

PROGRAMACIÓN SOBRE GRANDES VOLUMENES DE DATOS

BIG DATA

Magister - Efraín Alberto Oviedo
eaoc46@gmail.com

**UNIVERSIDAD DE ANTIOQUIA
FACULTAD DE INGENIERÍA**

ESPECIALIZACIÓN EN ANALÍTICA Y CIENCIA DE DATOS



**UNIVERSIDAD
DE ANTIOQUIA**

1 8 0 3

AGENDA

1. Introducción

- 2. Evolución de la ciencia de datos
- 3. Datos
- 4. Aplicaciones

¿QUÉ ES BIG DATA?



¿QUÉ ES BIG DATA?

“Conjunto de **técnicas** que permiten **analizar, procesar y gestionar** conjuntos de datos **extremadamente grandes** que pueden ser analizados informáticamente para **revelar patrones, tendencias y asociaciones**, especialmente en relación con la conducta humana y las interacciones con los usuarios”

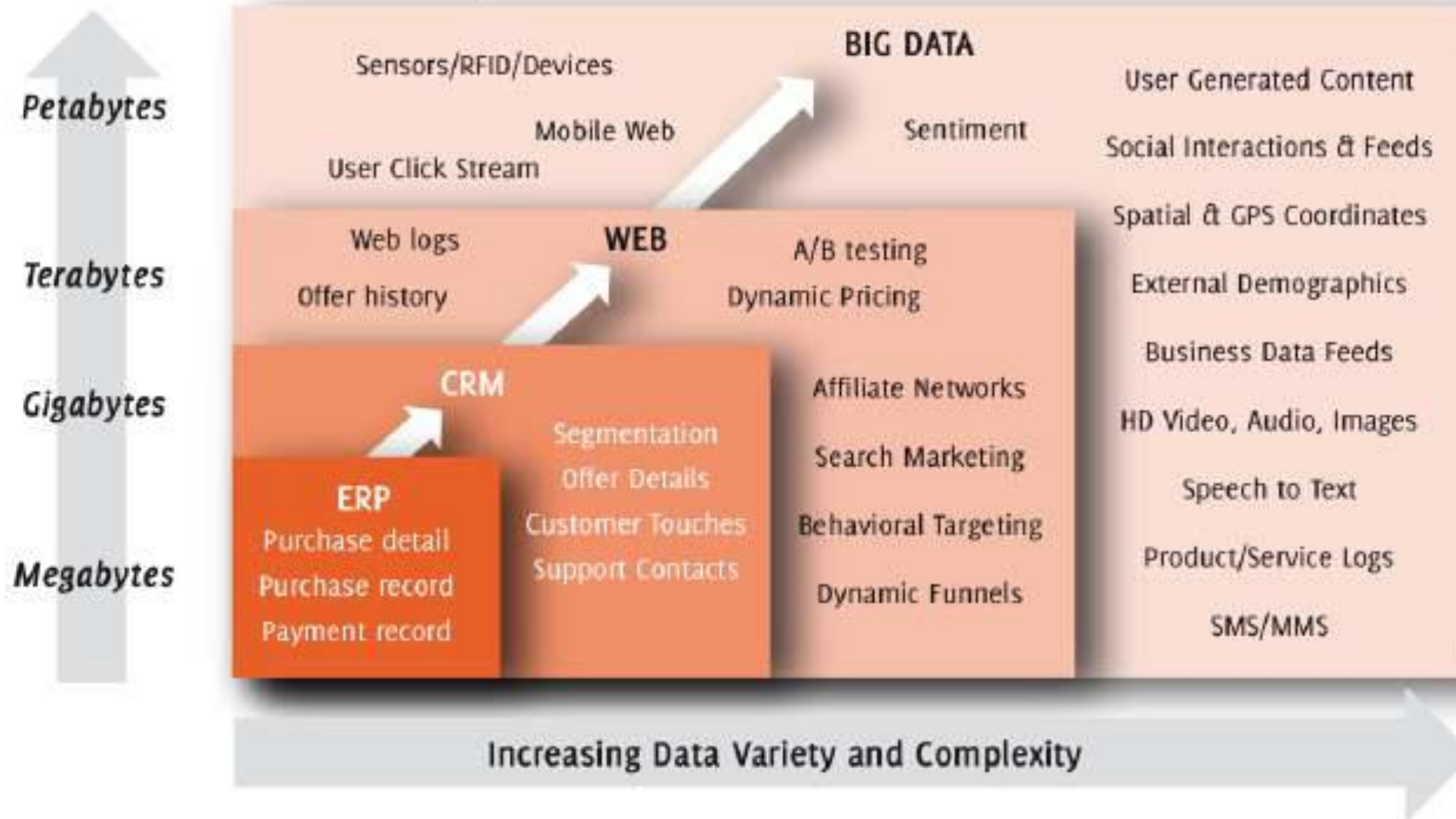


REAL ACADEMIA ESPAÑOLA

Unidades de Medidas de Almacenamiento

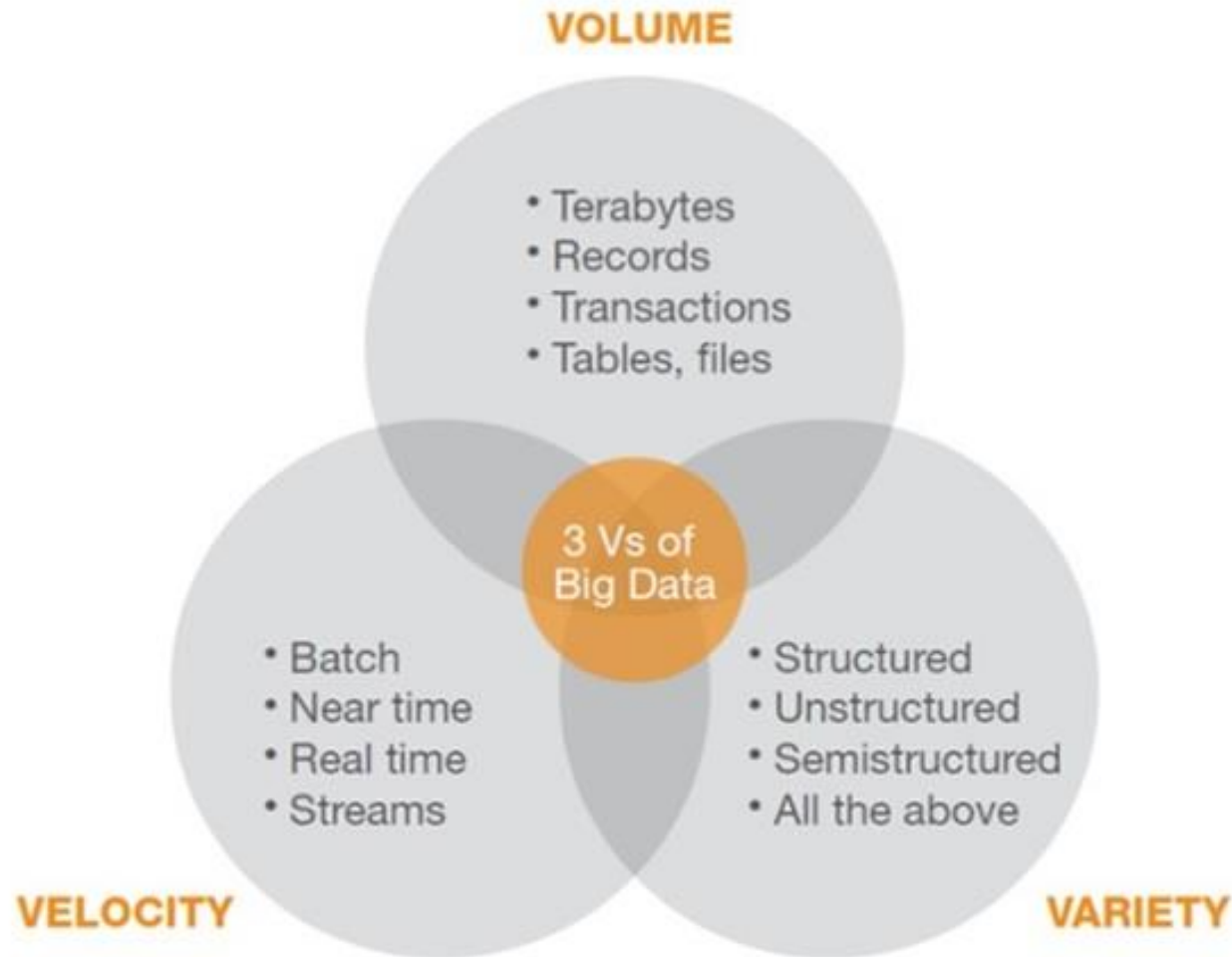
Medida	Simbologia	Equivalencia	Equivalente en Bytes
byte	b	8 bits	1 byte
kilobyte	Kb	1024 bytes	1 024 bytes
megabyte	MB	1024 KB	1 048 576 bytes
gigabyte	GB	1024 MB	1 073 741 824 bytes
terabyte	TB	1024 GB	1 099 511 627 776 bytes
Petabyte	PB	1024 TB	1 125 899 906 842 624 bytes
Exabyte	EB	1024 PB	1 152 921 504 606 846 976 bytes
Zetabyte	ZB	1024 EB	1 180 591 620 717 411 303 424 bytes
Yottabyte	YB	1024 ZB	1 208 925 819 614 629 174 706 176 bytes
Brontobyte	BB	1024 YB	1 237 940 039 285 380 274 899 124 224 bytes
Geopbyte	GB	1024 BB	1 267 650 600 228 229 401 496 703 205 376 bytes

Big Data = Transactions + Interactions + Observations



Source: Contents of above graphic created in partnership with Teradata, Inc.

