

Arquitecturas Cloud y Big Data



GOBIERNO
DE ESPAÑA

VICEPRESIDENCIA
PRIMERA DEL GOBIERNO
MINISTERIO
DE ASUNTOS ECONÓMICOS
Y TRANSFORMACIÓN DIGITAL

SECRETARÍA DE ESTADO
DE DIGITALIZACIÓN
E INTELIGENCIA ARTIFICIAL

red.es

Centro de
Referencia Nacional
en Comercio Electrónico
y Marketing

CRN
Digital





UNIÓN EUROPEA

Barrabés

The Valley

"El FSE invierte en tu futuro"
Fondo Social Europeo

Índice

1. Hoja de Ruta y Metodología
2. Big Data
3. Cloud computing. 
4. Aplicaciones Big Data 
5. Antecedentes: Hadoop vs Spark

1. Hoja de ruta y metodología



SOBRE MÍ



Docente:

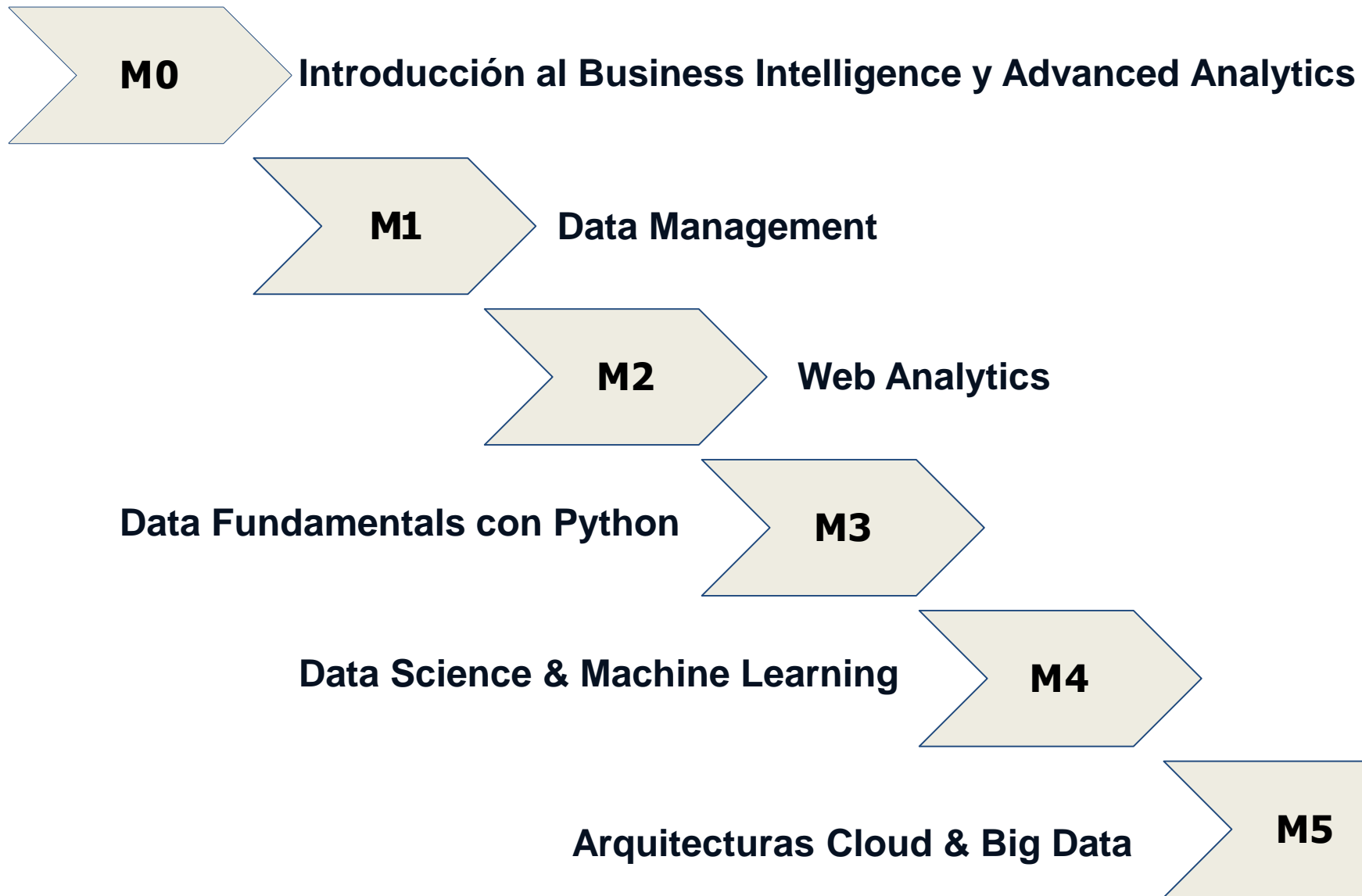
Eduardo Rivero Falcón

Formación:

- **Máster Universitario en Ciencia de Datos (UOC)**
- **Programa Superior en Big Data (I.T. Telefónica)**
- **Ingeniería Técnica de Telecomunicación (ULPGC)**

- **A finales de 2018 decidí dar un giro a mi desarrollo profesional hacia el mundo de la analítica y la ciencia de datos.**
- **Científico de datos: 2 años de experiencia**
- **Voluntariado: enseñanza programación grupos desfavorecidos (1.5 años)**

HOJA DE RUTA



METODOLOGÍA

- **Aula Virtual (20 horas “solo”)**

Si se atiende y se practica, tiempo suficiente para asentar fundamentos.

- **Conceptos teóricos**
(los necesarios y contexto)
- **Actividades en clase**
(learning by doing)
- **Evaluación final:**
Cuestionario
(una respuesta correcta)



