitter
100.000 tweets / min

Velocidad

- También hace referencia a la necesidad para extraer conocimiento de los datos en el **momento oportuno**
  - Ej. Datos de tipo financiero

Variedad

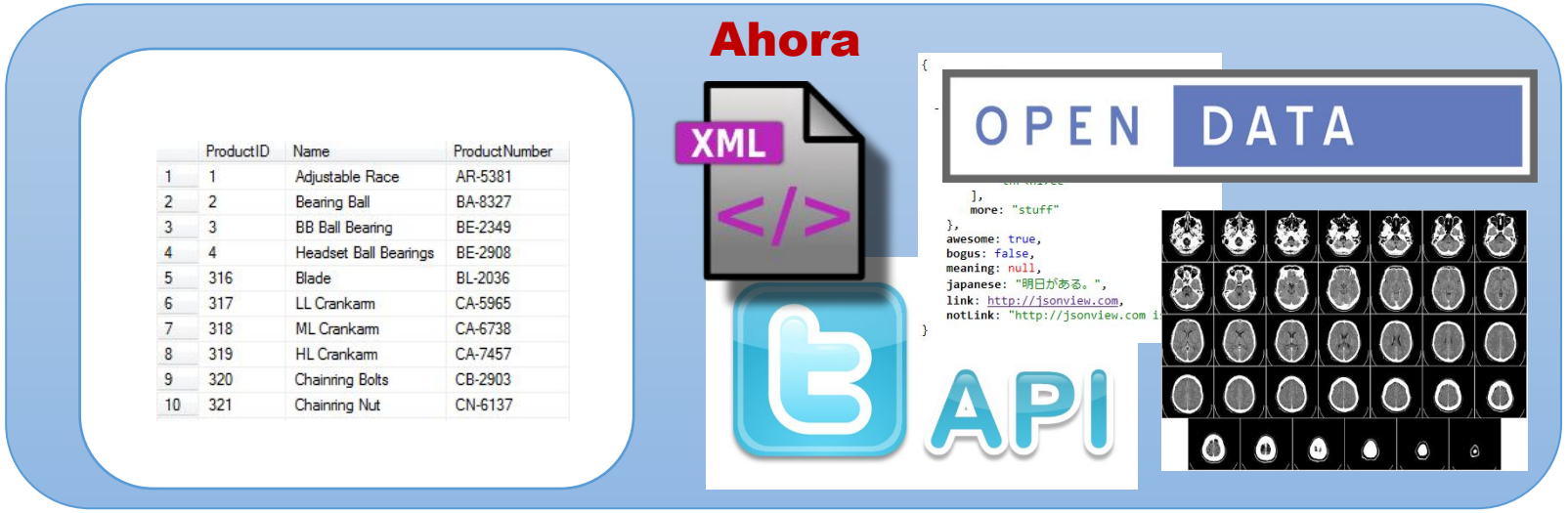
- La **Variedad** se refiere al importante aumento en la heterogeneidad de las fuentes de datos debido a diversos factores como:
  - Incremento en el **número** de fuentes disponibles
  - Fuentes con **distinto** nivel de **estructura**
  - **Diversidad** de **formatos** de distribución

Antes

	ProductID	Name	ProductNumber
1	1	Adjustable Race	AR-5381
2	2	Bearing Ball	BA-8327
3	3	BB Ball Bearing	BE-2349
4	4	Headset Ball Bearings	BE-2908
5	316	Blade	BL-2036
6	317	LL Crankarm	CA-5965
7	318	ML Crankarm	CA-6738
8	319	HL Crankarm	CA-7457
9	320	Chainring Bolts	CB-2903
10	321	Chainring Nut	CN-6137

Variedad

- La **Variedad** se refiere al importante aumento en la heterogeneidad de las fuentes de datos debido a diversos factores como:
  - Incremento en el **número** de fuentes disponibles
  - Fuentes con **distinto** nivel de **estructura**
  - **Diversidad** de **formatos** de distribución



Veracidad

- Aumento de la **incertidumbre** respecto de la **Veracidad** o calidad de los datos disponibles

Veracidad

- Aumento de la **incertidumbre** respecto de la **Veracidad** o calidad de los datos disponibles

Incertidumbre datos 

Veracidad

- Aumento de la **incertidumbre** respecto de la **Veracidad** o calidad de los datos disponibles