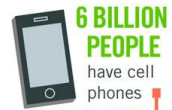
Las Vs del Big Data



40 ZETTABYTES

[43 TRILLION GIGABYTES]

of data will be created by 2020, an increase of 300 times from 2005



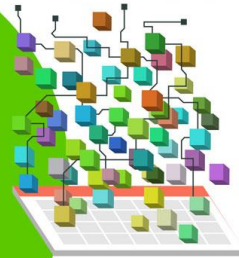
6 BILLION PEOPLE have cell phones

WORLD POPULATION: 7 BILLION

Volume SCALE OF DATA

It's estimated that
2.5 QUINTILLION BYTES

[2.3 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 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

[161 BILLION GIGABYTES]



**30 BILLION
PIECES OF CONTENT**

are shared on Facebook every month



Variety DIFFERENT FORMS OF DATA



By 2014, it's anticipated there will be

420 MILLION WEARABLE, WIRELESS HEALTH MONITORS

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



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



Veracity UNCERTAINTY OF DATA



Poor data quality costs the US economy around

\$3.1 TRILLION A YEAR



1 IN 3 BUSINESS LEADERS

don't trust the information they use to make decisions



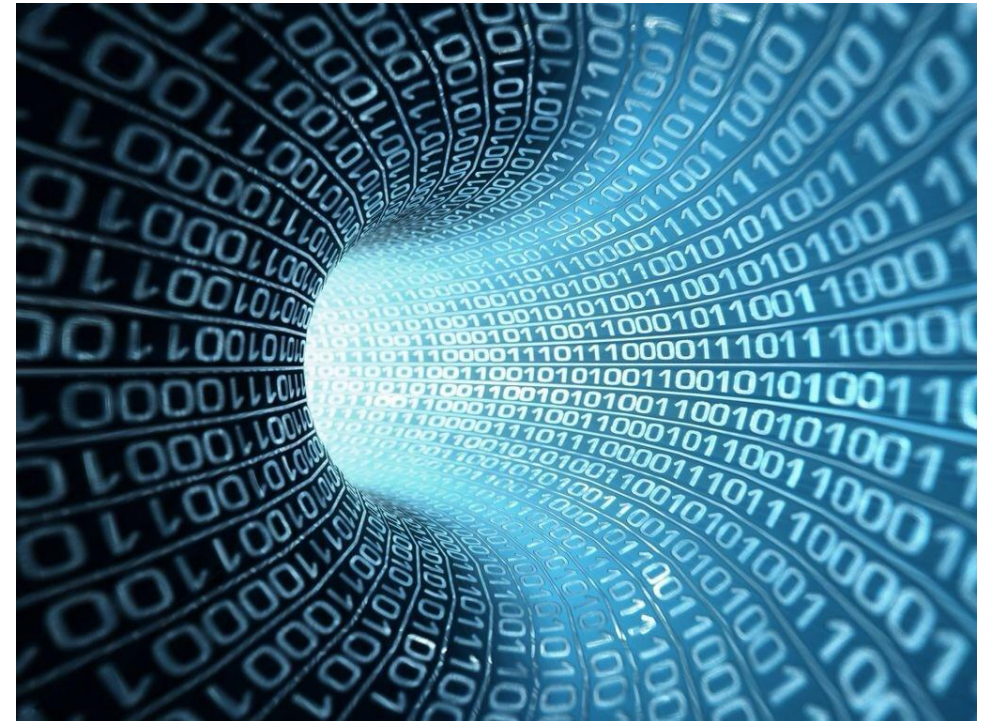
in one survey were unsure of how much of their data was inaccurate

Las Vs del Big Data

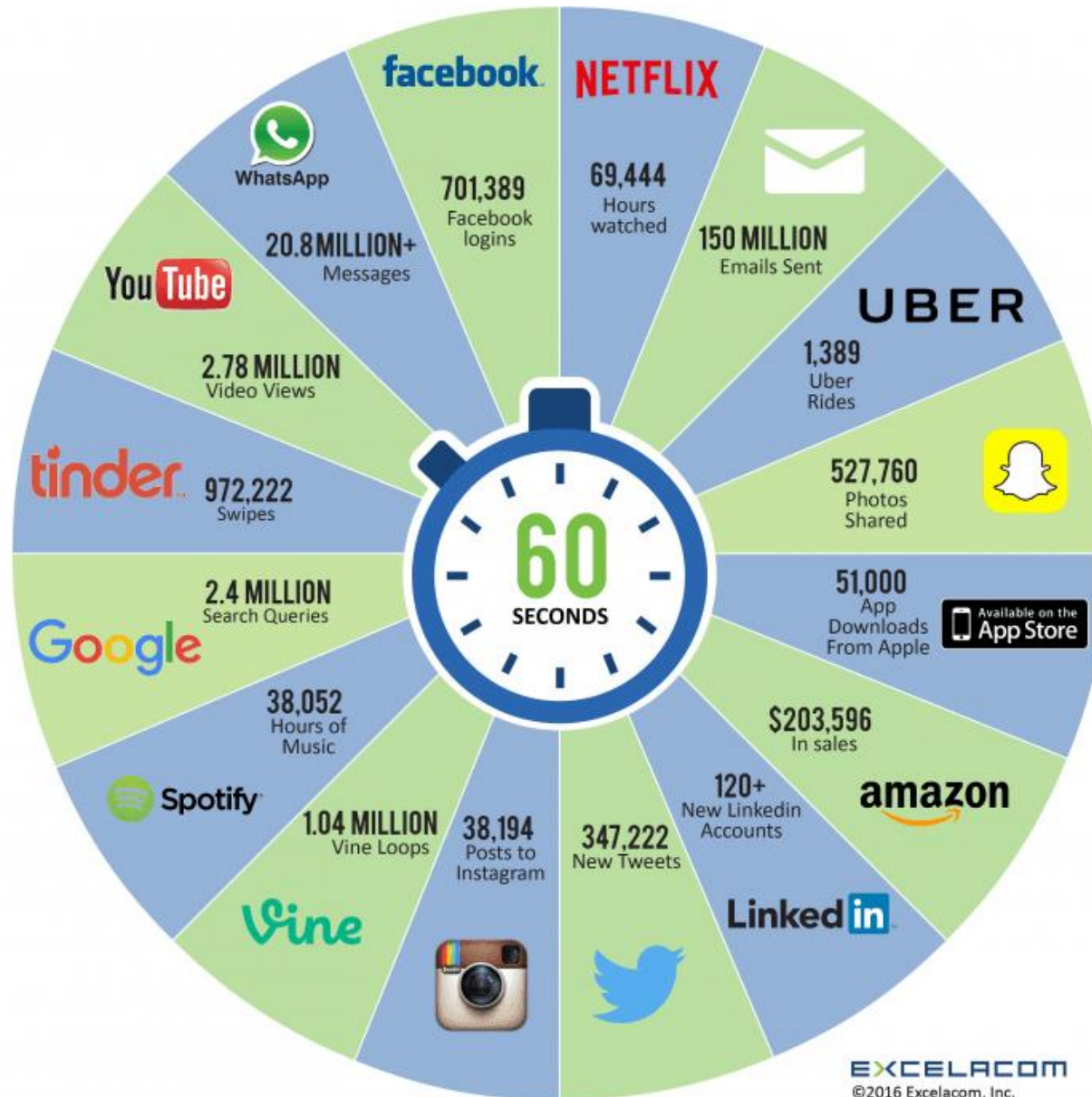


¿DE DONDE SALEN LOS DATOS?

- Redes sociales y Aplicaciones
- IoT



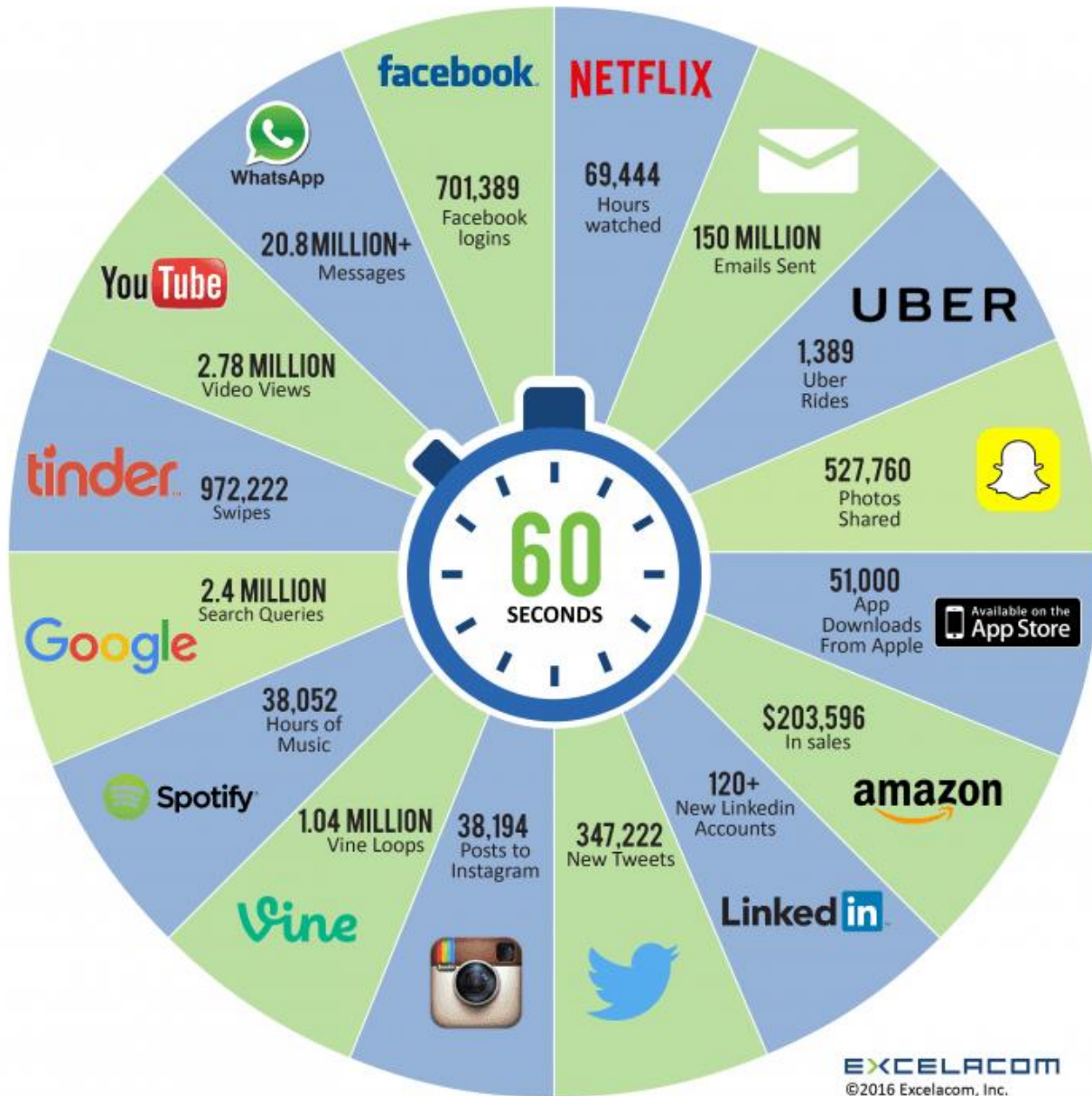
2016 What happens in an INTERNET MINUTE?