2. Big Data



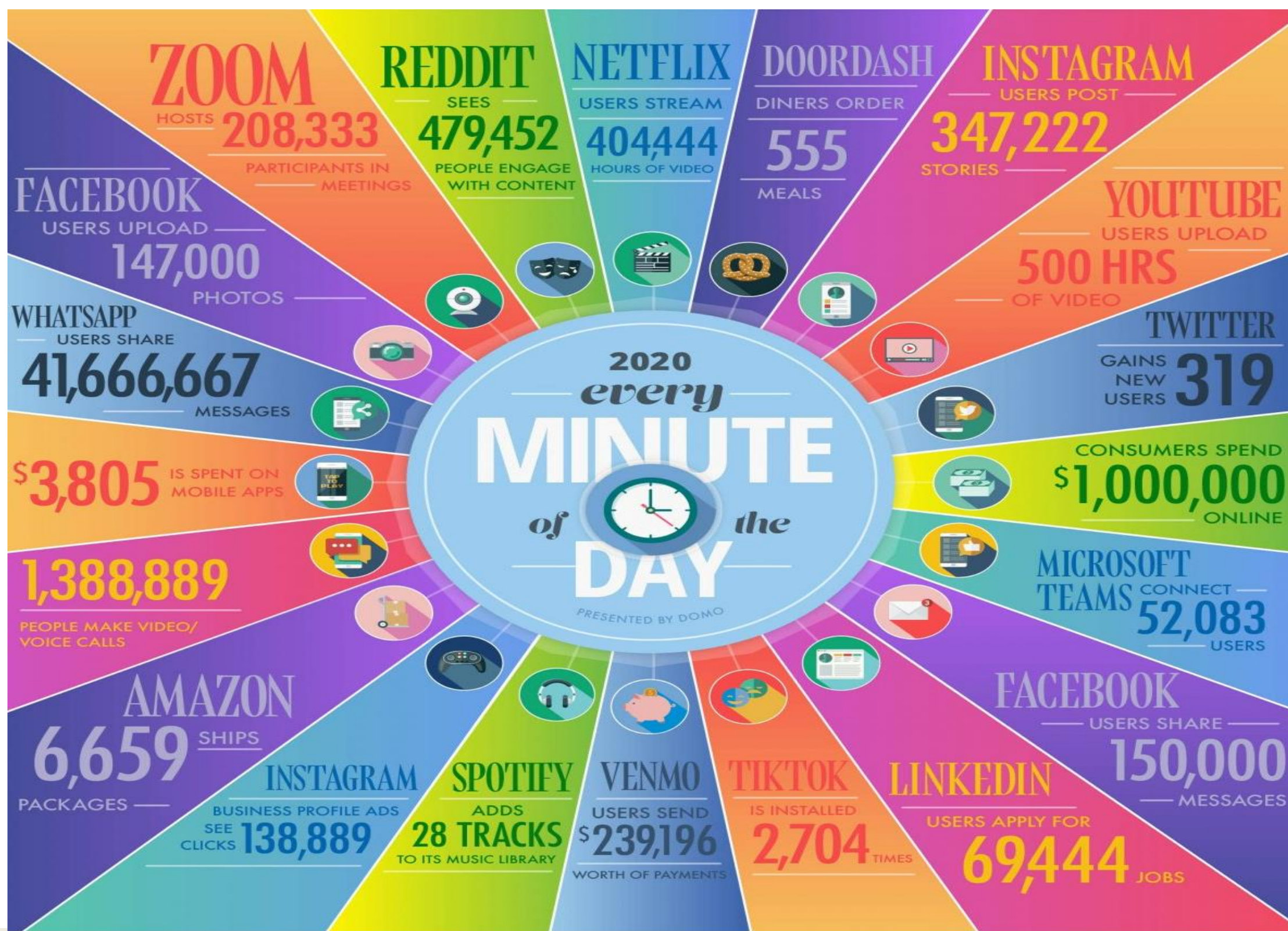
BIG DATA: definición

Podríamos definir “Big Data” como a todo conjunto de datos de tamaño tan grande que es difícil de procesar mediante las herramientas de gestión de bases de datos o las aplicaciones tradicionales de procesamiento de datos.

Ante estos casos surgieron los siguientes desafíos:

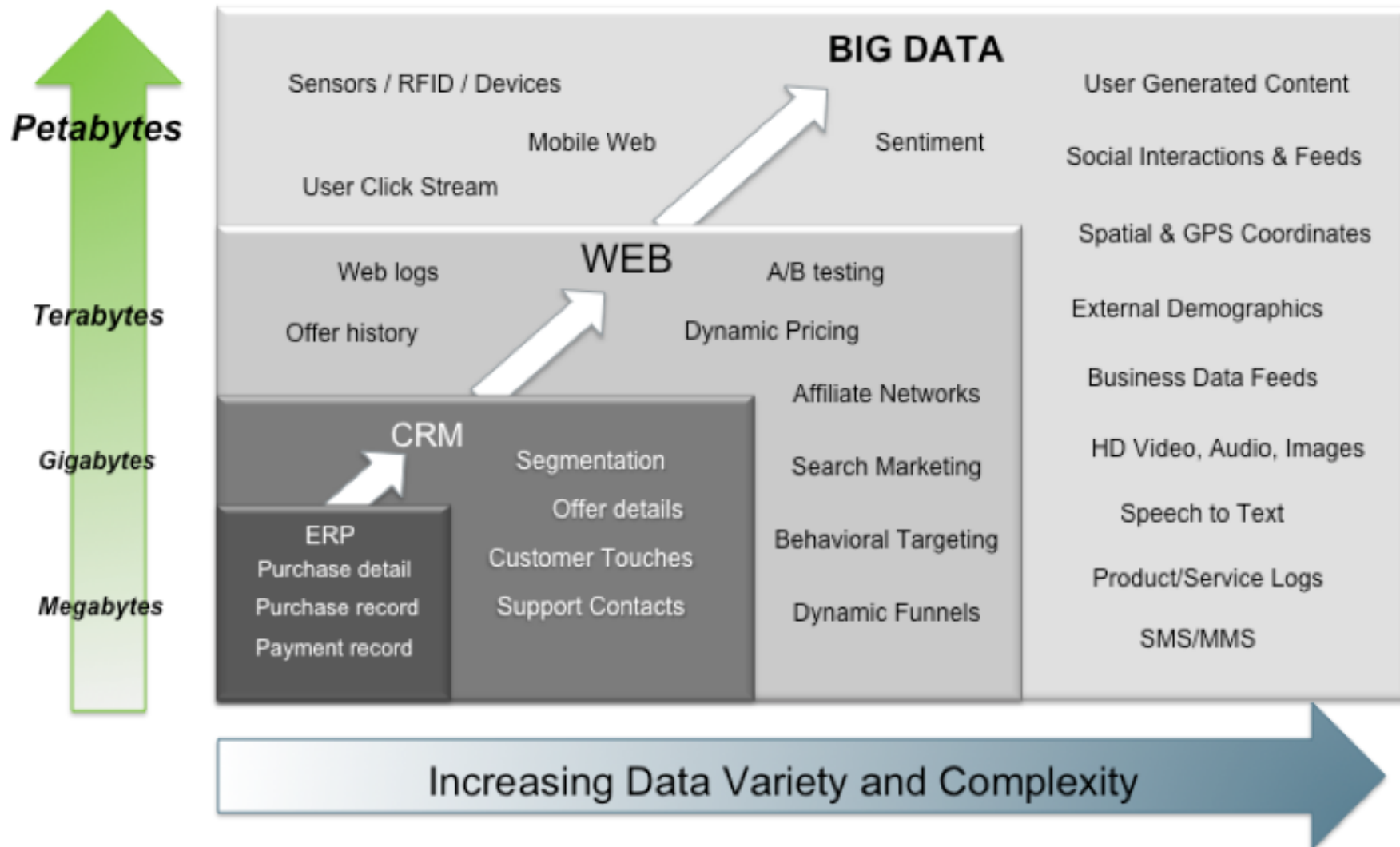
- ☐ captura
- ☐ almacenamiento
- ☐ búsqueda, consulta
- ☐ transferencia
- ☐ análisis
- ☐ visualización / interpretación
- ☐ proceso en tiempo real

¿¿Quién genera tal cantidad de datos??



