BIG DATA

•••

by Miguel Carrizo

Introducción
Big Data

Unidades de Medidas de Información

Símbolo	Denominación	10^n
b	Byte	10^0
Kb	Kilobyte	10^3
Mb	Megabyte	10^6
Gb	Gigabyte	10^9
Tb	Terabyte	10^{12}
Pb	Petabyte	10^{15}
Eb	Exabyte	10^{18}
Zb	Zettabyte	10^{21}
Yb	Yottabyte	10^{24}
Bb	Brontobyte	10^{27}
Gb	Geopbyte	10^{30}
Sb	Saganbyte	10^{33}
Jb	Jotabyte	10^{36}

Definición de Big Data

Conjunto de tecnologías que permiten
...
de grandes conjuntos de datos distribuidos.

- Recopilación
- Almacenamiento
- Gestión
- Análisis
- Visualización

Tecnologías diseñadas para el tratamiento de datos

- BATCH – por lotes
- STREAMING – Tiempo real

Datos

- Estructurados
- Semiestructurados
- No Estructurados

Tipos de procesamiento del Big Data

Procesamiento por lotes BATCH

Permite procesar volúmenes de datos en tiempos espaciados, por ejemplo cada 10 minutos, 1 hora o diario.

El sistema dispone de lotes o batch en el que almacena toda la información que va obteniendo hasta completar un periodo.

Procesamiento en tiempo real - STREAMING

Permite procesar volúmenes de datos en tiempos lo más parecido a tiempo real que se pueda, hablamos de ordenes de 100 mili segundos a segundos.

Tipos de datos en Big Data

Datos Estructurados

	nombre	color	edad	altura	peso
1:	Paco	Rojo	24	182	74.8
2:	Juan	Green	30	170	70.1
3:	Andres	Amarillo	41	169	60.0
4:	Natalia	Green	22	183	75.0
5:	Vanesa	Verde	31	178	83.9
6:	Miriam	Rojo	35	172	76.2
7:	Juan	Amarillo	22	164	68.0

Perfectamente definido la longitud, el formato y el tamaño de sus datos.

Se almacenan en formato tabla, hojas de cálculo o en bases de datos relacionales.

No estructurados

CAPÍTULO PRIMERO

Que trata de la condición y ejercicio del famoso hidalgo D. Quijote de la Mancha

En un lugar de la Mancha, de cuyo nombre no quiero acordarme, no ha mucho tiempo que vivía un hidalgo de los de lanza en astillero, adarga antigua, rocín flaco y galgo corredor. Una olla de algo más vaca que carnero, salpicón las más noches, duelos y quebrantos los sábados, lentejas los viernes, algún palomino de añadidura los domingos, consumían las tres partes de su hacienda. El resto della concluían sayo de velarte, calzas de velludo para las fiestas con sus pantuflos de lo mismo, los días de entre semana se honraba con su vellorí de lo más fino. Tenía en su casa una ama que pasaba de los cuarenta, y una sobrina que no llegaba a los veinte, y un mozo de campo y plaza, que así ensillaba el rocín como tomaba la podadera. Frisaba la edad de nuestro hidalgo con los cincuenta años, era de complexión recia, seco de carnes, enjuto de rostro; gran madrugador y amigo de la caza. Quieren decir que tenía el sobrenombre de Quijada o Quesada (que en esto hay alguna diferencia en los autores que deste caso escriben), aunque por conjeturas verosímiles se deja entender que se llama Quijana; pero esto importa poco a nuestro cuento; basta que en la narración dél no se salga un punto de la verdad.

