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

¿Por qué?

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

Departamento de Informática e Ingeniería de Sistemas
Universidad Zaragoza

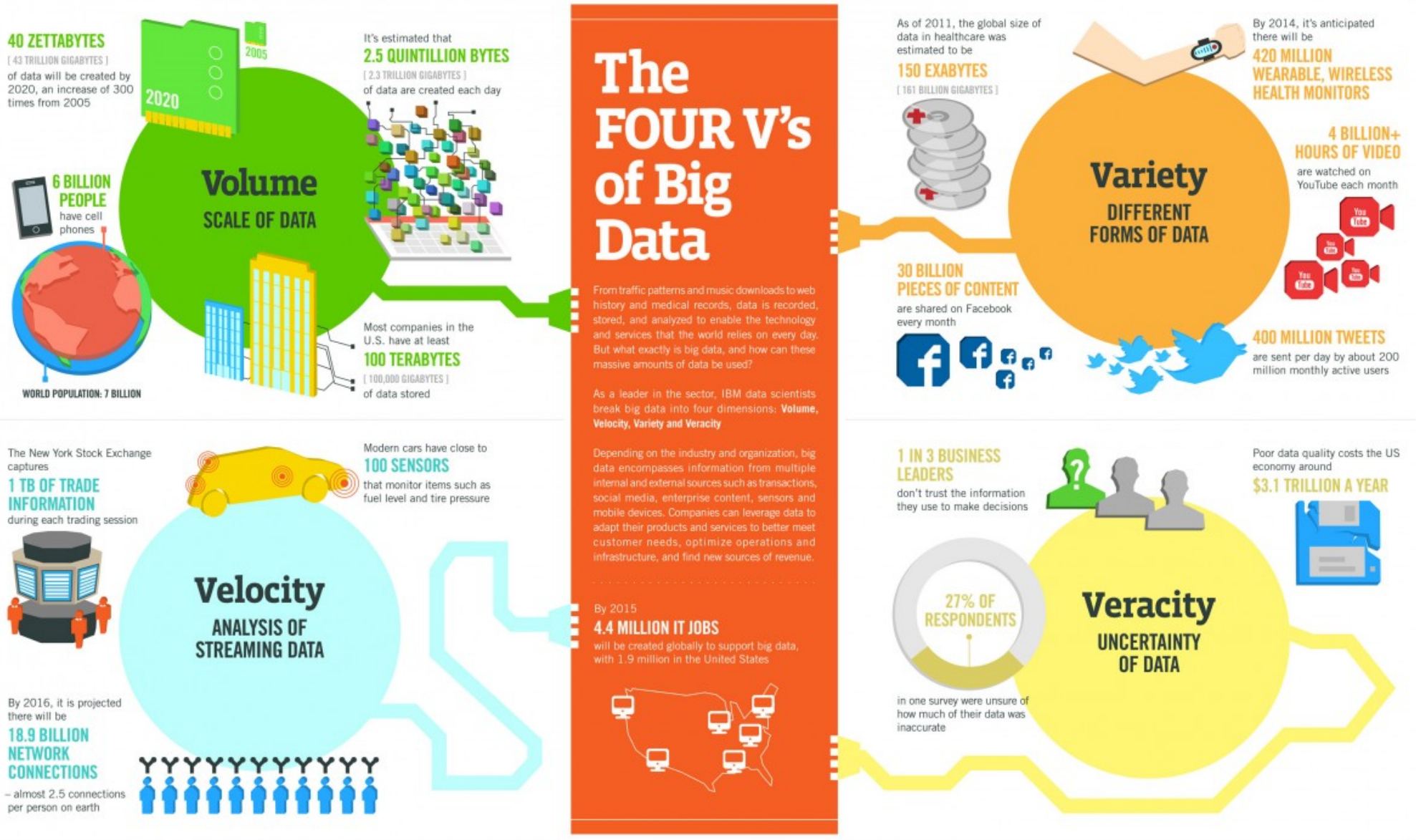
Diferencias

- 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?