BIG DATA: evolución fuentes y usos

Big Data = Transactions + Interactions + Observations



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

BIG DATA: evolución

- **1997:** Michael Cox y David Ellsworth (NASA) utilizan por primera vez el término Big Data (problema)
- **2003-2004:** Google publica GFS y MAPREDUCE (pilares de Hadoop, “antecesor” Spark)
- **2006:** Hadoop, código 100% abierto para Big Data
- **2009:** Nace Spark (Universidad de Berkeley)
- **2014-2015:** Auge tráfico móvil, datos geolocalizados, IOT, Smart cities

En la **actualidad**, se crean **2,5 exabytes** de información **al día**.

BIG DATA: nuevos perfiles

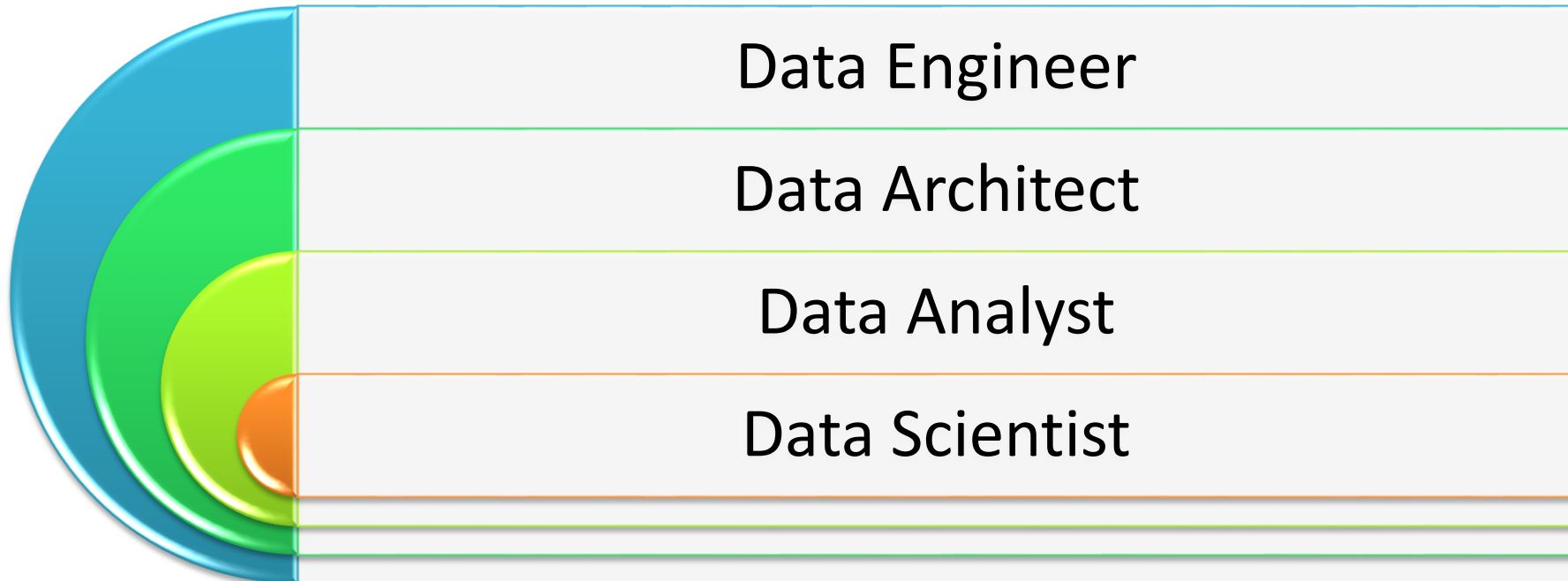
- Los **datos** se convierten en un **ACTIVO** más de las organizaciones y empresas (**ORGANIZACIONES DATA-DRIVEN**)
- Es evidente que se tienen que desarrollar **nuevos perfiles** para cuidar y sacar el máximo de esos activos.

¿CONOCÉIS ALGUNOS DE ESTOS
PUESTOS DE TRABAJO ESPECÍFICOS?

!!! Veamos cuales son !!!



BIG DATA: Roles (puestos de trabajo)



- **Data Engineer**

- Big Data Architect
- Data Analyst
- Data Scientist

**SALARIO ANUAL MEDIO EN
ESPAÑA: 35.000€**

DATA ENGINEER

Los Data Engineer **preparan todo el ecosistema** para que los demás puedan obtener sus datos limpios y preparados para su análisis.

ROL PRINCIPAL

Diseñan, desarrollan, construyen, prueban y mantienen los sistemas de procesamiento de datos

DEBEN CONOCER:



- Data Engineer
- **Data Architect**
- Data Analyst
- Data Scientist

**SALARIO ANUAL MEDIO
EN ESPAÑA: 41.000€**

DATA ARCHITECT

Data Architect es un Data Engineer con una **visión más global, y más orientada a la integración,** centralización y el mantenimiento de todas las fuentes de datos.

DEBEN CONOCER:



- Data Engineer
- Data Architect
- **Data Analyst**
- Data Scientist

**SALARIO ANUAL MEDIO
EN ESPAÑA: 28.000€**

DATA ANALYST

Perfil más orientados al **análisis de datos**, el Data Analyst es un perfil previo al de Data Scientist.

ROL PRINCIPAL

Minería, obtención y/o recuperación de datos así como su procesado, estudio avanzado y visualización.

DEBEN CONOCER:



- Data Engineer
- Data Architect
- Data Analyst
- **Data Scientist**

**SALARIO ANUAL MEDIO
EN ESPAÑA: 35.000€**

DATA SCIENTIST

Es la “evolución del Data Analyst”.
Un rol **más específico** y menos
alineado con la visión de negocio.