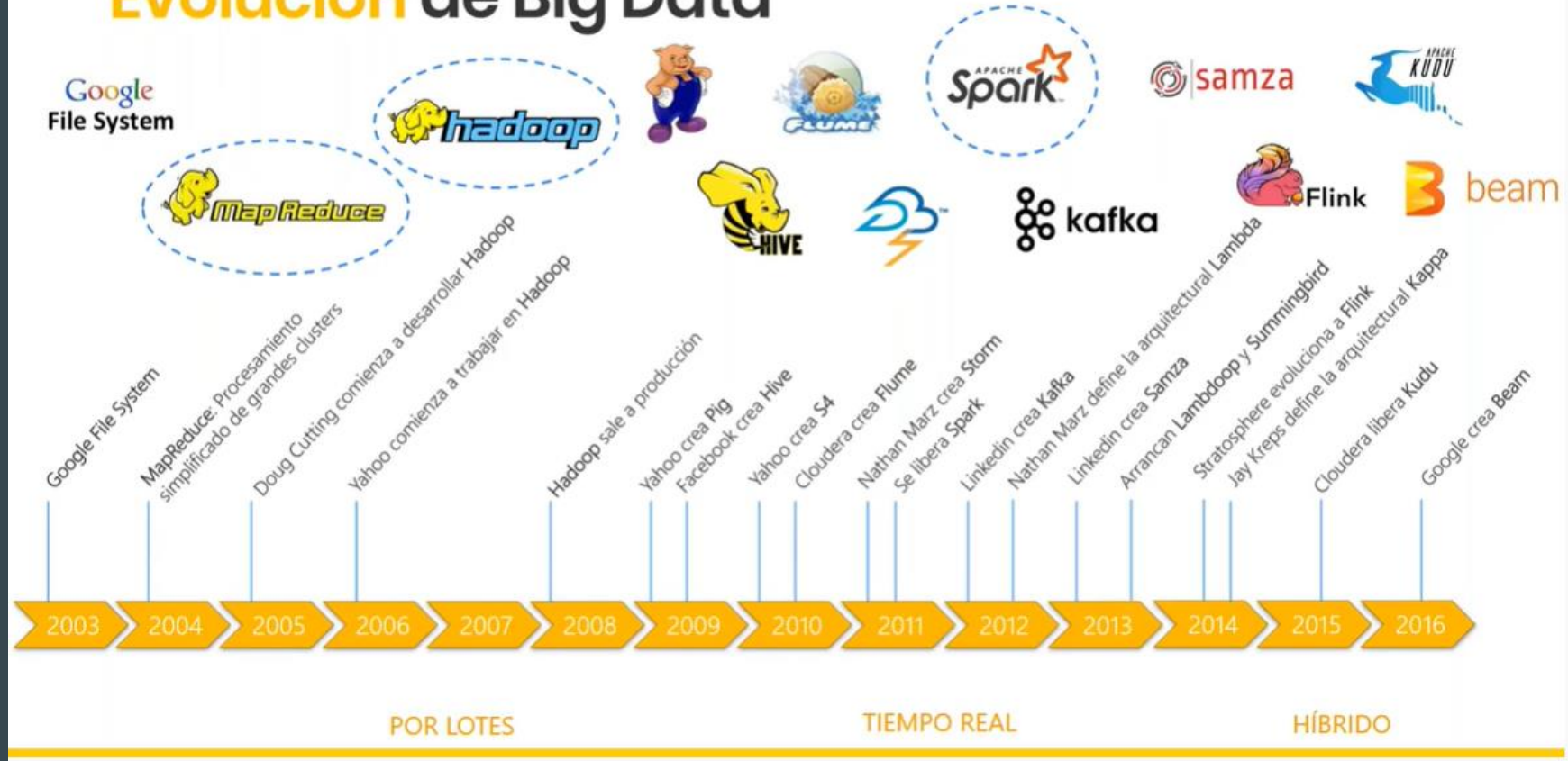
No tienen un formato específico.
Se almacenan en múltiples formatos como documentos PDF o Word, correos electrónicos, ficheros multimedia de imagen, audio o video,...

