

Curso: (62949) Internet para las cosas

Fernando Tricas García

Departamento de Informática e Ingeniería de Sistemas
Universidad de Zaragoza

<http://webdiis.unizar.es/~ftricas/>

<http://moodle.unizar.es/>

ftricas@unizar.es

Big Data

Fernando Tricas García

Departamento de Informática e Ingeniería de Sistemas
Universidad de Zaragoza

<http://webdiis.unizar.es/~ftricas/>

<http://moodle.unizar.es/>
ftricas@unizar.es



Departamento de
Informática e Ingeniería
de Sistemas
Universidad Zaragoza

¿Por qué?

- ▶ Generado automáticamente
- ▶ Típicamente una nueva fuente de datos
- ▶ No diseñado para ser amistoso (no diseñado)
- ▶ Puede ser de poco valor



Diferencias

- ▶ Son datos igual que los pequeños
- ▶ Otras necesidades técnicas (arquitectura, gestión, ...)
- ▶ De ¿Qué datos almacenamos?
A: ¿Qué podemos hacer si tenemos más datos?
- ▶ Mejor datos más diversos que más datos
- ▶ Volumen, variedad, velocidad, ¿veracidad?



3 (+1) V's

40 ZETTAYTES

(43 TRILLION GIGABYTES) of data will be created by 2020, an increase of 300 times from 2005

6 BILLION PEOPLE have cell phones



Volume SCALE OF DATA

It's estimated that **2.5 QUINTILLION BYTES** (2.5 TRILLION GIGABYTES) of data are created each day



Most companies in the U.S. have at least **100 TERABYTES** (100,000 GIGABYTES) of data stored

The New York Stock Exchange captures **1 TB OF TRADE INFORMATION** during each trading session



Velocity ANALYSIS OF STREAMING DATA

Modern cars have close to **100 SENSORS** that monitor items such as fuel level and tire pressure



By 2016, it is projected there will be **18.9 BILLION NETWORK CONNECTIONS** - almost 2.5 connections per person on earth



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015 **4.4 MILLION IT JOBS** will be created globally to support big data, with 1.9 million in the United States



As of 2011, the global size of data in healthcare was estimated to be **150 EXABYTES** (150 BILLION GIGABYTES)



30 BILLION PIECES OF CONTENT are shared on Facebook every month



By 2014, it's anticipated there will be **420 MILLION WEARABLE, WIRELESS HEALTH MONITORS**

Variety DIFFERENT FORMS OF DATA

4 BILLION+ HOURS OF VIDEO are watched on YouTube each month



400 MILLION TWEETS are sent per day by about 200 million monthly active users



1 IN 3 BUSINESS LEADERS don't trust the information they use to make decisions



in one survey were unsure of how much of their data was inaccurate



Veracity UNCERTAINTY OF DATA

Poor data quality costs the US economy around **\$3.1 TRILLION A YEAR**



Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPEEC, QAS

<http://www.ibmbigdatahub.com/infographic/four-vs-big-data>



de Sistemas
Universidad Zaragoza

Volumen

IDC → Universo Digital será de 35 Zetabytes en 2020

Multiples of bytes				V·T·E	
Decimal		Binary			
Value	Metric	Value	IEC	JEDEC	
1000	kB kilobyte	1024	KiB kibibyte	KB kilobyte	
1000 ²	MB megabyte	1024 ²	MiB mebibyte	MB megabyte	
1000 ³	GB gigabyte	1024 ³	GiB gibibyte	GB gigabyte	
1000 ⁴	TB terabyte	1024 ⁴	TiB tebibyte	–	
1000 ⁵	PB petabyte	1024 ⁵	PiB pebibyte	–	
1000 ⁶	EB exabyte	1024 ⁶	EiB exbibyte	–	
1000 ⁷	ZB zettabyte	1024 ⁷	ZiB zebibyte	–	
1000 ⁸	YB yottabyte	1024 ⁸	YiB yobibyte	–	

Orders of magnitude of data

1,000,000,000,000,000,000,000,000

<https://en.wikipedia.org/wiki/Zettabyte>



“The combined space of all computer hard drives in the world was estimated at approximately 160 exabytes in 2006. As of 2009, the entire World Wide Web was estimated to contain close to 500 exabytes. This is one half zettabyte. This has increased rapidly however, as Seagate Technology reported selling a total capacity of 330 exabytes of hard drives during the 2011 Fiscal Year.