DIFERENCIA CON RESPECTO A D.A.

lo que diferencia al Data Scientist
es que es el encargado de sacarle
valor a los datos. Tiene un rol **más
enfocado a la predicción**

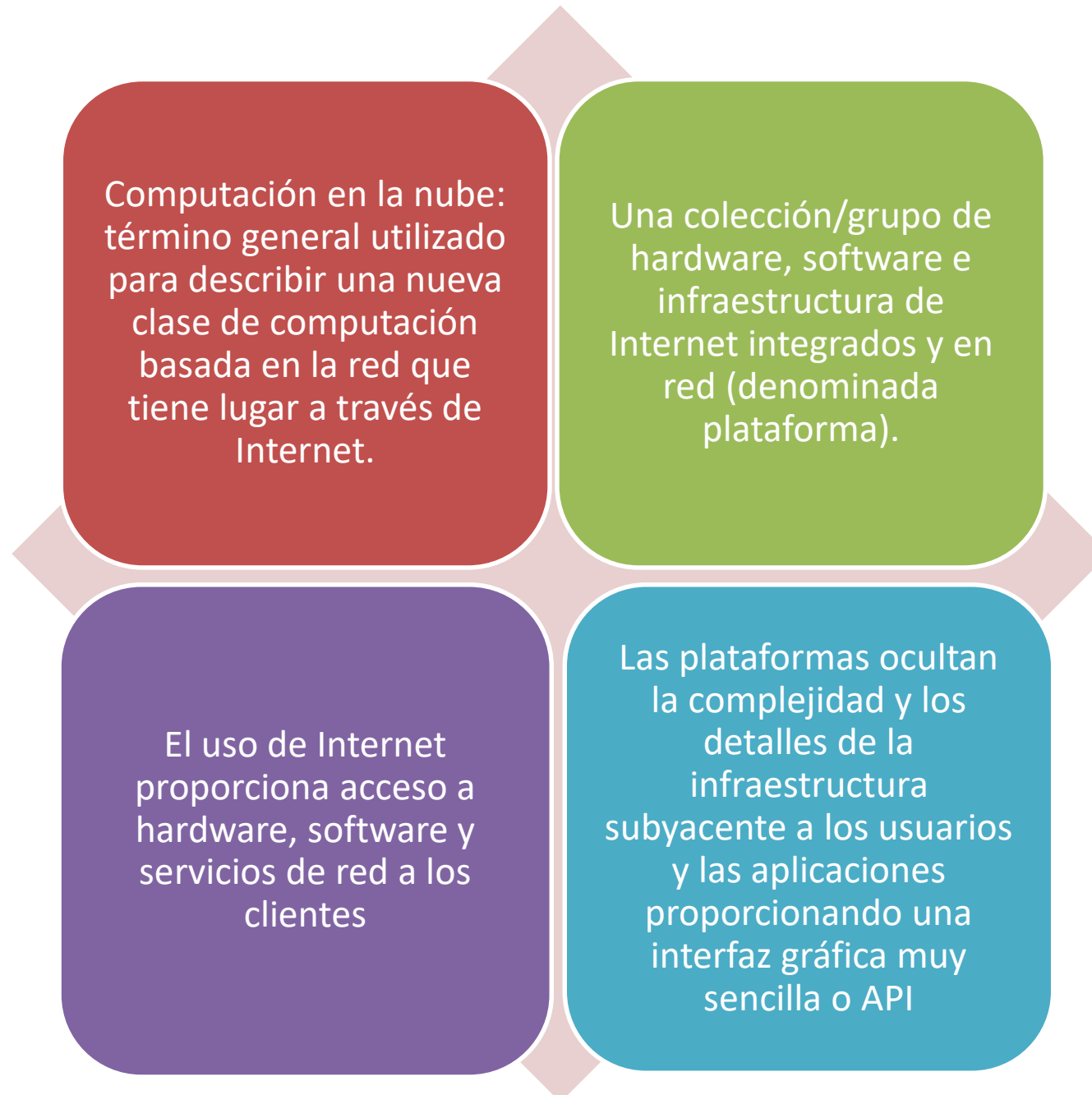
DEBEN CONOCER:



3. Cloud computing (nube)



COMPUTACIÓN EN LA NUBE



PREGUNTA: ¿Nosotros utilizamos la nube?

Google Drive

Videoconferencia (Skype, Teams o Zoom)

Tiempo de ocio (Spotify, Netflix ...)

Mensajería (Whatsapp ...)

SERVICIOS EN LA NUBE

- **Infraestructura como servicio (IaaS)**

Oferta de servicios relacionados con el hardware utilizando los principios de la computación en nube. Almacenamiento (base de datos o almacenamiento en disco) o servidores virtuales. Amazon EC2, Amazon S3