Semiestructurados

```
{
    "marcadores": [
      {
        "latitude": 40.416875,
        "longitude": -3.703308,
        "city": "Madrid",
        "description": "Puerta del Sol"
      },
      {
        "latitude": 40.417438,
        "lonaitude": -3.693363,
      },
    ]
}
```

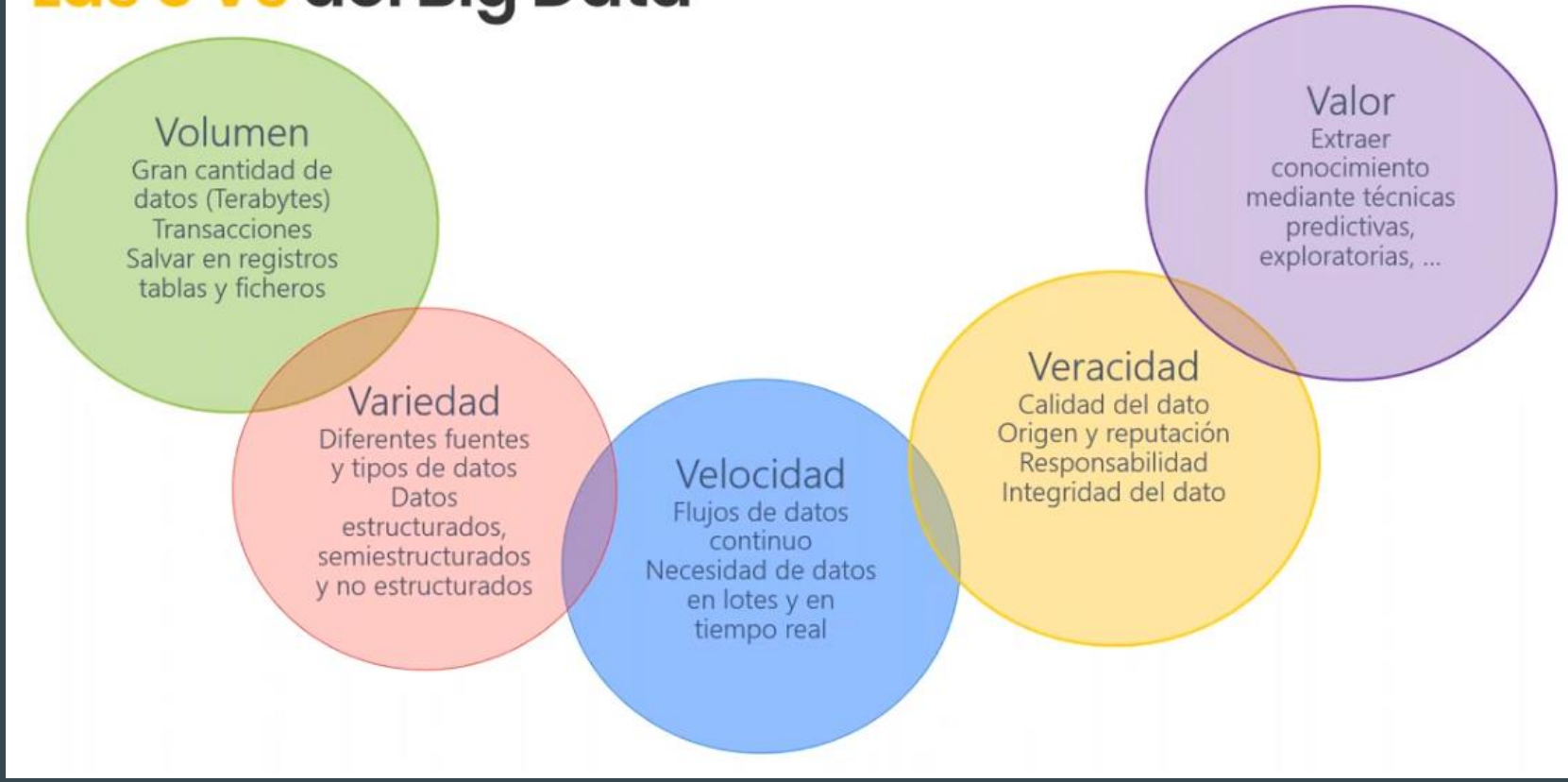
Mezcla de los anteriores, presentan una estructura flexible.
Se almacenan en formatos HTML, XML o JSON.