<https://www.trecebits.com/2019/04/03/minuto-internet-infografia/>

<https://www.reasonwhy.es/actualidad/digital/esto-es-lo-que-pasa-en-internet-en-un-minuto-2016-2016-05-04>

2016 What happens in an INTERNET MINUTE?



2019 This Is What Happens In An Internet Minute

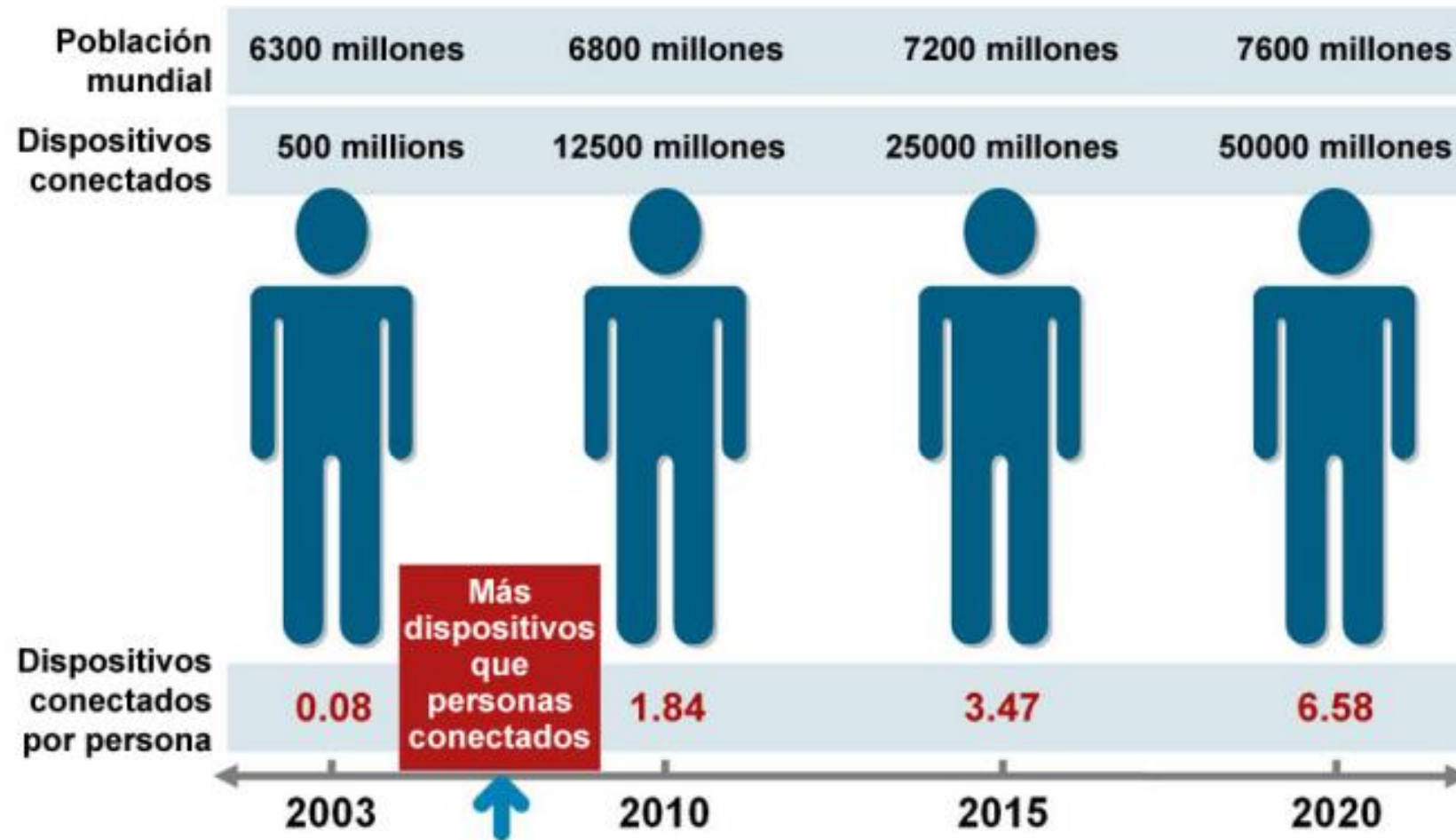


IoT (Internet Of Things)

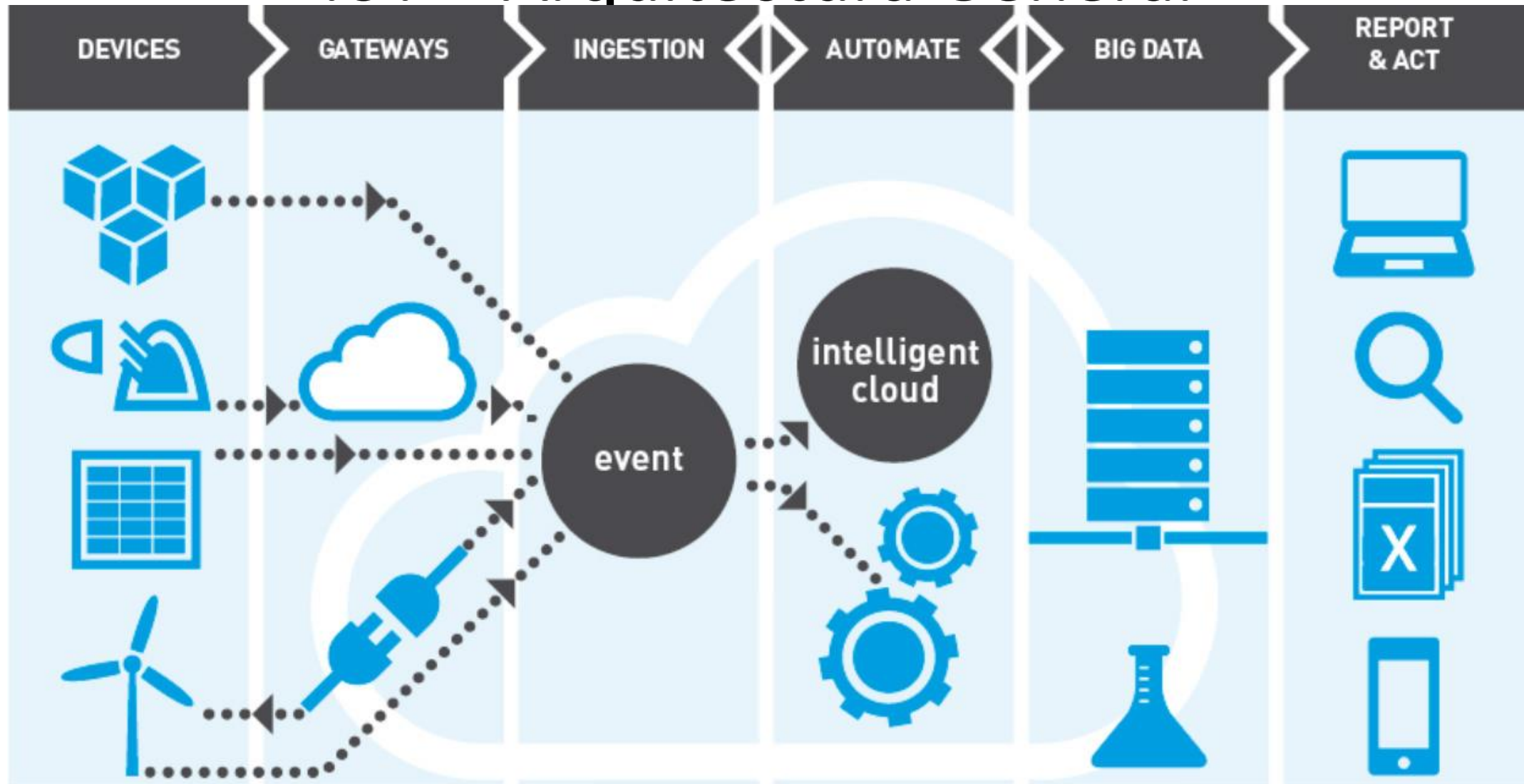
Tres letras que lo están cambiando todo

- Antes: Los seres humanos eran quienes capturaban la información y la subían a internet
- Ahora: Las cosas (dispositivos instalados sobre cualquier objeto físico, con capacidad de medición y comunicación) se encargan de capturar y subir la información a internet