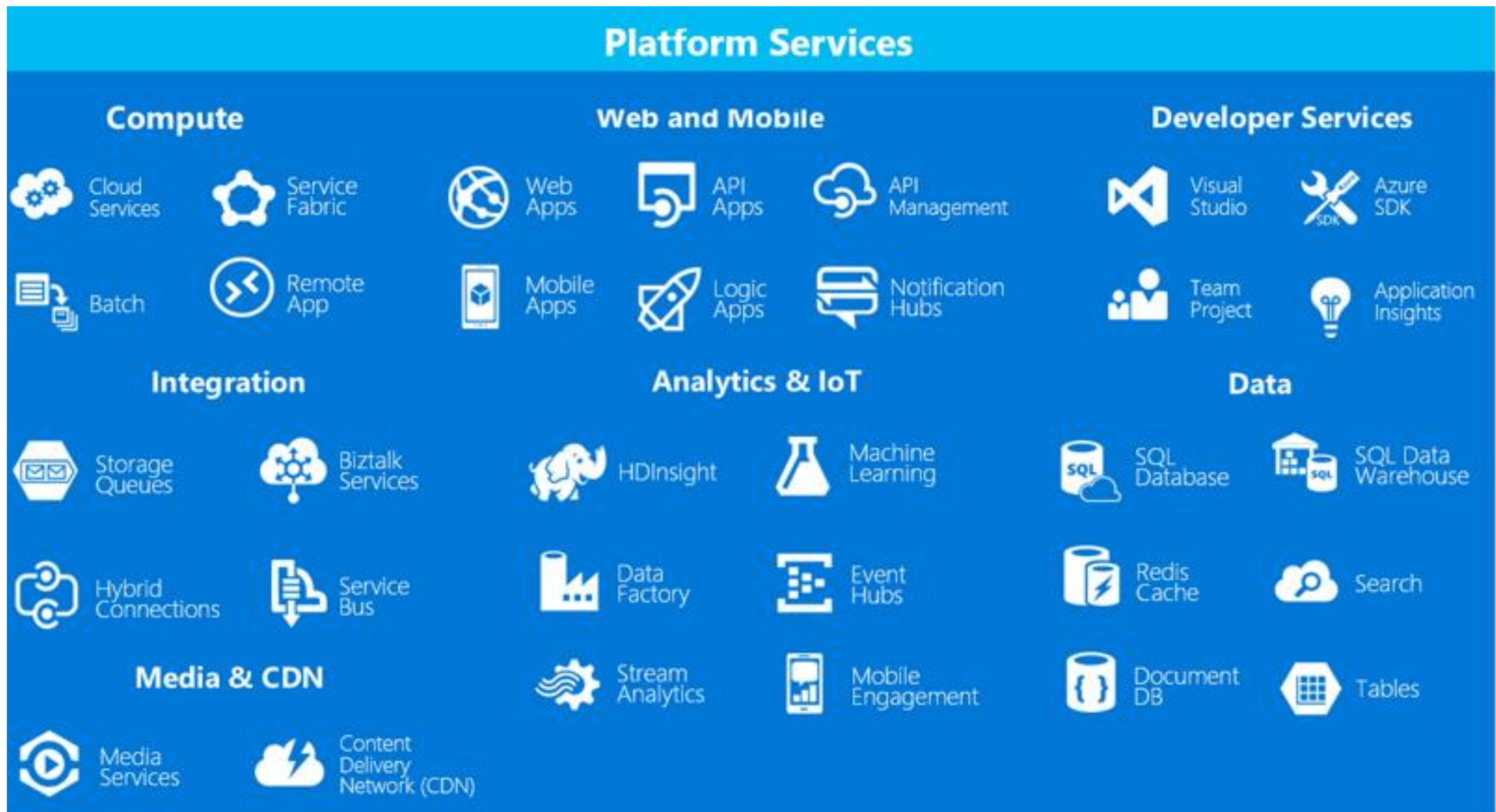
- **Plataforma como servicio (PaaS)**

Ofrecen una plataforma de desarrollo en la nube. Application Engine de Google

- **Software como servicio (SaaS)**

Incluye una oferta completa de software en la nube. Los usuarios pueden acceder a una aplicación de software alojada por el proveedor de la nube en régimen de pago por uso. Se trata de un sector muy consolidado. Microsoft Office 365

SERVICIOS EN LA NUBE: AZURE (Microsoft)



SERVICIOS EN LA NUBE: AWS (Amazon Web Services)

Servicios AWS

Deployment & Management

Application Services



Amazon SQS



Amazon ElasticTranscoder



Amazon SES



Amazon AppStream



Amazon CloudSearch

Mobile Services



Amazon Cognito



Amazon Mobile Analytics



Amazon SNS

Enterprise Applications



Amazon WorkDocs



Amazon WorkSpaces



Amazon WorkMail

Application Services

Administration & Security



AWS DirectoryService



AWS IAM



AWS Trusted Advisor



AWS Config



AWS CloudTrail



Amazon CloudWatch

Deployment & Management



Amazon CloudFormation



AWS OpsWorks



AWS CodeDeploy

Analytics



Amazon Kinesis



AWS Data Pipeline



Amazon EMR

Foundation Services

Compute



Amazon EC2



AWS Lambda

Storage & Content Delivery



Amazon CloudFront



Amazon Glacier



AWS Storage Gateway



Amazon Content Delivery

Database



Amazon Dynamo DB



Amazon RDS



Amazon Redshift



Amazon Elastic Cache

Networking



Amazon Route 53



Amazon VPC



AWS Direct Connect

SERVICIOS EN LA NUBE: GCP (Google Cloud Platform)

Compute



App Engine



Compute Engine



Container Engine

Storage



Cloud Storage



Cloud Datastore



Cloud SQL



Cloud Bigtable

Big Data



BigQuery



Cloud Dataflow



Cloud Dataproc



Cloud Pub/Sub

Services



Cloud Endpoints



Translate API



Prediction API

SERVICIOS EN LA NUBE: Ejemplo

Amazon Web Services (AWS) ofrece un conjunto de servicios de computación en la nube que conforman una plataforma de computación bajo demanda:

Amazon Elastic Compute Cloud (EC2): Máquinas virtuales para ejecutar software personalizado

Amazon Simple Storage Service (S3): Almacén simple de valores clave, accesible como servicio web

Amazon Elastic MapReduce (EMR): Computación MapReduce escalable

Amazon DynamoDB: Base de datos NoSQL distribuida, una de varias en AWS

SERVICIOS EN LA NUBE: Ventajas

- ☐ La nube nos está beneficiando en cuanto a ahorros de costes y precios de los servicios.
- ☐ No hay una necesidad de mantenimiento del producto o es tan simple su mantenimiento que una persona no experta podría actualizar los programas con un simple clic.
- ☐ Compartir archivos es más rápido y cómodo.
- ☐ Los accesos son ahora más rápidos desde cualquier sitio y cuando queramos. (Dispositivo + Internet)
- ☐ Acceso a través de múltiples dispositivos.

SERVICIOS EN LA NUBE: Inconvenientes

- ☐ Para acceder a nuestra vida en la nube es necesario Internet y no siempre hay posibilidad de disponer de él.
- ☐ Los datos que introducimos en la nube están en peligro de ser robados o usados sin nuestro consentimiento...
- ☐ En el caso de que tengamos todas nuestras aplicaciones centralizadas en un único sitio: si falla el servicio, falla todo nuestro sistema.

BIG DATA Y CLOUD van unidos



4. Aplicaciones Big data



APLICACIONES BIG DATA

- ❑ Marketing, Advertising (segmentación de clientes)
- ❑ Aplicaciones financieras (análisis de riesgos préstamos, evaluación experiencia del cliente)
- ❑ Medicina (decodificación del ADN, la detección de posibles enfermedades)
- ❑ Deporte (estado físico, rutinas de entrenamiento, tácticas)
- ❑ Ciberseguridad (detectar patrones de conducta y prevenir amenazas a la seguridad, prevenir ataques de hackers)
- ❑ Urbanismo y ciudades Inteligentes (gestión del transporte público a través del análisis de datos, gestión del tráfico)
- ❑ Mercado inmobiliario (adquisición de viviendas para su venta, coste actual y su evolución a lo largo de los años)

- ❑ Mapeando fenología desde el cielo
(PYCONES 2021)
- ❑ Seguimiento de falsificaciones en la web
(PYCONES 2021)