Incertidumbre datos  Incertidumbre conocimiento extraído

Veracidad

- Aumento de la **incertidumbre** respecto de la **Veracidad** o calidad de los datos disponibles

Incertidumbre datos → Incertidumbre conocimiento extraído

- Es uno los retos principales del nuevo contexto de datos
  - El uso de datos incorrectos supone **grandes pérdidas**

Valor

- El **Valor** es la medida de la utilidad de los datos seleccionados para nuestros objetivos finales

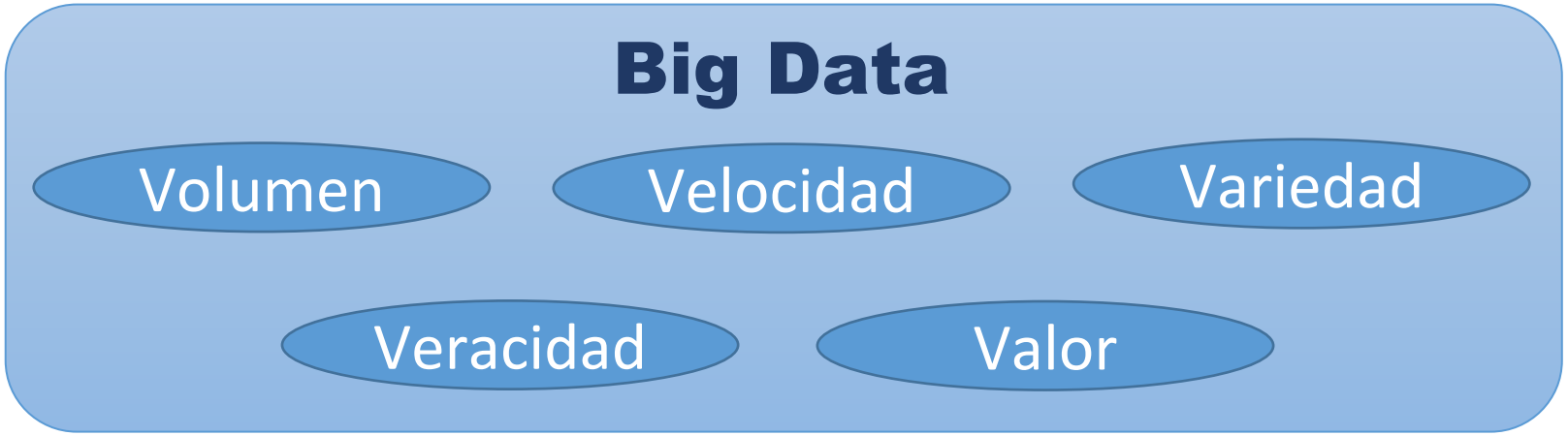
Valor

- El **Valor** es la medida de la utilidad de los datos seleccionados para nuestros objetivos finales

- **Determinar** dicha **utilidad** a priori puede ser realmente **complicado**
  - Debido a las características anteriores de Volumen, Velocidad, Variedad y Veracidad

Definición de Big Data

- **Big Data**: Datos que cumplen una o más de las características anteriores, junto con las técnicas y tecnologías para su correcto procesamiento y análisis
- De esta forma Big Data queda definido por las **5 V's** o características anteriores



Definición de Big Data

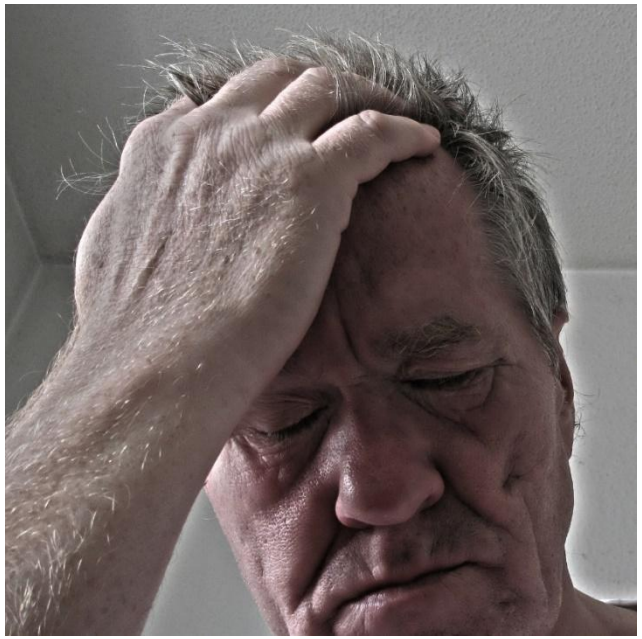
- **Definición 2**: Big Data surge como respuesta a las carencias, respecto a las 5v's, de las tecnologías de procesamiento y análisis previamente existentes
  - Ej. Almacenes de datos (DW) sobre BD's relacionales