<https://en.wikipedia.org/wiki/Zettabyte>



Variedad

- ▶ No sólo números, fechas, cadenas
- ▶ 80 % datos no estructurados (datos geoespaciales, imagen, sonido, vídeo, . . .).
- ▶ Estructura impredecible



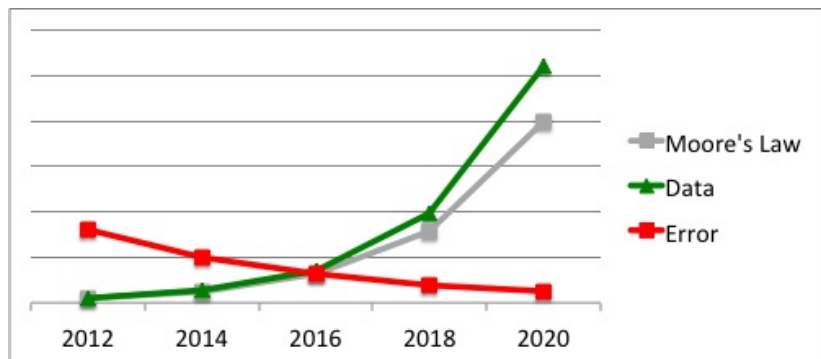
Velocidad

Tiempo real, incluso sin llegar a almacenar

- ▶ Clickstream
- ▶ Operaciones de bolsa, en tiempo real
- ▶ M2M con muchos dispositivos
- ▶ Infraestructura, sensores
- ▶ Juegos en línea



Moore vs big data



[https://amplab.cs.berkeley.edu/
for-big-data-moores-law-means-better-decisions/](https://amplab.cs.berkeley.edu/for-big-data-moores-law-means-better-decisions/)



¿Veracidad?

- ▶ Datos complejos
- ▶ Precisión y calidad poco controlable



Un mundo de V's

- ▶ Valor
- ▶ Validez
- ▶ Volatilidad
- ▶ Variabilidad
- ▶ Viabilidad
- ...



Riesgos

- ▶ Demasiado de todo
- ▶ Coste crece rápido
- ▶ Privacidad (regulaciones y autocontrol)



Beneficios

- ▶ Poder tomar mejores decisiones en el momento adecuado
- ▶ Poder conservar información que aún no sabemos si utilizaremos
- ▶ Acceso a la información independientemente de la forma en que está
- ▶ Beneficio desde el punto de vista de los clientes (ofrecer mejor servicio)
- ▶ Construir un ecosistema mejor de información



Datos

- ▶ Actividad
- ▶ Conversación
- ▶ Fotografía e imagen
- ▶ Sensores
- ▶ IoT



Datos

- ▶ Actividad
- ▶ Conversación
- ▶ Fotografía e imagen
- ▶ Sensores
- ▶ IoT

Y entonces...

- ▶ Selección de fuentes
- ▶ Eliminación de datos redundantes (y ruido)



Aplicaciones

- ▶ Salud
- ▶ Tráfico
- ▶ Seguridad
- ▶ Fabricación
- ▶ Ventas
- ▶ Telecomunicaciones
- ▶ Bolsa
- ▶ Buscadores



Objetivos

- ▶ Modelos predictivos
- ▶ Comportamiento clientes
- ▶ Mejora de procesos
- ▶ Mejora de salud
- ▶ Detección de fraude
- ▶ Urbanismo, ciudades,...

¿más ideas?



¿Qué se hace?

- ▶ Regresión (relaciones)
- ▶ Clasificación
- ▶ Clustering (agrupamiento)
- ▶ Asociación
- ▶ Resumen
- ▶ Detección de anomalías
- ▶ Machine learning // Data mining



Big data Analytics

- ▶ Examinar grandes cantidades de datos
- ▶ Información apropiada
- ▶ Identificación de patrones ocultos, relaciones no conocidas
- ▶ Ventaja competitiva
- ▶ Decisiones de negocio: estratégicas y de operaciones
- ▶ Marquetin
 - ▶ segmentación, Estimación de gasto, análisis de pérdida de clientes, optimización de cartera de productos, recomendaciones, fidelización, descuentos
- ▶ Recursos humanos
 - ▶ identificación/monitorización/retención de talento, formación, abandono