5. Antecedentes: Hadoop vs Spark

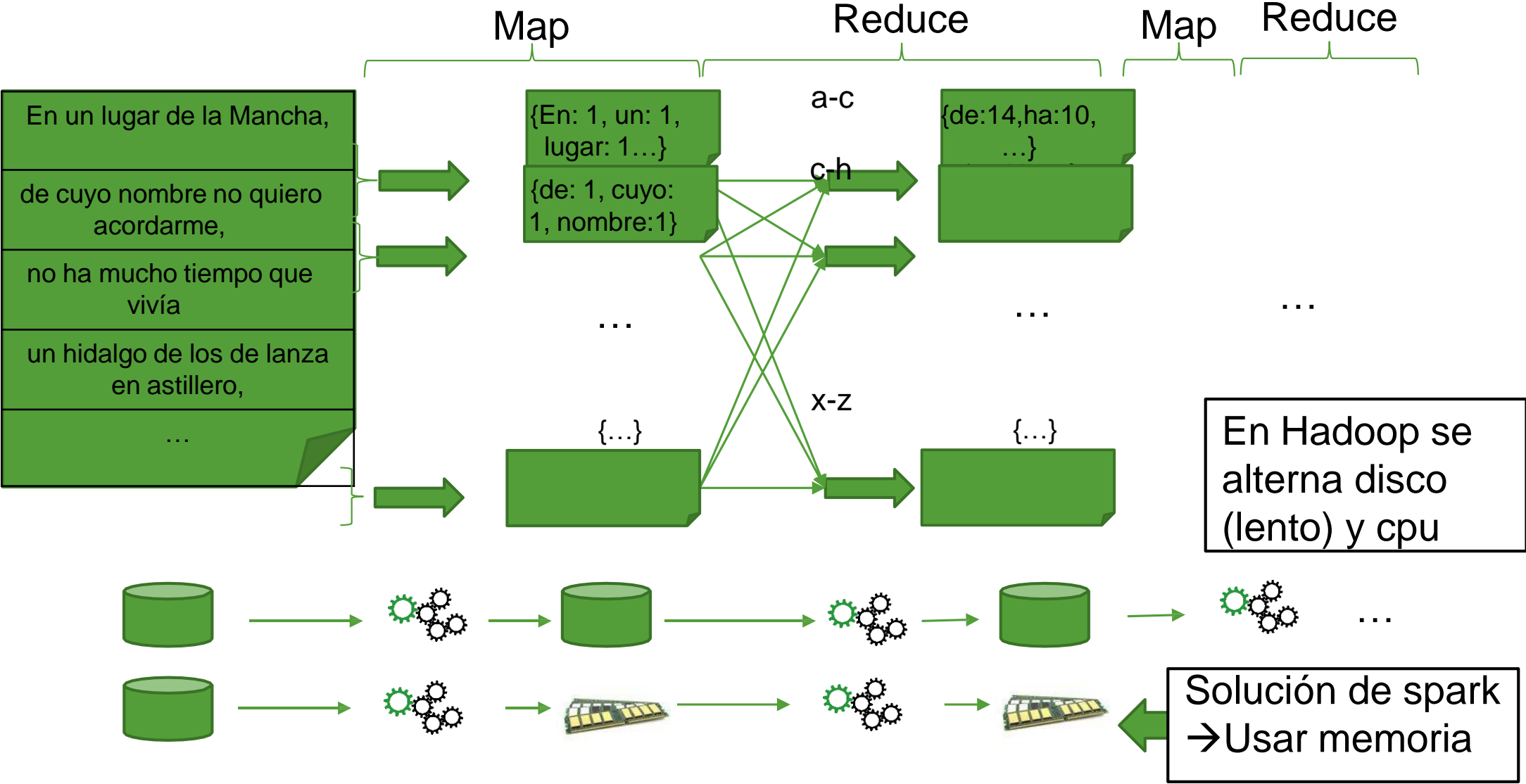


ANTECEDENTES: Hadoop

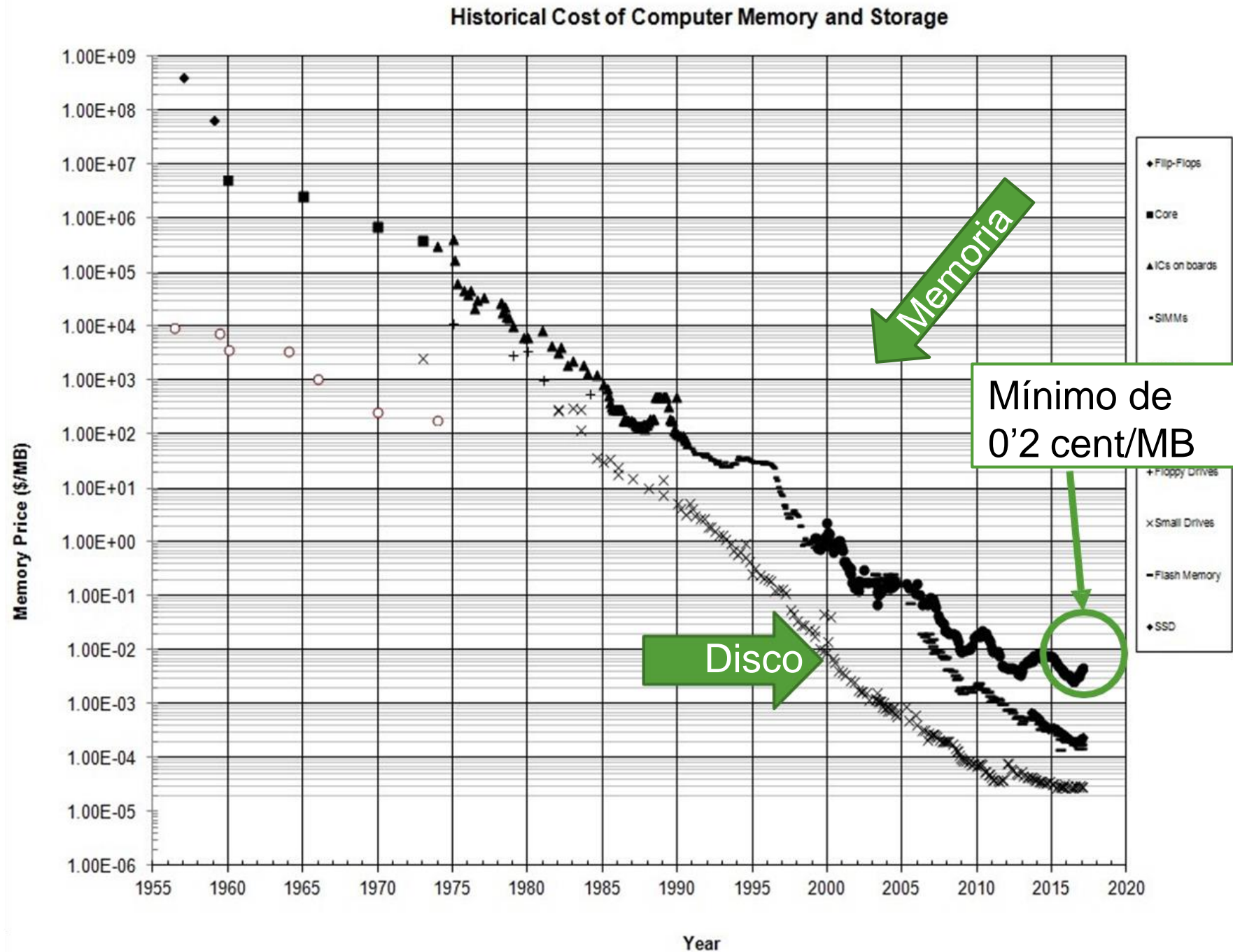
- **1997:** Michael Cox y David Ellsworth (NASA) utilizan por primera vez el término Big Data (problema)
- **2003-2004:** Google publica GFS y MAPREDUCE (pilares de Hadoop, “antecesor” Spark)
- **2006:** Hadoop, código 100% abierto para Big Data
- **2009:** Nace Spark (Universidad de Berkeley)
- **2014-2015:** Auge tráfico móvil, datos geolocalizados, IOT, Smart cities