Definición de Big Data

- **Definición 2**: Big Data surge como respuesta a las carencias, respecto a las 5v's, de las tecnologías de procesamiento y análisis previamente existentes
  - Ej. Almacenes de datos (DW) sobre BD's relacionales
- **Definición 3**: "Forma de afrontar el procesamiento o análisis de grandes volúmenes de información que por su naturaleza desestructurada no pueden ser analizados, en un tiempo aceptable, usando los procesos y herramientas tradicionales de BI" (Zikopoulos and Eaton. Understanding Big Data. IBM, 2011)
  - ¿Qué es BI? Definiremos este concepto en el módulo 2

Problemática

El **aprovechamiento** de Big Data para la extracción de conocimiento útil de los datos **no es una tarea trivial** ...



Las 5v’s son un problema en sí mismas

Problemática

El **aprovechamiento** de Big Data para la extracción de conocimiento útil de los datos **no es una tarea trivial** ...

Las 5v’s son un problema en sí mismas

Problemática

El **aprovechamiento** de Big Data para la extracción de conocimiento útil de los datos **no es una tarea trivial** ...

Las 5v’s son un problema en sí mismas

Problemática

El **aprovechamiento** de Big Data para la extracción de conocimiento útil de los datos **no es una tarea trivial** ...

Las 5v’s son un problema en sí mismas

Problemática

El **aprovechamiento** de Big Data para la extracción de conocimiento útil de los datos **no es una tarea trivial** ...

Las 5v’s son un problema en sí mismas

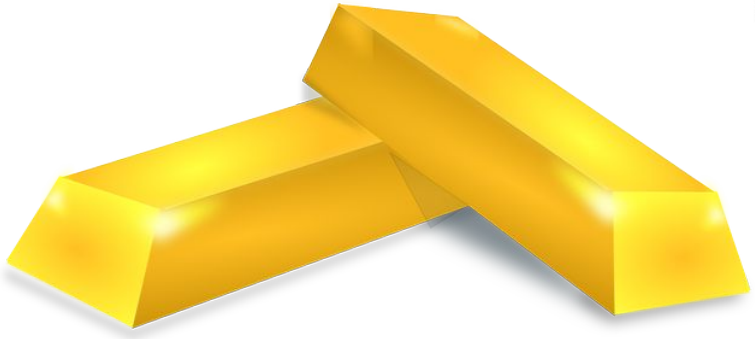
Problemática

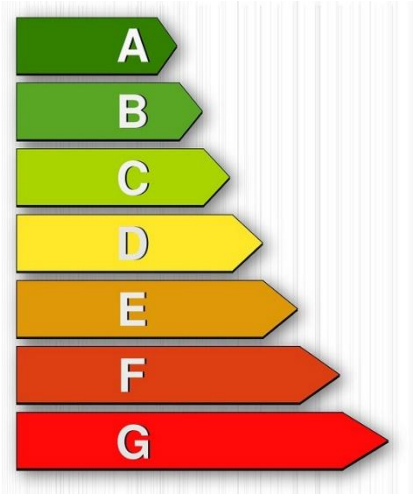
El **aprovechamiento** de Big Data para la extracción de conocimiento útil de los datos **no es una tarea trivial** ...

Las 5v's son un problema en sí mismas

Problemática

- Sin embargo, si tenemos éxito en dicha tarea podemos lograr **importantes beneficios**





Contacto

- Dudas, comentarios y noticias acerca del curso
  - Email: ibigdata@ibigdata.es
  - Twitter: @lucentialab
  - Página web del curso: http://ibigdata.es
- Otras webs
  - Grupo de investigación Lucentia http://www.lucentia.es/
  - Canal de YouTube https://www.youtube.com/user/lucentialab
  - Plataforma UniMOOC http://unimooc.com/