Departamento de Informática e Ingeniería de Sistemas
Universidad Zaragoza

3 (+1) V's



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

Volumen

IDC $\longrightarrow$ 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

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

Departamento de Informática e Ingeniería de Sistemas
Universidad Zaragoza

Volumen

> *"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, texto,...).
- Estructura impredecible

Departamento de Informática e Ingeniería de Sistemas
Universidad Zaragoza

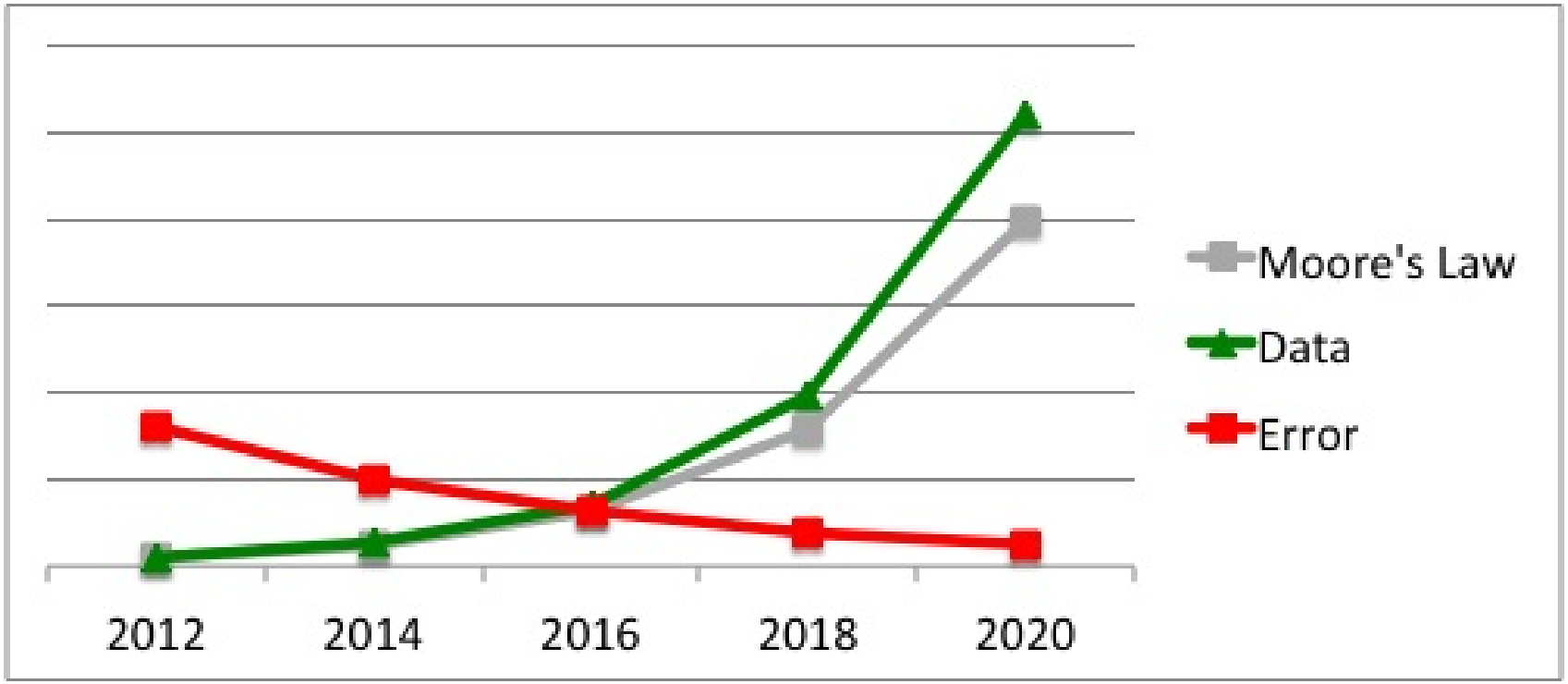
Velocidad

Tiempo real, incluso sin llegar a almacenar

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

Departamento de Informática e Ingeniería de Sistemas
Universidad Zaragoza

Moore vs big data



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

...