En la actualidad, se crean 2,5 exabytes de información al día.

Cómo trabaja Hadoop vs Spark



Hadoop vs Spark: Coste disco - memoria

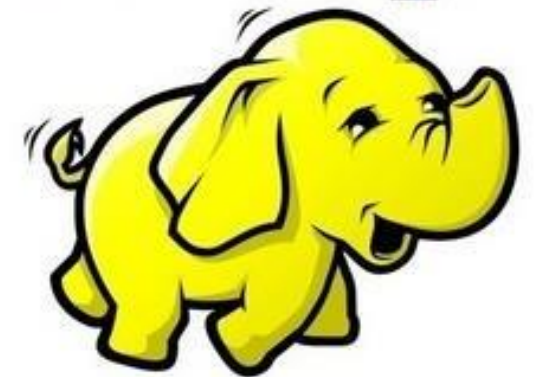


SPARK: definición y aportaciones

- Spark es una plataforma de computación para clústers (grandes cantidades de datos)
- Trabaja en memoria (más rápido)
- Es de propósito general. Distintas funcionalidades integradas (SQL, ML, grafos, etc.)
- Desarrollo simplificado (simplicidad de las APIs)
- Versatilidad de lenguajes (APIs Python, R, Java, Scala) frente a Hadoop (Java)
- Ejecución Batch, interactiva, streaming vs solo batch en Hadoop



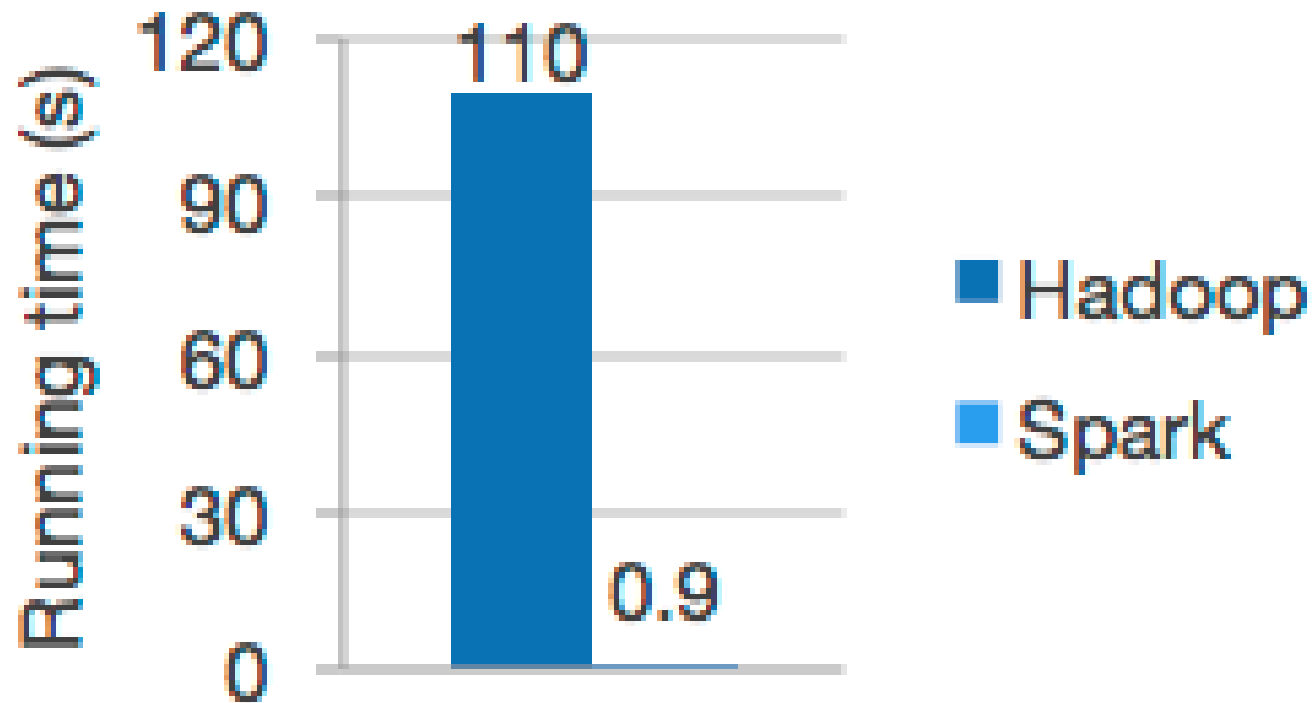
hadoop



SPARK: “demoledor” velocidad

- Puede ser hasta 100x más rápido que Hadoop

Logistic regression in Hadoop and Spark



Hadoop vs. Spark: tarea ordenar un dataset

	Hadoop MR Record	Spark Record	Spark 1 PB
Data Size	102.5 TB	100 TB	1000 TB
Elapsed Time	72 mins	23 mins	234 mins
# Nodes	2100	206	190
# Cores	50400 physical	6592 virtualized	6080 virtualized
Cluster disk throughput	3150 GB/s (est.)	618 GB/s	570 GB/s
Sort Benchmark Daytona Rules	Yes	Yes	No
Network	dedicated data center, 10Gbps	virtualized (EC2) 10Gbps network	virtualized (EC2) 10Gbps network
Sort rate	1.42 TB/min	4.27 TB/min	4.27 TB/min
Sort rate/node	0.67 GB/min	20.7 GB/min	22.5 GB/min