IoT - Dispositivos



IoT – Arquitectura General



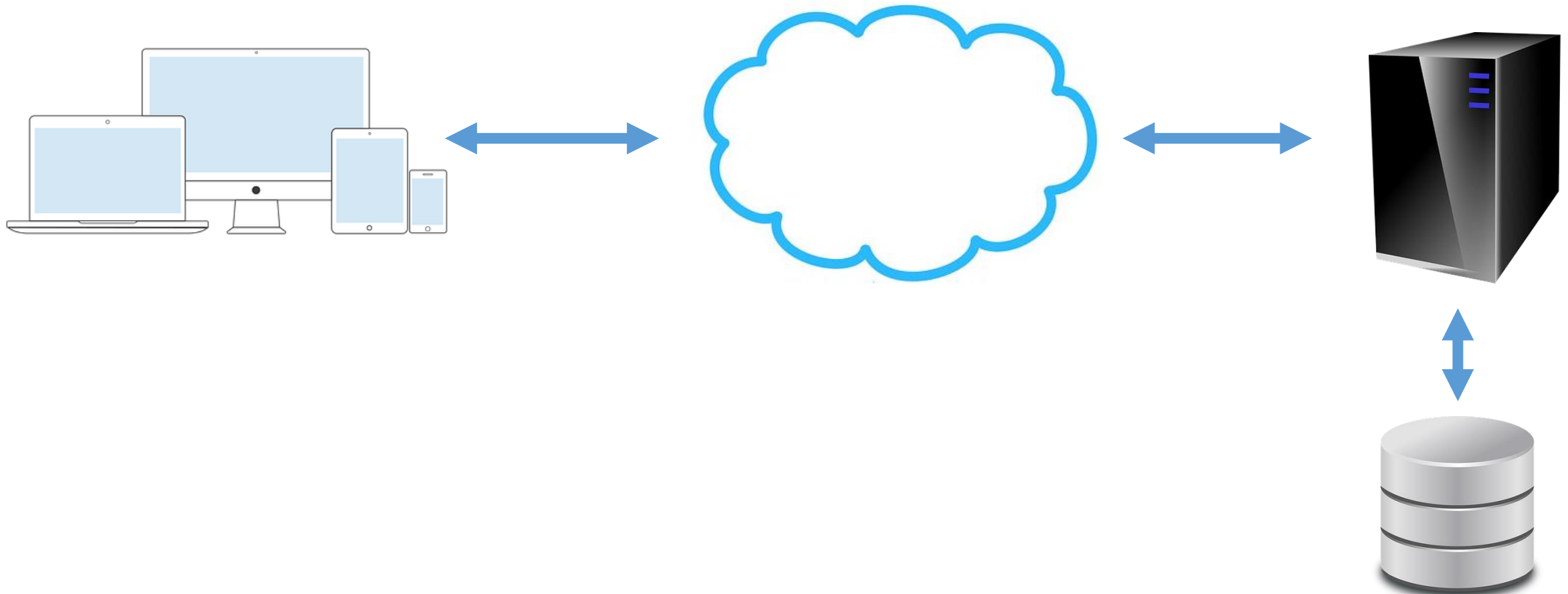
AGENDA

1. Introducción
- 2. Evolución de la ciencia de datos**
3. Datos
4. Aplicaciones

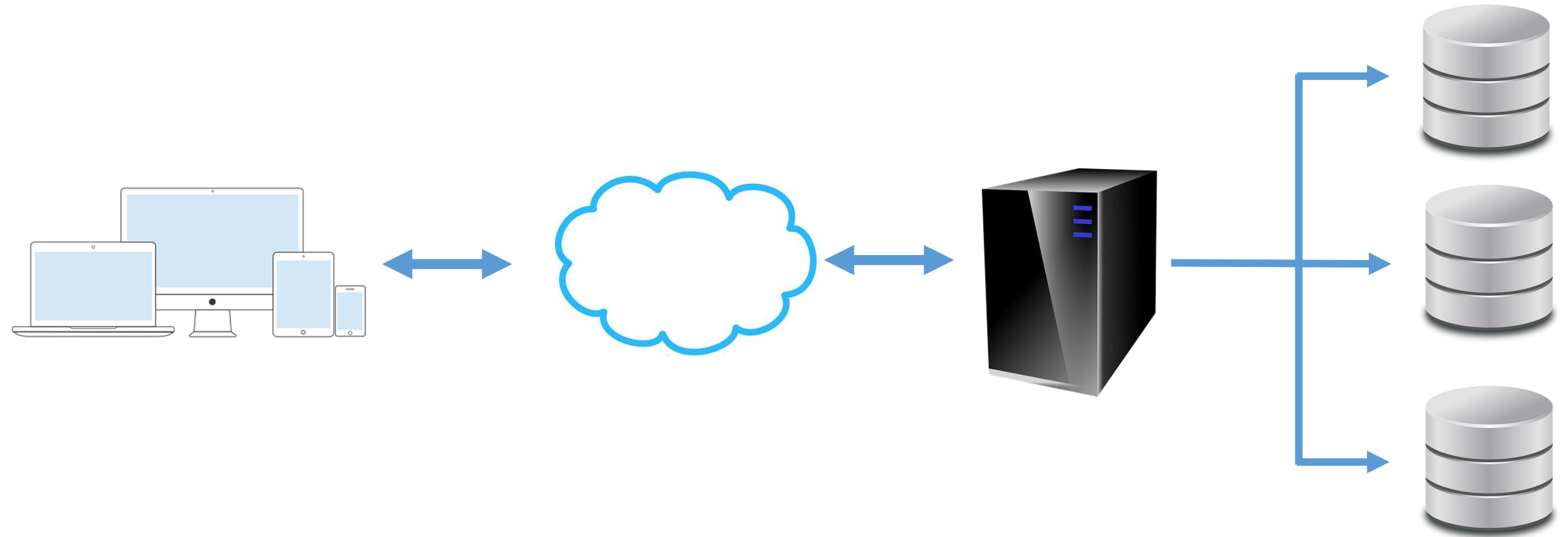
Arquitecturas de Sistemas de Información

- Sistemas Centralizados
- Sistemas de Almacenamiento Distribuido
- Sistemas de Procesamiento Distribuido
- Sistemas de Almacenamiento y procesamiento distribuido

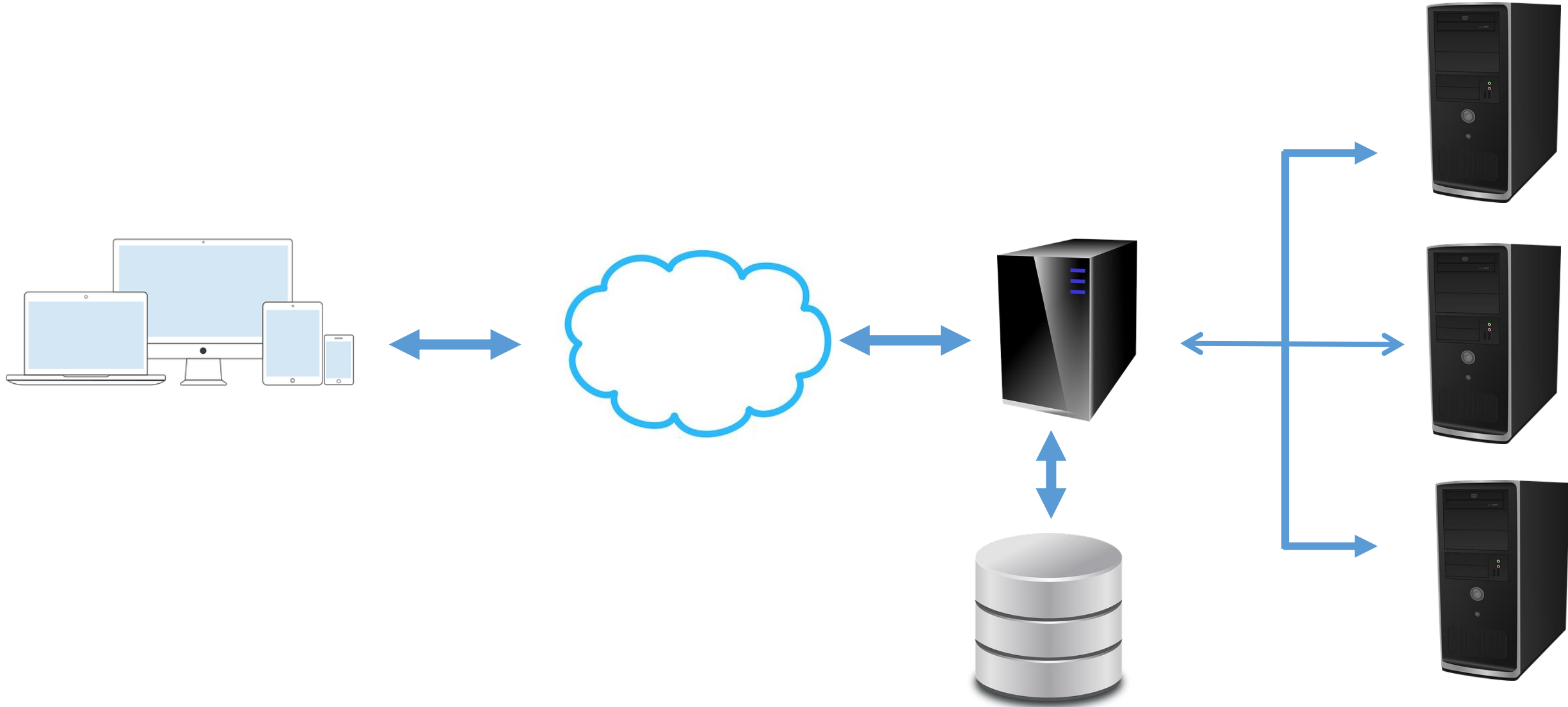
Sistemas Centralizados



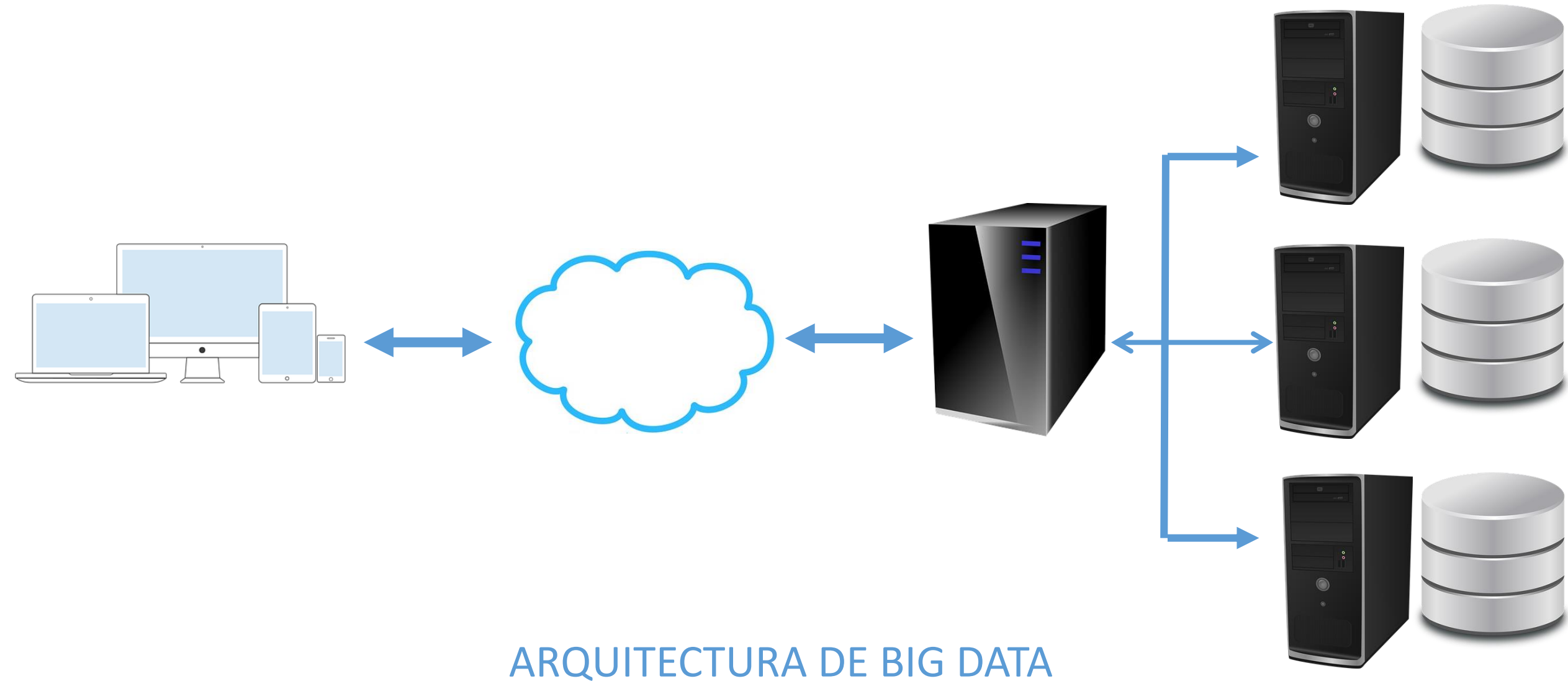
Almacenamiento Distribuido



Procesamiento Distribuido



Almacenamiento y Procesamiento Distribuido

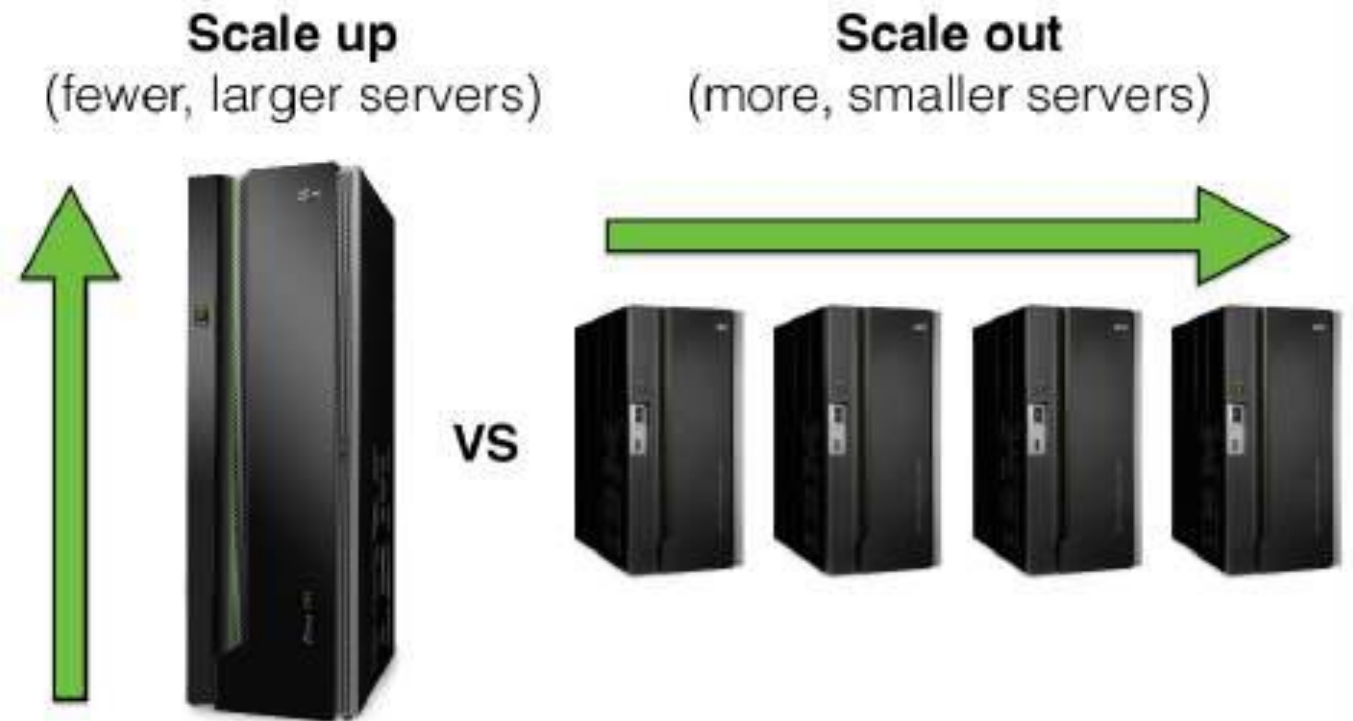


Escalabilidad

Añadir capacidad a un sistema para garantizar el servicio a los usuarios

Vertical: Añadir recursos a una máquina

Horizontal: Añadir una o varias máquinas mas



Escalabilidad

Ley de Amdahl: Representa de forma matemática la influencia de una mejora de uno o varios componentes de una computadora en el rendimiento global de la misma.