- ▶ Estructurados (DBRM, Tablas)
- ▶ Semi-estructurados (XML, json)
- ▶ No estructurados (texto, imágenes, vídeo)
 - ▶ Datos no estructurados + metadatos



Tablas

CUSTOMER		
NAME	DATATYPE	NULLABLE?
CUSTOMER_ID	VARCHAR	NO
FIRST_NAME	VARCHAR	NO
LAST_NAME	VARCHAR	NO
BIRTH_DAY	TIMESTAMP	NO
ADDRESS	VARCHAR	NO
ADDRESS2	VARCHAR	YES
STATE	VARCHAR	NO
ZIP_CODE	INTEGER	NO

CUST_ORDER		
NAME	DATATYPE	NULLABLE?
ORDER_ID	VARCHAR	NO
CUSTOMER_ID	VARCHAR	NO
STATUS	VARCHAR	NO
ORDER_AMOUNT	DECIMAL	NO

PRODUCT		
NAME	DATATYPE	NULLABLE?
PRODUCT_ID	VARCHAR	NO
CATEGORY	VARCHAR	NO
LIST_PRICE	DECIMAL	NO

https://docs.oracle.com/cd/E13167_01/aldsp/docs21/xquery/sql_pushdown.html



► XML (Extensible Markup Language)

```

<?xml version="1.0" encoding="UTF-8" ?>
<!DOCTYPE Edit_Mensaje SYSTEM "Edit_Mensaje.dtd">

<Edit_Mensaje>
  <Mensaje>
    <Remitente>
      <Nombre>Nombre del remitente</Nombre>
      <Mail> Correo del remitente </Mail>
    </Remitente>
    <Destinatario>
      <Nombre>Nombre del destinatario</Nombre>
      <Mail>Correo del destinatario</Mail>
    </Destinatario>
    <Texto>
      <Asunto>
        Este es mi documento con una estructura muy sencilla
        no contiene atributos ni entidades...
      </Asunto>
      <Parrafo>
        Este es mi documento con una estructura muy sencilla
        no contiene atributos ni entidades...
      </Parrafo>
    </Texto>
  </Mensaje>
</Edit_Mensaje>

```

Y otros ...

▶ JSON (JavaScript Object Notation)

```
{
  "firstName": "John",
  "lastName": "Smith",
  "age": 25,
  "address": {
    "streetAddress": "21 2nd Street",
    "city": "New York",
    "state": "NY",
    "postalCode": "10021"
  },
  "phoneNumber": [
    { "type": "home", "number": "212 555-1234" },
    { "type": "fax", "number": "646 555-4567" }
  ]
}
```

▶ CSV (Comma Separated Values)

```
1997,Ford,E350,"ac, abs, moon",3000.00
1999,Chevy,"Venture ""Extended Edition""",4900.00
1999,Chevy,"Venture ""Extended Edition, Very Large""",5000.00
1996,Jeep,Grand Cherokee,"MUST SELL!
air, moon roof, loaded",4799.00
```

(Hay más)



Datos y su interpretación

Datos con código (?)

- ▶ Middleware (presentar los datos según las necesidades)
- ▶ Conectar y extraer datos del almacenamiento
- ▶ Transformar los datos
- ▶ Subdividirlos para su procesado



Infraestructura

- ▶ Servidores distribuidos/nube
- ▶ Almacenamiento distribuido
- ▶ Procesamiento distribuido (MapReduce, Hadoop)
- ▶ Bases de datos especializadas (menos estructura, más prestaciones)
- ▶ Interpretación de los datos (semántica)



Tecnologías



Almacenamiento

HDFS (Hadoop Distributed File System)

- ▶ Grandes ficheros divididos en trozos
- ▶ Se mueven partes de los ficheros al clúster
- ▶ Tolerancia a fallos mediante replicación
- ▶ Registro mediante NameNode (metadata), acceso mediante DataNode (data)
- ▶ Escribe una vez, utiliza varias

Otros: Ceph, Swift, Dispersed Storage Network (Cleversafe), GPFS (IBM), Isilon (EMC), Lustre, MapR File System