Departamento de Informática e Ingeniería de Sistemas
Universidad Zaragoza

Riesgos

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

Departamento de Informática e Ingeniería de Sistemas
Universidad Zaragoza

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

Departamento de Informática e Ingeniería de Sistemas
Universidad Zaragoza

Datos

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

Y entonces...

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

Departamento de Informática e Ingeniería de Sistemas
Universidad Zaragoza

Aplicaciones

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

Departamento de Informática e Ingeniería de Sistemas
Universidad Zaragoza

Objetivos

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

¿más ideas?

Departamento de Informática e Ingeniería de Sistemas
Universidad Zaragoza

¿Qué se hace?

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

Departamento de Informática e Ingeniería de Sistemas
Universidad Zaragoza

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

Datos

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

Departamento de Informática e Ingeniería de Sistemas
Universidad Zaragoza

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

- XML (**E**xtensible **M**arkup **L**anguage)

```
<?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)

Departamento de Informática e Ingeniería de Sistemas
Universidad Zaragoza

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

Departamento de Informática e Ingeniería de Sistemas
Universidad Zaragoza

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

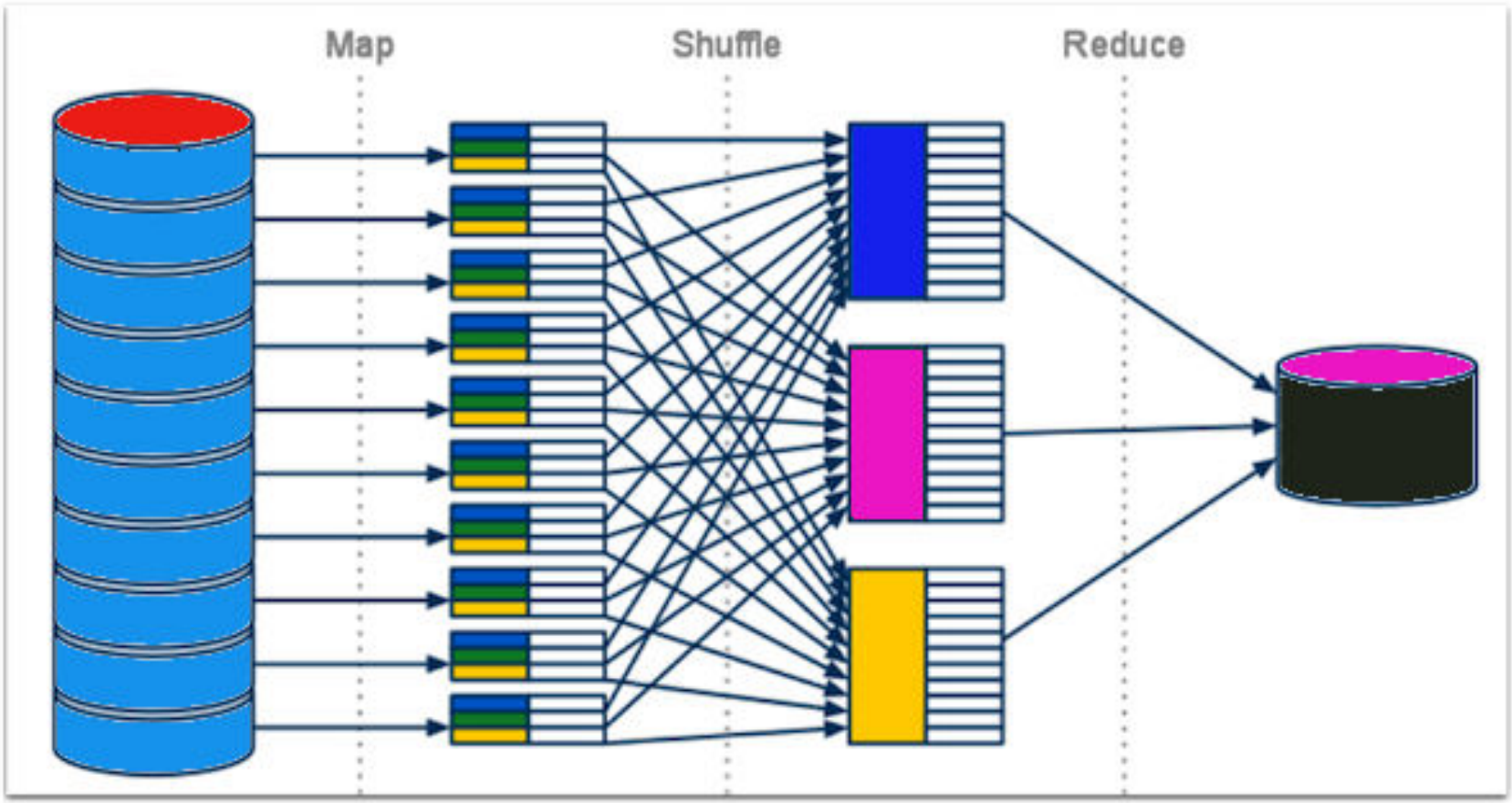
Otros: Ceph, Swift, Dispersed Storage Network (Cleversafe), GPFS (IBM), Isilon (EMC), Lustre, MapR File System

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)

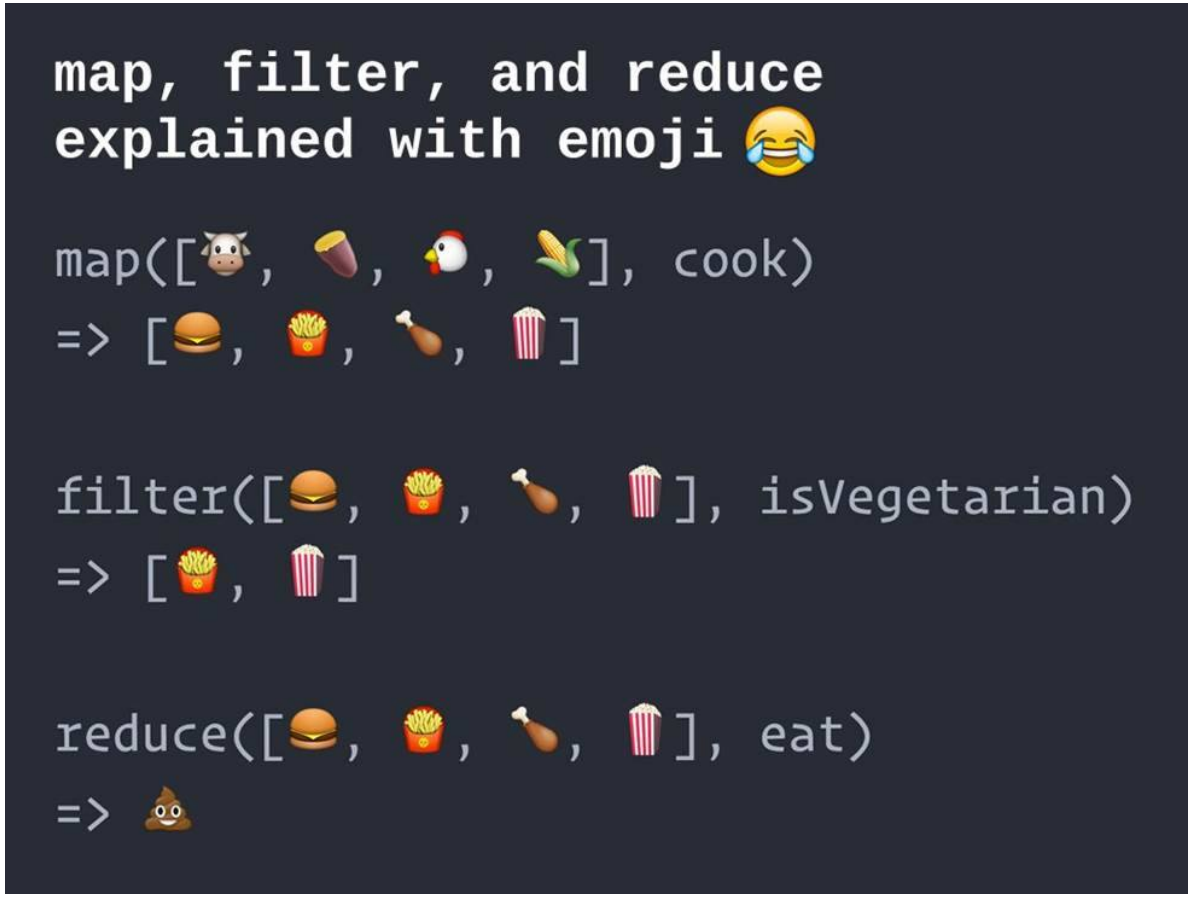
Origen: multiplicaciones de grandes matrices para cálculo del *PageRank*

MapReduce



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