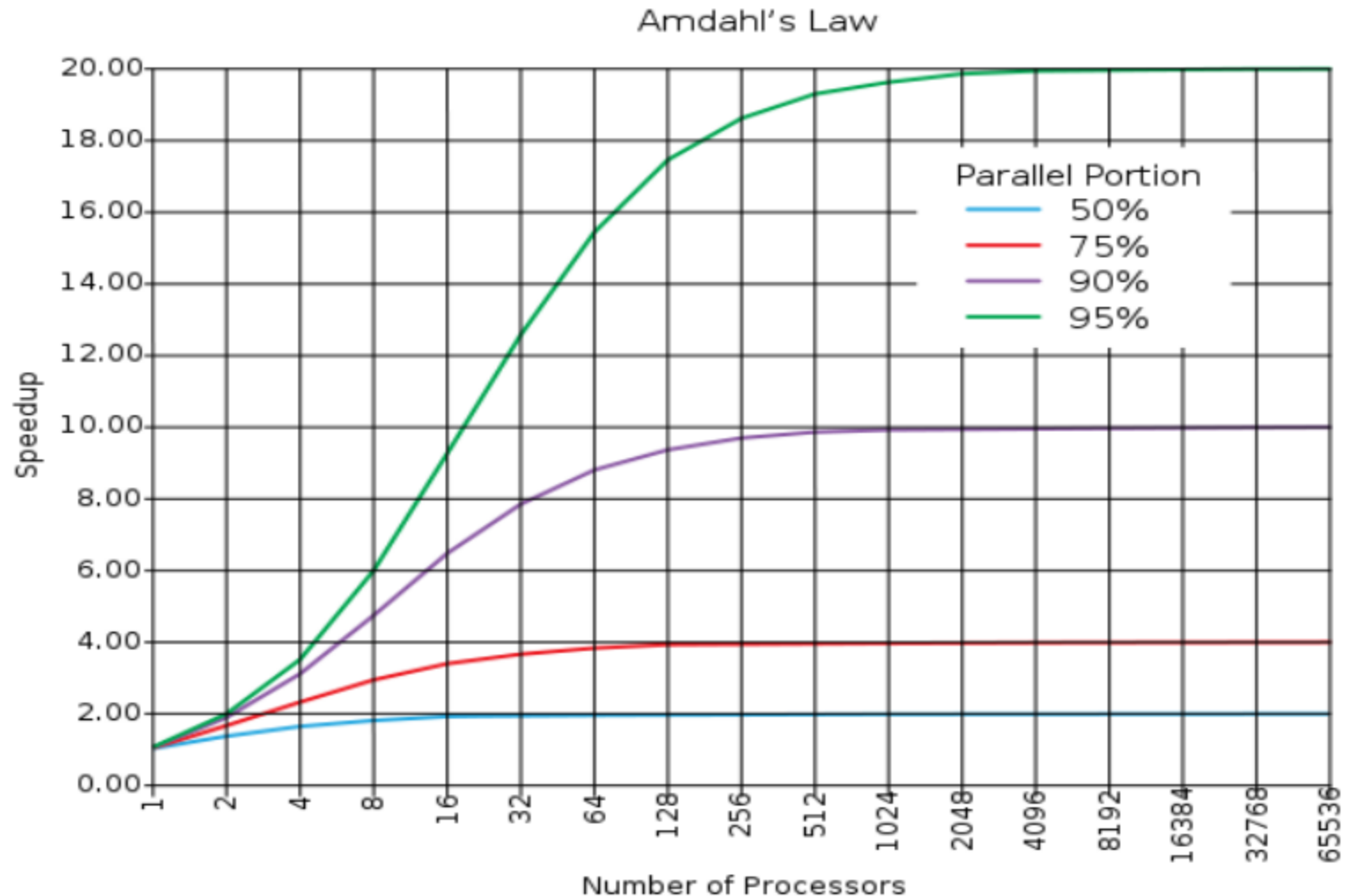


Escalabilidad

Y si aumentamos el número de procesadores para paralelizar la aplicación?

$$\frac{1}{1-P}$$

La aceleración de un programa paralelo está limitada por la porción serial del mismo



AGENDA

1. Introducción
2. Evolución de la ciencia de datos
- 3. Datos**
4. Aplicaciones

Datos

- Representación simbólica de la información
- Pueden clasificarse en:
 - Datos Estructurados
 - Datos No estructurados
 - Datos Semiestructurados
- Se almacenan en Bases de Datos



Datos Estructurados

- Tienen una estructura definida
- Se conocen sus propiedades: tamaño, longitud.
- Suelen representarse en tablas. Son los que podemos encontrar en la mayoría de bases de datos

ID	Nombre	Edad	Sexo	Profesión	Salario
1	Juan	33	M	Ingeniero	4.500.000
2	Ana	38	F	Arquitecta	6.200.000
3	María	7	F	Abogada	9.600.000

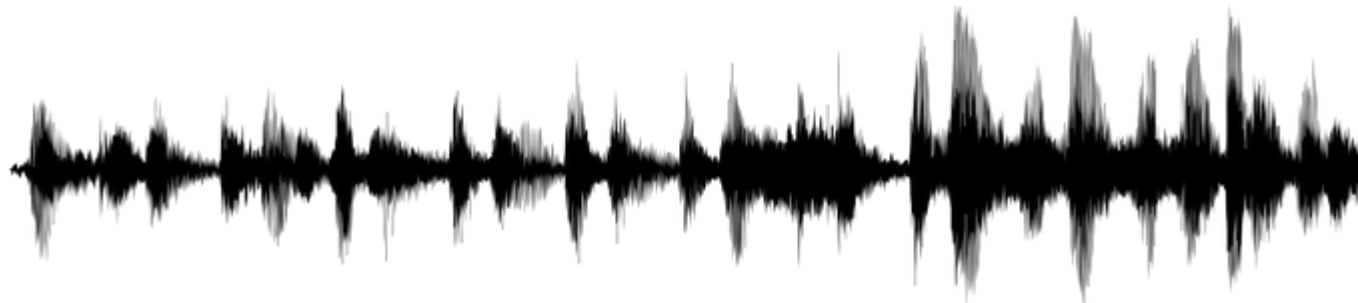
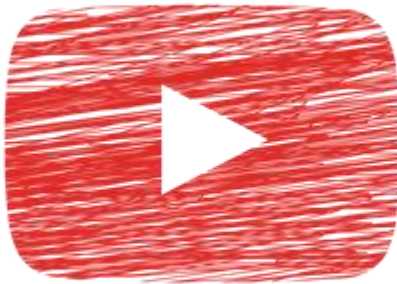
Datos Semiestructurados

- Combinación entre datos estructurados y no estructurados
- No están completamente estructurados pero contienen metadatos para describir sus objetos y relaciones
- Ejemplo:
 - XML
 - JSON

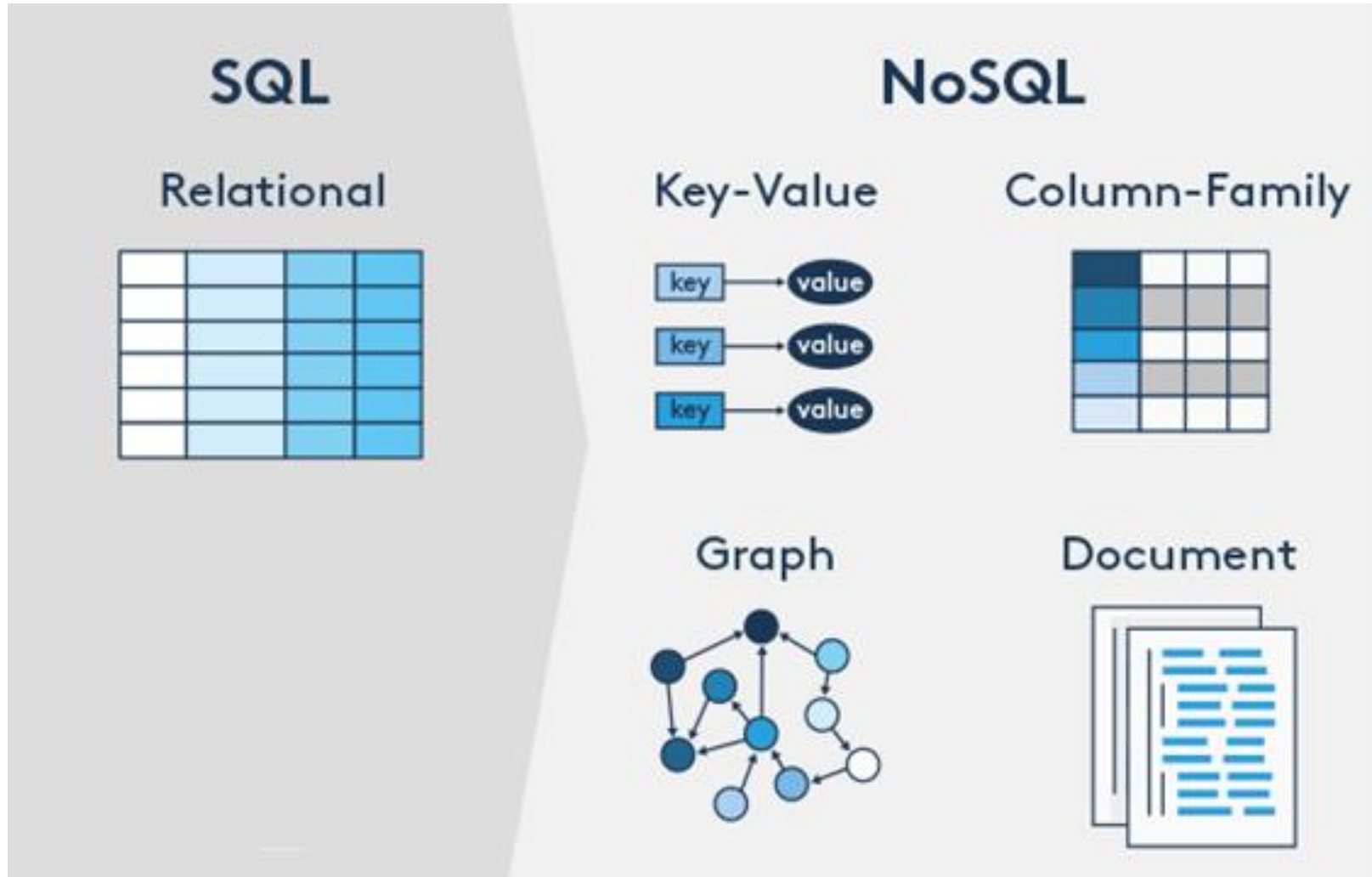
```
{
  "Ciudades": [
    {
      "Nombre": "Medellín"
      "Departamento": "Antioquia"
      "Población":2.500.000
    },
    {
      "Nombre": "Bogotá"
      "Departamento": "Cundinamarca"
      "Población":8.200.000
    }
  ]
}
```


