Evolución de Big Data

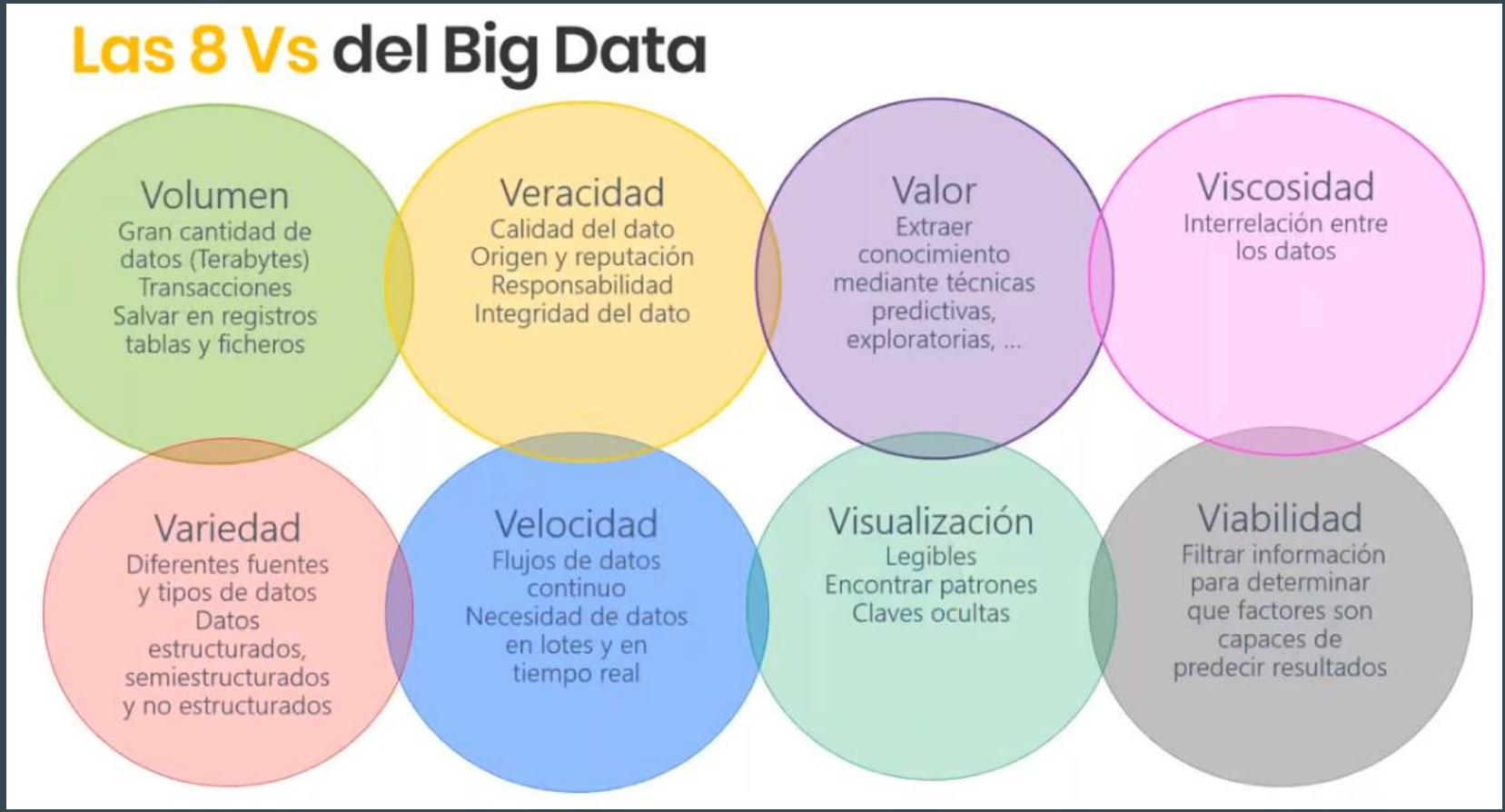


Características
Big Data

Las 5 Vs del Big Data



Las 8 Vs del Big Data



Herramientas
Big Data

Hadoop - Definición

Apache Hadoop es un sistema distribuido que permite realizar procesamiento de grandes volúmenes de datos a través de clúster, fácil de escalar.

A grandes rasgos se puede decir que Hadoop está compuesto por dos partes:

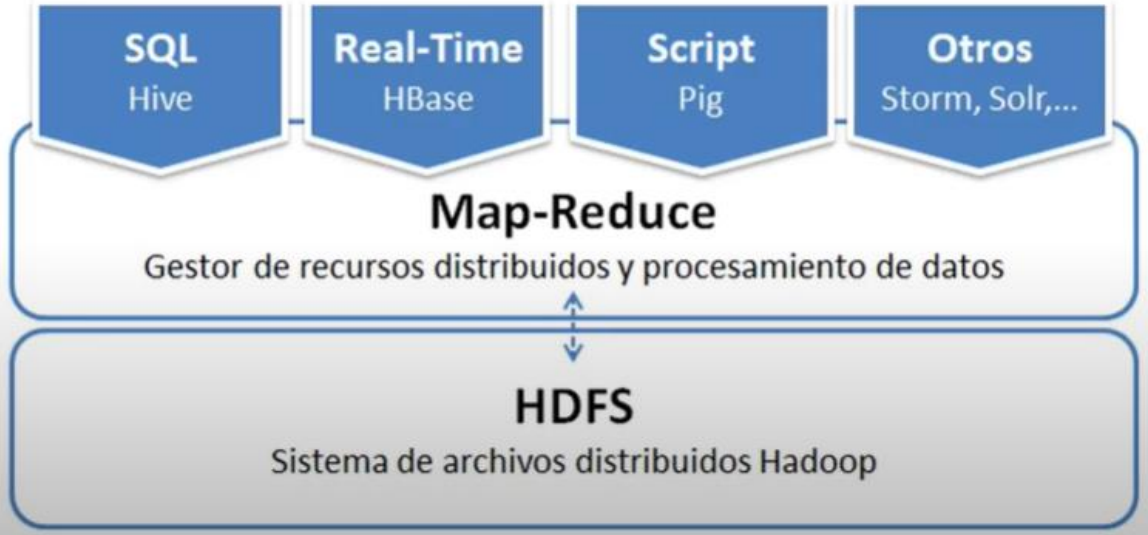
1. Se ocupa del **almacenamiento de datos** de distintos tipos (HDFS)
2. Realiza las tareas de **procesamiento de los datos** de manera distribuida (MapReduce).

Hadoop esta basado en una arquitectura maestro-esclavo o **Master-Slave**.

Hadoop es ampliamente utilizado en Big Data, porque trabajar con **grandes volúmenes** de información a **muy bajo coste**.