MapReduce



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/

Not Only SQL
También: *non-relational*

Departamento de Informática e Ingeniería de Sistemas
Universidad Zaragoza

SQL

Structured Query Language

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

Departamento de Informática e Ingeniería de Sistemas
Universidad Zaragoza

ACID

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

Departamento de Informática e Ingeniería de Sistemas
Universidad Zaragoza

noSQL

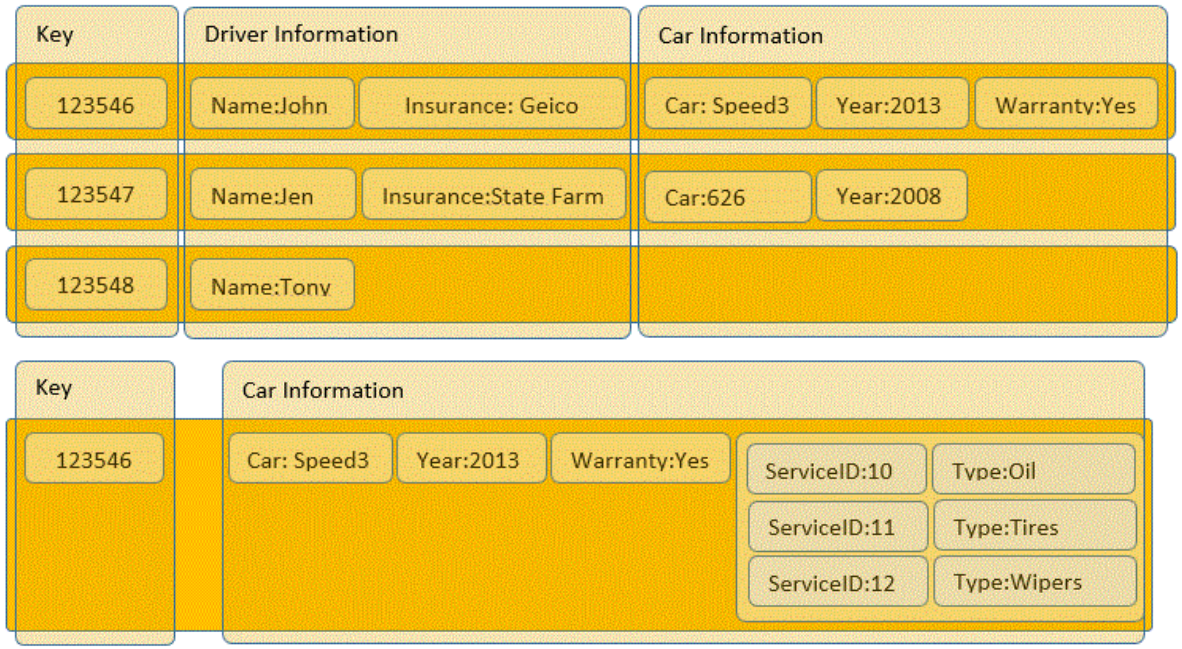
- Tecnología antigua (1960's) (anterior a RDBMS)
  - Nombre del siglo XXI (Google, Amazon, Facebook, Twitter, ... web 2.0)
- Ficheros secuenciales
- BD jerárquica
- Base de datos en red
- Distribuida
- Simplicidad, escalabilidad horizontal
- Consistencia eventual (disponibilidad, tolerancia a la partición,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

noSQL

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