Datos NO Estructurados

- No tienen una estructura definida
- Ejemplo:
 - Audio
 - Imágenes
 - Texto



Almacenamiento de Datos



Document Database	Graph Databases
   	 
Wide Column Stores	Key-Value Databases
   	     

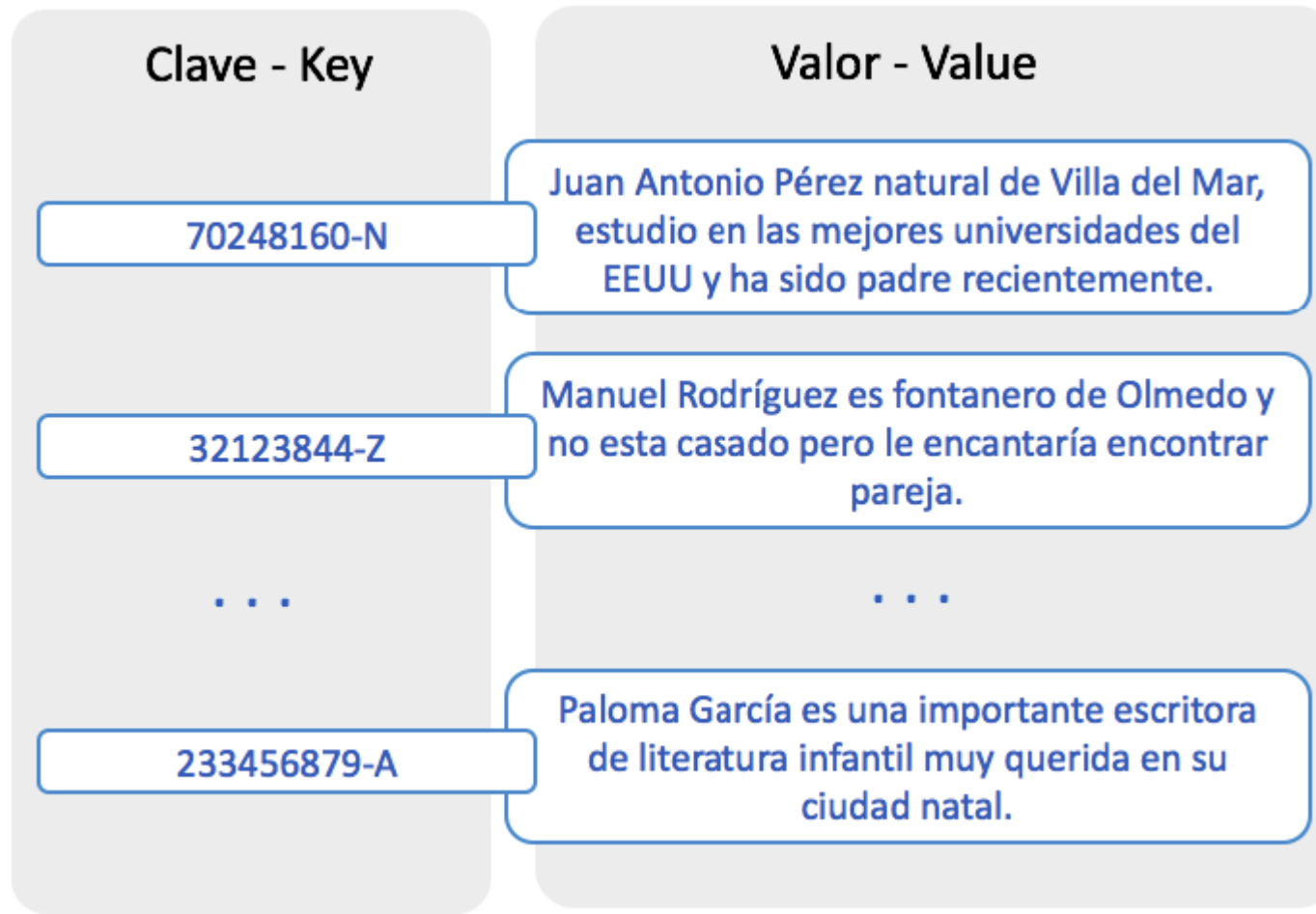
NoSQL: Clave - Valor

- Modelo de base de datos NoSQL mas popular y mas sencillo
- Cada elemento está identificado por una llave única
- Eficiente en lectura y escritura



NoSQL: Clave - Valor

El valor puede ser cualquier tipo de dato como una imagen, un archivo, una página web, un código de programación



<http://www.diegocalvo.es/base-de-datos-clave-valor/>

NoSQL: Documental

- Almacena la información como un documento
- Utiliza documentos que contienen estructuras simples (XML, JSON)
- La indexación de los documentos se hace bajo una clave única
- Permite crear índices



NoSQL: Documental

XML

ID	Nombre	Edad	Sexo	Profesión	Salario
1	Juan	33	M	Ingeniero	4.500.000
2	Ana	38	F	Arquitecta	6.200.000

```
<?xml version="1.0" encoding="UTF-8" ?>
<empleados>
  <Id>1</Id>
  <Nombre>Juan</Nombre>
  <Edad>33</Edad>
  <Sexo>M</Sexo>
  <Profesión>Ingeniero</Profesión>
  <Salario>4500000</Salario>
</empleados>
<empleados>
  <Id>2</Id>
  <Nombre>Ana</Nombre>
  <Edad>38</Edad>
  <Sexo>F</Sexo>
  <Profesión>Arquitecta</Profesión>
  <Salario>6200000</Salario>
</empleados>
```


NoSQL: Documental

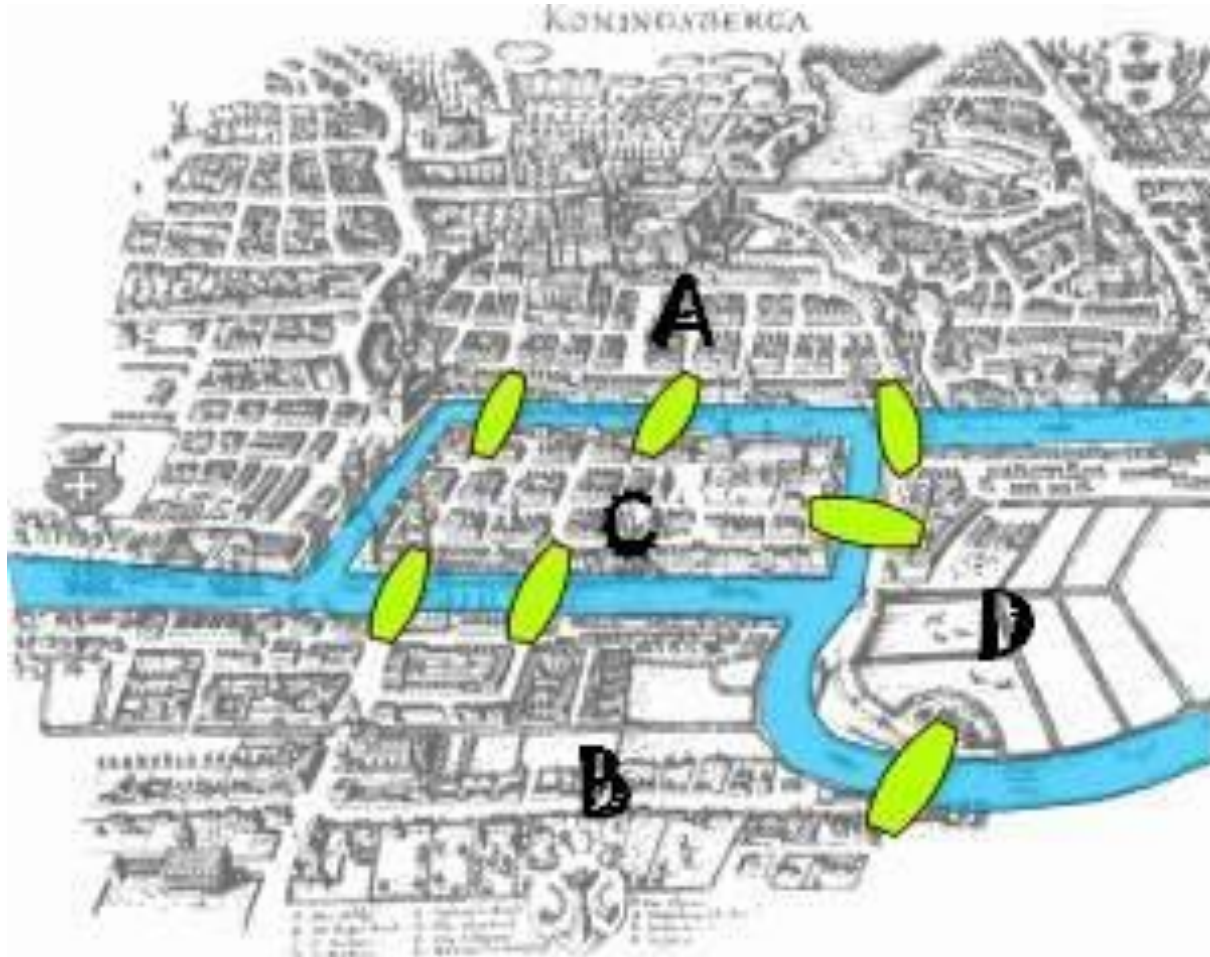
JSON

ID	Nombre	Edad	Sexo	Profesión	Salario
1	Juan	33	M	Ingeniero	4.500.000
2	Ana	38	F	Arquitecta	6.200.000

```
{
  "empleados": [
    {
      "Id": 1,
      "Nombre": "Juan",
      "Edad": 33,
      "Sexo": "M",
      "Profesión": "Ingeniero",
      "Salario": 4500000
    },
    {
      "Id": 2,
      "Nombre": "Ana",
      "Edad": 38,
      "Sexo": "F",
      "Profesión": "Arquitecta",
      "Salario": 6200000
    }
  ]
}
```