Hadoop - Arquitectura

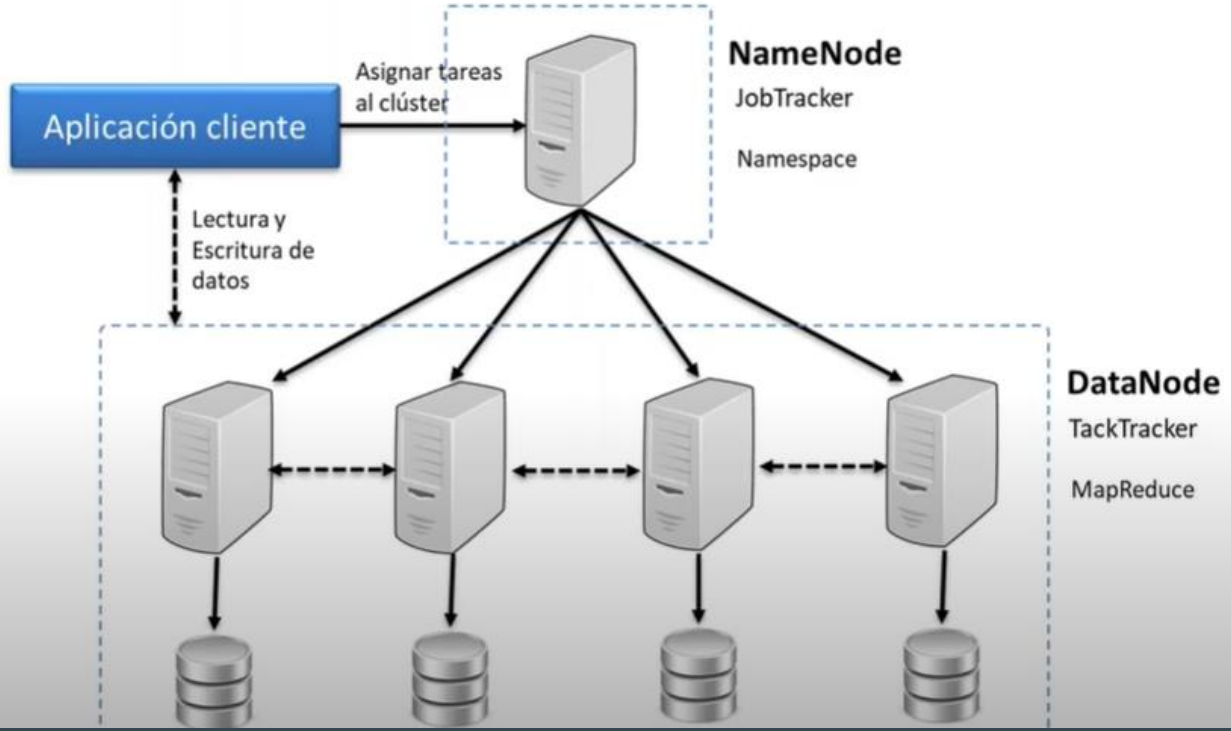
Básica



Hadoop - Arquitectura

Evolución de arquitectura con Yarn

Hadoop - Componentes



Hadoop - Componentes

- **NameNode**: es el kernel de sistemas Hadoop y se encarga de gestionar el cluster. Conoce la ubicación de los datanodes y el contenido de estos. Almacena toda esa información es un espacio de nombres o namespace.
- **Secondary NameNode** (opcional): se encarga de replicar periódicamente el *namespace del datanode y en caso de que falle pasa a sustituirlo.*
- **DataNodes**: se encargan de almacenar físicamente los bloques de datos en el cluster y de entregar la información cuando se la soliciten.
- **JobTracker**: se encarga de coordinar los trabajos (jobs) solicitados, para ello crea tareas MapReduce y las asigna a los TaskTrackers de los DataNodes.
- **TaskTracker**: se encarga de ejecutar las tareas MapReduce asignadas por el JobTracker y reportar el estado de la tarea a este.

Spark - Definición

Apache Spark es un sistema de computación distribuida de software libre, que permite procesar grandes conjuntos de datos sobre un conjunto de máquinas de forma simultánea, proporcionando escalabilidad horizontal y la tolerancia a fallos.

Para cumplir con estas características proporciona un modelo de desarrollo de programas que permite ejecutar código de forma distribuida de tal manera que cada máquina se ocupe de realizar una parte de la tarea y entre todos realicen la tarea global.