SPARK: Java, Scala, Python, R

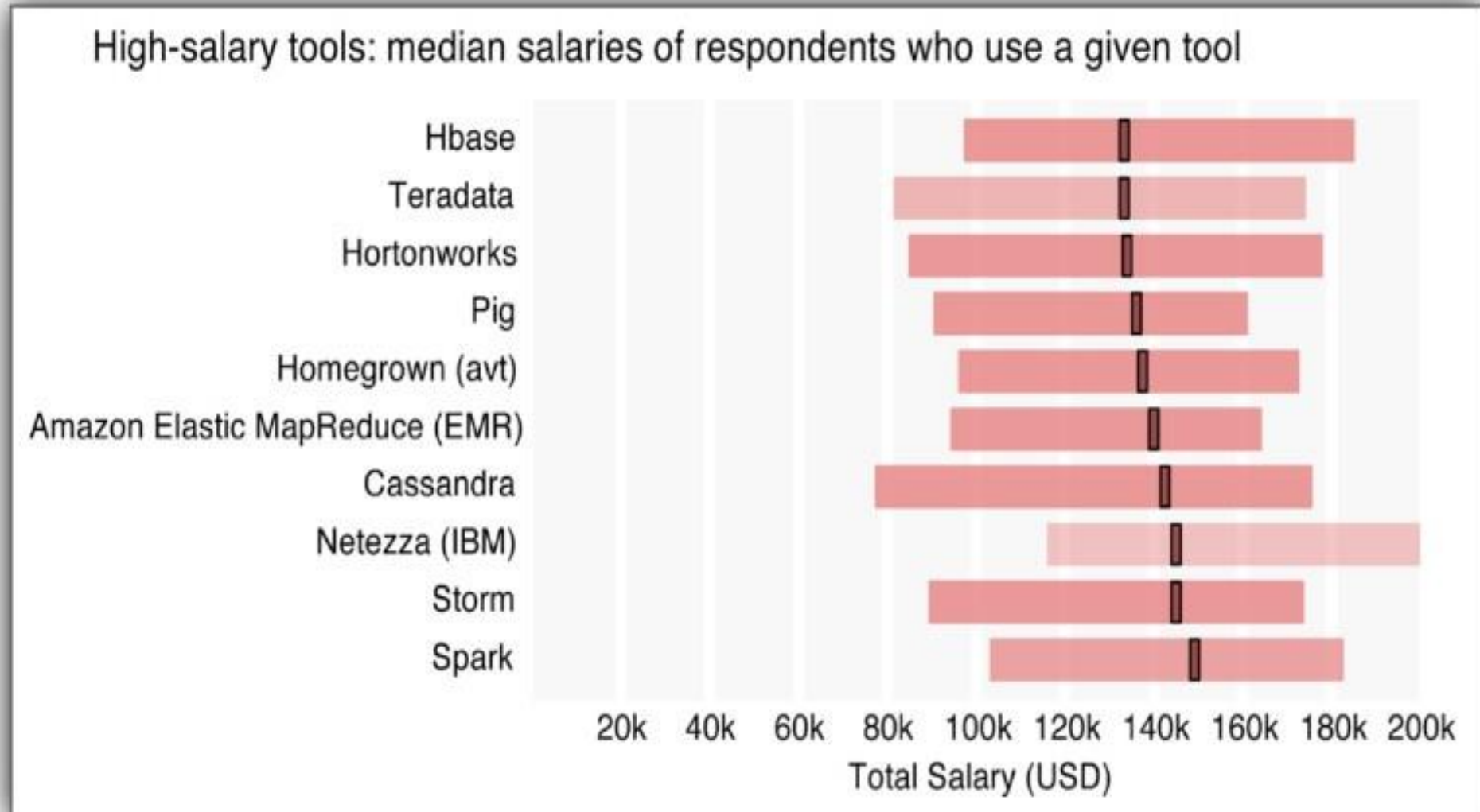
```
1 package org.myorg;
2
3 import java.io.IOException;
4 import java.util.*;
5
6 import org.apache.hadoop.fs.Path;
7 import org.apache.hadoop.conf.*;
8 import org.apache.hadoop.io.*;
9 import org.apache.hadoop.mapreduce.*;
10 import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
11 import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
12 import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
13 import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;
14
15 public class WordCount {
16
17     public static class Map extends Mapper<LongWritable, Text, Text, IntWritable> {
18         private final static IntWritable one = new IntWritable(1);
19         private Text word = new Text();
20
21         public void map(LongWritable key, Text value, Context context) throws IOException {
22             String line = value.toString();
23             StringTokenizer tokenizer = new StringTokenizer(line);
24             while (tokenizer.hasMoreTokens()) {
25                 word.set(tokenizer.nextToken());
26                 context.write(word, one);
27             }
28         }
29     }
30
31     public static class Reduce extends Reducer<Text, IntWritable, Text, IntWritable> {
32
33         public void reduce(Text key, Iterable<IntWritable> values, Context context)
34             throws IOException, InterruptedException {
35             int sum = 0;
36             for (IntWritable val : values) {
37                 sum += val.get();
38             }
39             context.write(key, new IntWritable(sum));
40         }
41     }
42
43     public static void main(String[] args) throws Exception {
44         Configuration conf = new Configuration();
45
46         Job job = new Job(conf, "wordcount");
47
48         job.setOutputKeyClass(Text.class);
49         job.setOutputValueClass(IntWritable.class);
50
51         job.setMapperClass(Map.class);
52         job.setReducerClass(Reduce.class);
53
54         job.setInputFormatClass(TextInputFormat.class);
55         job.setOutputFormatClass(TextOutputFormat.class);
56
57         FileInputFormat.addInputPath(job, new Path(args[0]));
58         FileOutputFormat.setOutputPath(job, new Path(args[1]));
59
60         job.waitForCompletion(true);
61     }
62
63 }
```

Contar palabras en Hadoop (Java)

Contar palabras en Spark (Python API)

```
text_file = sc.textFile("hdfs://...")
counts = text_file.flatMap(lambda line: line.split(" ")) \
    .map(lambda word: (word, 1)) \
    .reduceByKey(lambda a, b: a + b)
counts.saveAsTextFile("hdfs://...")
```

SPARK: Salarios



800 respuestas de 53 países y 41 estados de EEUU



red.es



UNIÓN EUROPEA

"El FSE invierte en tu futuro"

Fondo Social Europeo