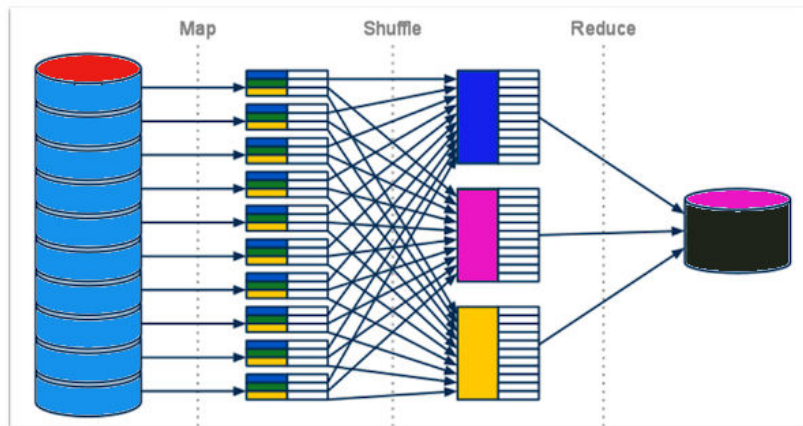
MapReduce

- ▶ Algoritmos próximos a los datos
- ▶ Datos/algoritmos preparados para la paralelización
- ▶ Commodity computing
- ▶ Simplicidad:
 - ▶ Fase Map (de los datos, a una lista de clave, valor)
 - ▶ Fase Reduce (agrupar datos con la misma clave)

Origen: multiplicaciones de grandes matrices para cálculo del *PageRank*



MapReduce



<http://hadoopproject.com/mapreduce-projects/>



MapReduce

map, filter, and reduce
explained with emoji 😂

```
map([🐮, 🍷, 🐔, 🌽], cook)  
=> [🍔, 🍟, 🍗, 🍿]
```

```
filter([🍔, 🍟, 🍗, 🍿], isVegetarian)  
=> [🍟, 🍿]
```

```
reduce([🍔, 🍟, 🍗, 🍿], eat)  
=> 🤩
```

https://www.reddit.com/r/ProgrammerHumor/comments/5rf9xf/map_filter_and_reduce_explained/

https:

[//css-tricks.com/an-illustrated-and-musical-guide-to-map-reduce-and-filter-array-methods/](https://css-tricks.com/an-illustrated-and-musical-guide-to-map-reduce-and-filter-array-methods/)



Not Only SQL

También: *non-relational*



Structured Query Language

- ▶ Tablas
- ▶ Estructura
- ▶ Vistas, uniones, ...
- ▶ Índices, consistencia, transacciones, búsqueda,...



ACID

- ▶ Atomicity
- ▶ Consistency
- ▶ Isolation (entre operaciones)
- ▶ Durability



- ▶ Tecnología antigua (1960's) (anterior a RDBMS)
 - ▶ Nombre del siglo XXI (Google, Amazon, Facebook, Twitter, ... web 2.0)
- ▶ Ficheros secuenciales
- ▶ BD jerárquica
- ▶ Base de datos en red
- ▶ Distribuida
- ▶ Simplicidad, escalabilidad horizontal
- ▶ Consistencia eventual (disponibilidad, tolerancia a la partición, velocidad ...)

- ▶ Columnas
- ▶ Documentos
- ▶ Grafos
- ▶ clave, valor
- ▶ Multi-modelo

noSQL. Columnas

Key	Driver Information	Car Information
123546	Name:John Insurance: Geico	Car: Speed3 Year:2013 Warranty:Yes
123547	Name:Jen Insurance:State Farm	Car:626 Year:2008
123548	Name:Tony	

Key	Car Information
123546	Car: Speed3 Year:2013 Warranty:Yes ServiceID:10 Type:Oil ServiceID:11 Type:Tires ServiceID:12 Type:Wipers

Key	Name	Insurance	Car	Year	Warranty	ServiceID	Type	Key
123456	John	Geico	Speed3	2013	Yes	10	Oil	123456
123457	Jen	State Farm	626	2008	NULL	11	Tires	123456
123458	Tony	NULL	NULL	NULL	NULL	12	Wipers	123456

<http://www.ingenioussql.com/2013/02/28/rules-of-engagement-nosql-column-data-stores/>