Grafos

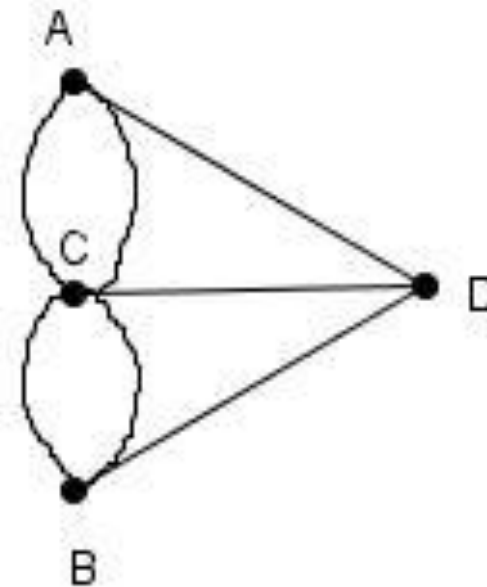
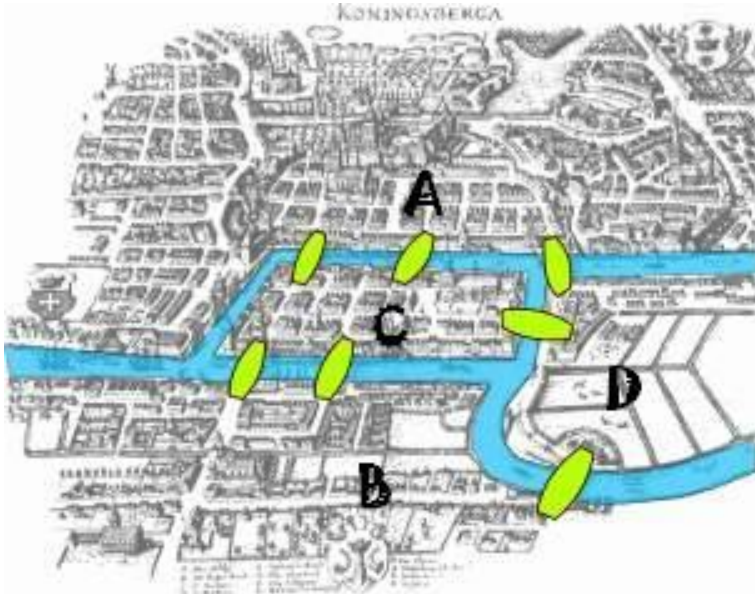
Problema de los puentes de Königsberg



Empezar en un punto, pasar por los siete puentes sin repetir ninguno y volver al punto de partida

Grafos

- Representar la ciudad de Königsberg como un grafo



- Vértices: Las cuatro zonas de la ciudad (A,B,C,D)
- Aristas: Puentes de la ciudad

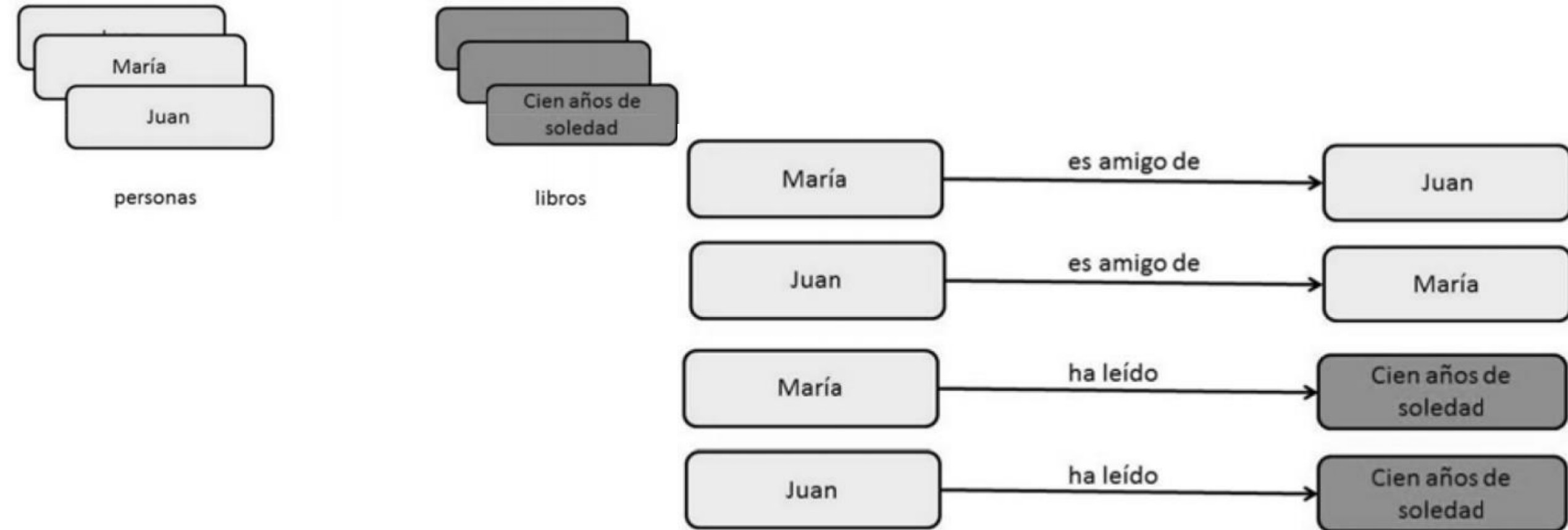
NoSQL: Orientado a Grafos

- Se almacena la información como nodos de un grafo y sus relaciones con otros nodos
- Presenta mejor rendimiento, permite una navegación mas eficiente entre los nodos y sus relaciones que en un modelo relacional
- Permite una gran flexibilidad



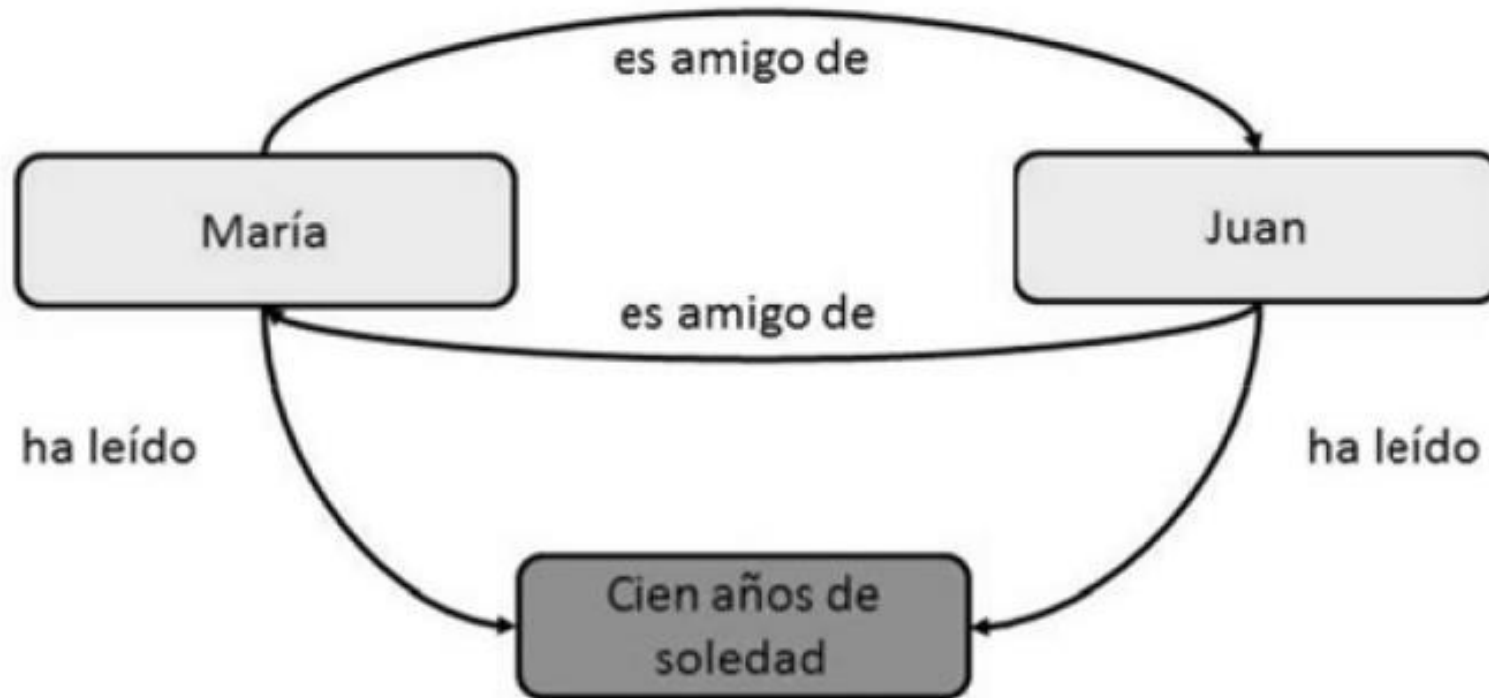
NoSQL: Orientado a Grafos

Ejemplo: María y Juan son amigos y ambos han leído el libro cien años de soledad



NoSQL: Orientado a Grafos

Ejemplo: María y Juan son amigos y ambos han leído el libro cien años de soledad



NoSQL: Orientado a Columnas

- Almacena los datos en forma de Columnas bajo una clave única
- Cada registro puede convertirse en una o mas columnas
- Cada columna puede contener distintas estructuras de datos
- Se utilizan para consultar datos de tipo histórico



NoSQL: Orientado a Columnas

Row key1	Column Key1	Column Key2	Column Key3	...
	Column Value1	Column Value2	Column Value3	
⋮				

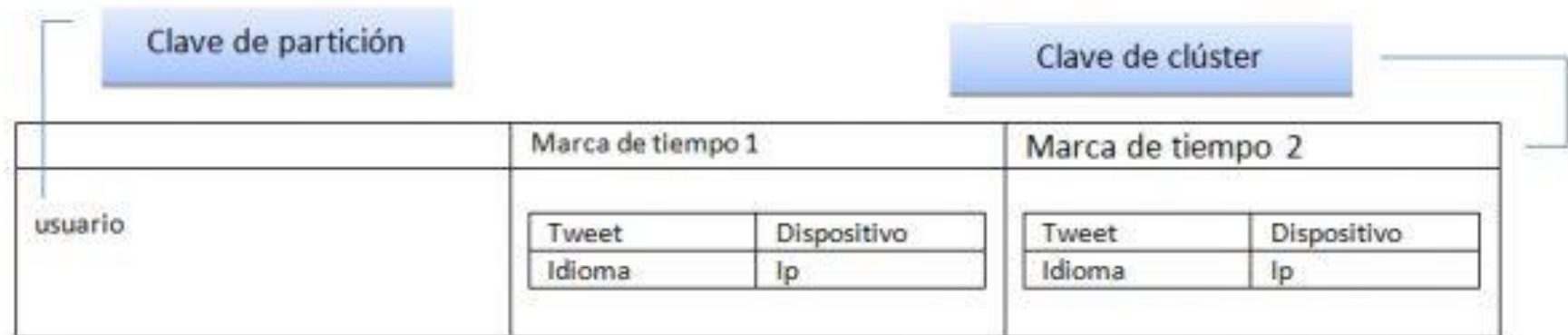
Row key1	Super Column key1			Super Column key2			...
	Subcolumn Key1	Subcolumn Key2	...	Subcolumn Key3	Subcolumn Key4	...	
	Column Value1	Column Value2		Column Value3	Column Value4		
⋮							