Departamento de Informática e Ingeniería de Sistemas
Universidad Zaragoza

noSQL. Columnas



Key	Name	Insurance	Car	Year	Warranty
123456	John	Geico	Speed3	2013	Yes
123457	Jen	State Farm	626	2008	NULL
123458	Tony	NULL	NULL	NULL	NULL

ServiceID	Type	Key
10	Oil	123456
11	Tires	123456
12	Wipers	12345

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

noSQL. Documentos

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

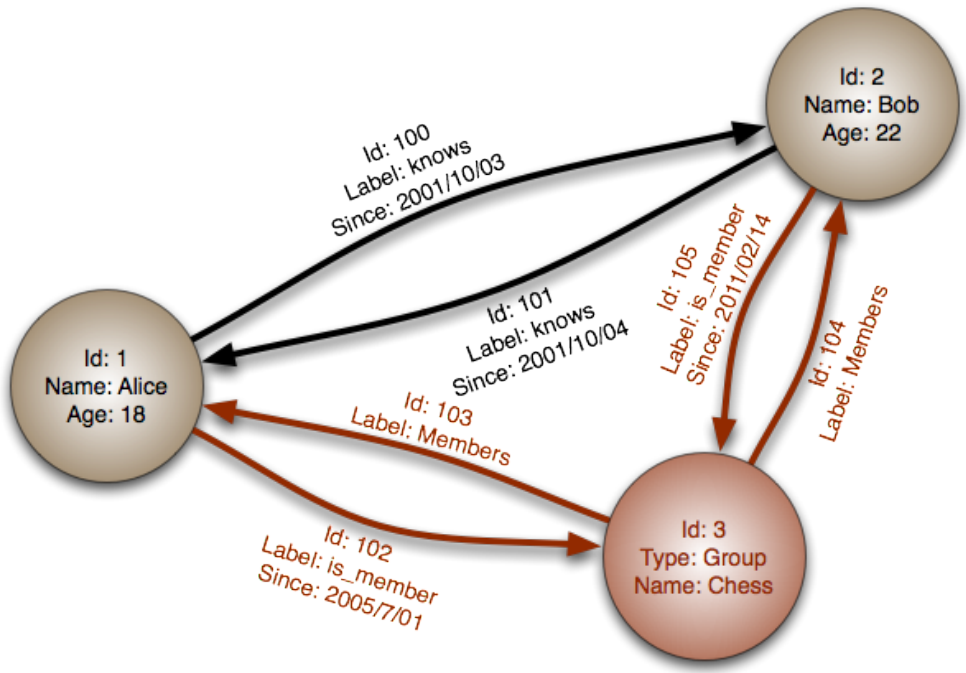
  ...

mongoDB, couchDB

Departamento de Informática e Ingeniería de Sistemas
Universidad Zaragoza

noSQL. Grafos

- 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

Departamento de Informática e Ingeniería de Sistemas
Universidad Zaragoza

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 ...

Departamento de Informática e Ingeniería de Sistemas
Universidad Zaragoza

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/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/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)