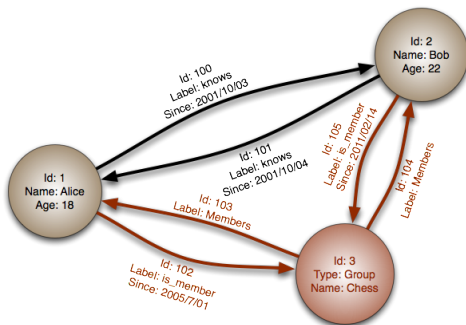
noSQL. Documentos

- ▶ Información semi-estructurada
- ▶ Metadatos
- ▶ Organización
 - ▶ Colecciones
 - ▶ Etiquetas
 - ▶ Directorios
 - ...

mongoDB, couchDB



- ▶ Estructuras de grafos para *queries* semánticas
 - ▶ Nodos (entidades: gente, negocios, cuentas, ...)
 - ▶ Propiedades
 - ▶ Arcos (Conectan nodos entre sí o nodos con propiedades)



Neo4j



noSQL. Clave-valor

Key	Value
K1	AAA,BBB,CCC
K2	AAA,BBB
K3	AAA,DDD
K4	AAA,2,01/01/2015
K5	3,ZZZ,5623

https://en.wikipedia.org/wiki/Key-value_database#/media/File:KeyValue.PNG

Cassandra



Departamento de
Informática e Ingeniería
de Sistemas
Universidad Zaragoza

noSQL. Clave-valor

Key	Value
K1	AAA,BBB,CCC
K2	AAA,BBB
K3	AAA,DDD
K4	AAA,2,01/01/2015
K5	3,ZZZ,5623

- ▶ Cada clave puede tener asociados datos de diferente tipo (no definido)
- ▶ más flexible
- ▶ A veces menos espacio y más prestaciones

Niveles

- ▶ Batch layer
 - ▶ cálculos arbitrarios
 - ▶ escalable horizontalmente
 - ▶ mayor latencia
 - ▶ Map/Reduce
 - ▶ Sólo añadir (copia maestra)
- ▶ Speed layer
 - ▶ Para compensar la alta latencia del otro
 - ▶ Algoritmos incrementales
 - ▶ Horas de datos en lugar de ...



Niveles

- ▶ Batch layer
 - ▶ cálculos arbitrarios
 - ▶ escalable horizontalmente
 - ▶ mayor latencia
 - ▶ Map/Reduce
 - ▶ Sólo añadir (copia maestra)
- ▶ Speed layer
 - ▶ Para compensar la alta latencia del otro
 - ▶ Algoritmos incrementales
 - ▶ Horas de datos en lugar de ...
- ▶ Serving layer (resultados)



Fuentes de datos

- ▶ Social network profiles
- ▶ Social influencers (reseñas, análisis, ...)
- ▶ Activity-generated data
- ▶ Software as a Service (SaaS) and cloud applications
- ▶ Public (open source intelligence)
- ▶ Hadoop MapReduce application results
- ▶ Data warehouse appliances
- ▶ Columnar/NoSQL data sources
- ▶ Network and in-stream monitoring technologies
- ▶ Legacy documents

<http://www.zdnet.com/article/top-10-categories-for-big-data-sources-and-mining-technologies/>



Referencias

<http://www.slideshare.net/nasrinhussain1/big-data-ppt-31616290>

http:

[//www.slideshare.net/BernardMarr/140228-big-data-slide-share/](http://www.slideshare.net/BernardMarr/140228-big-data-slide-share/)

<http://www.slideshare.net/outerthought/big-data>

<http://www.slideshare.net/PhilippeJulio/hadoop-architecture/>

<http://www.slideshare.net/zanorte/big-data-para-dummies>

<http://www.slideshare.net/nasrinhussain1/big-data-ppt-31616290>

http:

[//www.slideshare.net/BernardMarr/140228-big-data-slide-share/](http://www.slideshare.net/BernardMarr/140228-big-data-slide-share/)

<http://www.slideshare.net/outerthought/big-data>

<http://www.slideshare.net/PhilippeJulio/hadoop-architecture/>

<http://www.slideshare.net/zanorte/big-data-para-dummies>



Y más cosas....

- ▶ No hemos hablado de visualización
- ▶ No hemos hablado de análisis de redes sociales (SNA)