Relational Model	Cassandra Model
Database	Keyspace
Table	Column Family (CF)
Primary key	Row key
Column name	Column name/key
Column value	Column value

NoSQL: Orientado a Columnas

Ejemplo: Crear una tabla para almacenar tweets

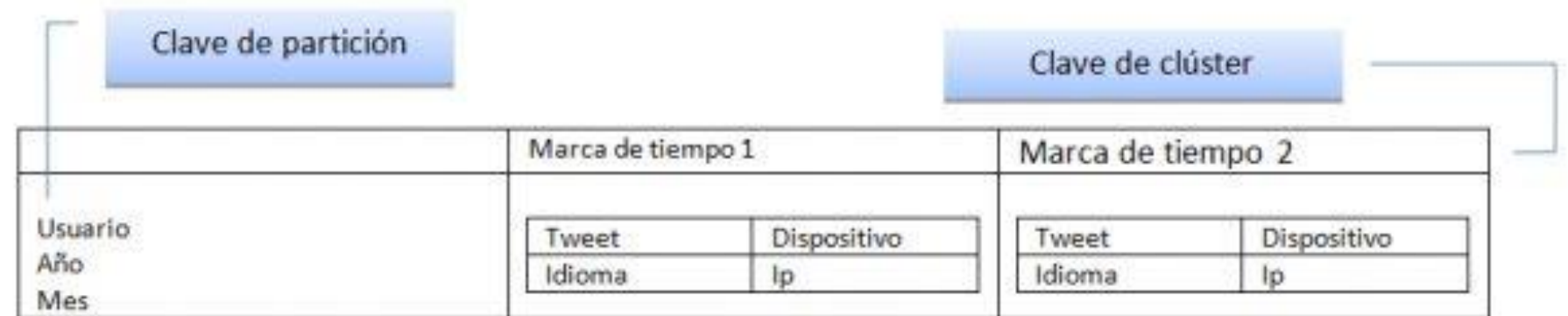
```
CREATE TABLE tweet (  
  usuario bigint,  
  timestamp timestamp,  
  tweet text,  
  dispositivo text,  
  idioma text,  
  ip text,  
  PRIMARY KEY (usuario, timestamp)  
) WITH CLUSTERING ORDER BY (timestamp DESC);
```



NoSQL: Orientado a Columnas

Y si un solo usuario crea muchos tweets al mes?

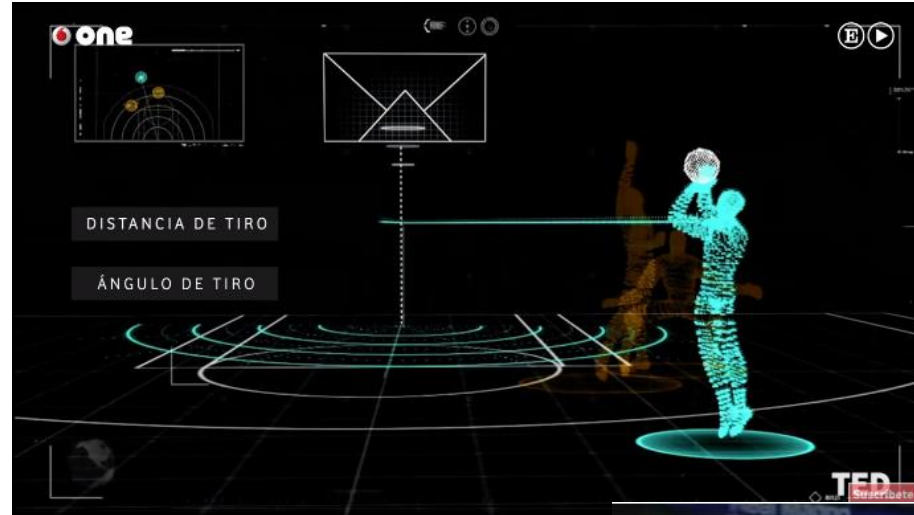
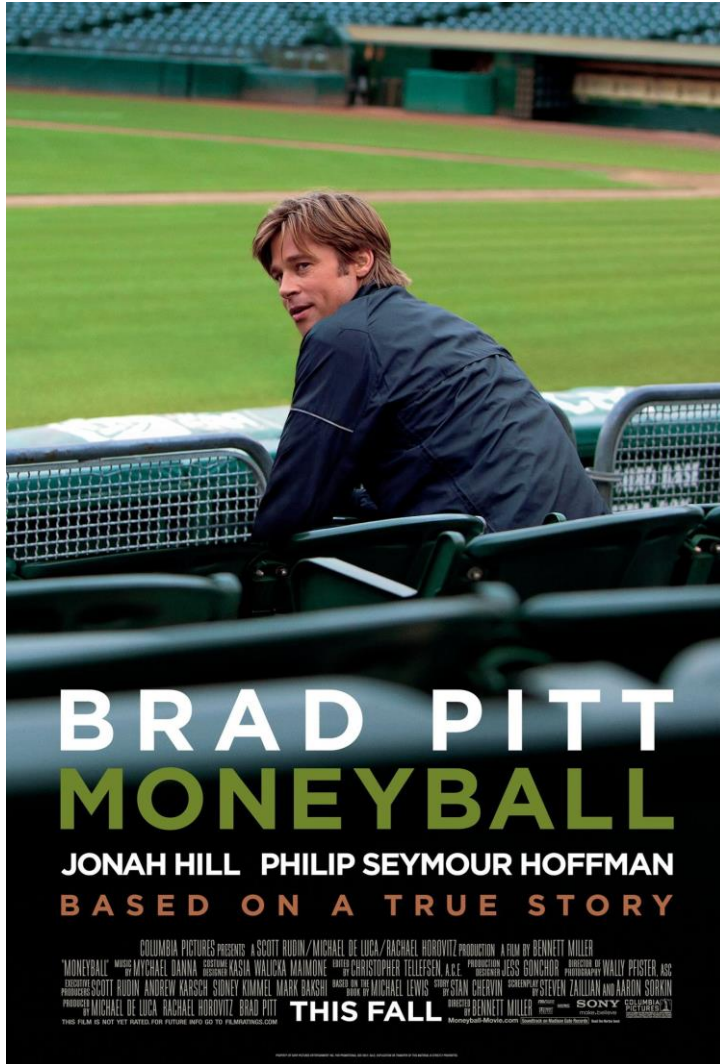
```
CREATE TABLE tweet (  
  usuario bigint,  
  year int,  
  month int,  
  timestamp timestamp,  
  tweet text,  
  dispositivo text,  
  idioma text,  
  ip text,  
  PRIMARY KEY ((usuario, year, month), timestamp)  
) WITH CLUSTERING ORDER BY (timestamp DESC);
```



AGENDA

1. Introducción
2. Evolución de la ciencia de datos
3. Datos
- 4. Aplicaciones**

Aplicaciones



<https://www.youtube.com/watch?v=EKmWMjvWFPg>

<https://www.youtube.com/watch?v=DXq30dvE0Xg>



<https://www.youtube.com/watch?v=-4R3m4ybDz4>

Aplicaciones



<https://www.youtube.com/watch?v=ku78zo9fhoI>



<https://www.estrategiasdeinversion.com/analisis/bolsa-y-mercados/el-experto-opina/el-big-data-nos-ayuda-a-predecir-el-comportamiento-n-323203>

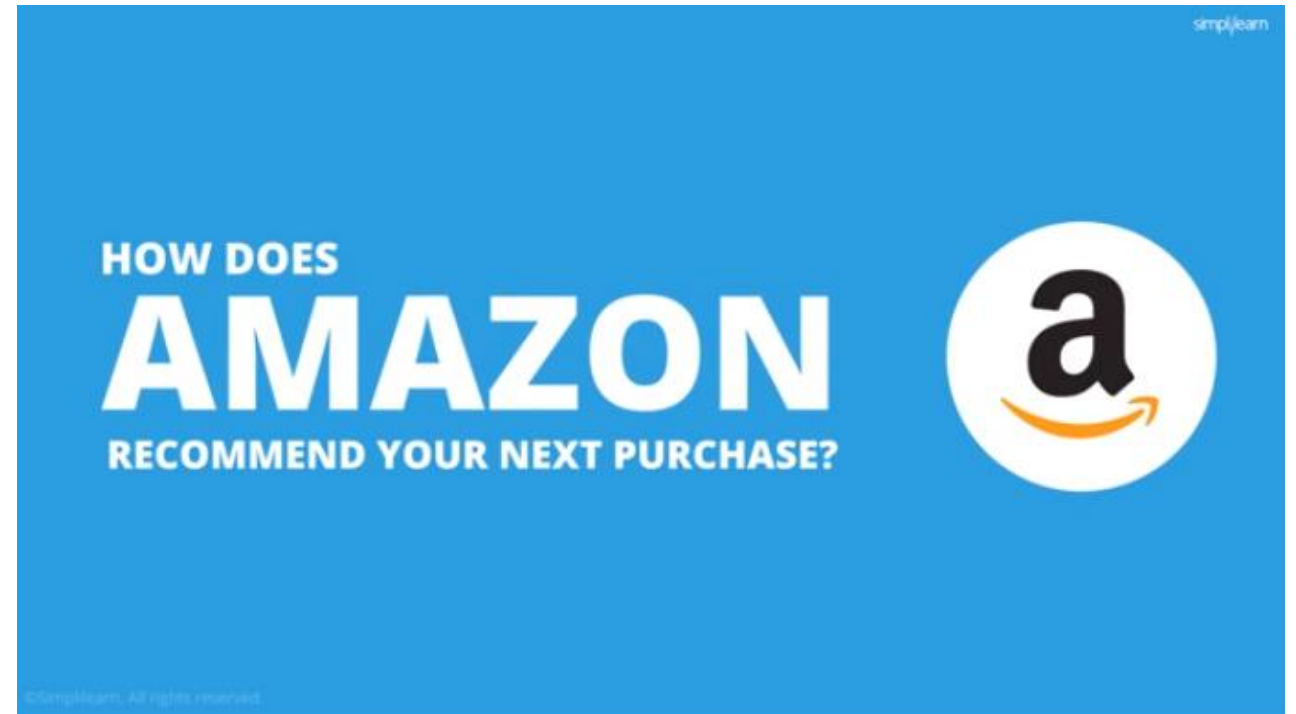


<https://www.dw.com/es/predecir-el-futuro-con-el-big-data/av-17816690>

Aplicaciones



<https://www.youtube.com/watch?v=fO7G6gRFLLM>



<https://www.youtube.com/watch?v=ImN4FTarqfo&t=53s>

Aplicaciones



[Video](#)

<https://colombiadigital.net/actualidad/noticias/item/9006-caoba-centro-de-excelencia-y-apropiacion-en-big-data-y-data-analytics.html>

EVALUACIÓN

Actividad	%
Taller I: Map-reduce	30
Taller II: RDD	30
Taller III: Dataframe	40