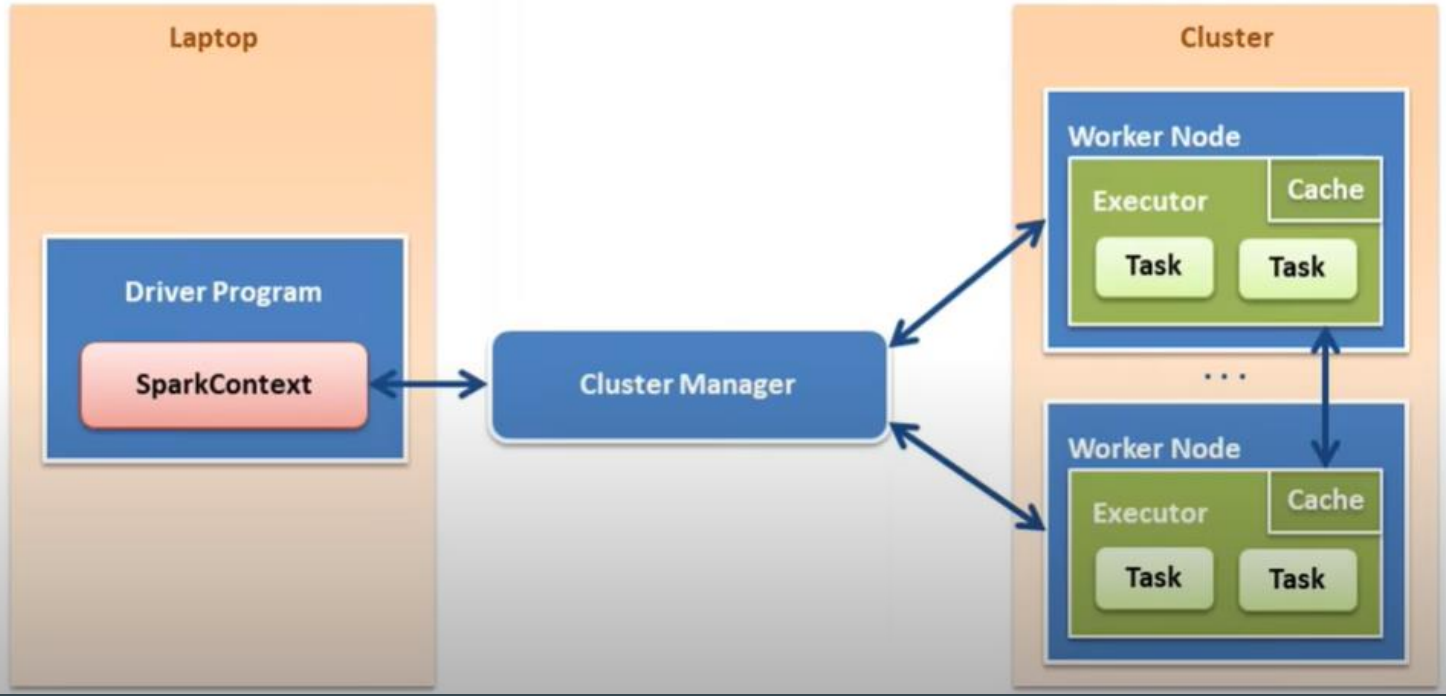
Spark – Características

Utiliza **estructuras de datos RDD** (Resilient Data Distributed) que son conjunto de datos de solo lectura, que están distribuidos a lo largo del clúster, mantenidos de manera tolerante a fallos. La disponibilidad de RDDs facilita la implementación de algoritmos iterativos que accedan varias veces a los mismos datos y para el análisis exploratorio de datos.

Para su correcto funcionamiento Spark necesita:

- **Gestión de recursos**, soporta **Standalone**, YARN o **Apache Mesos**.
- **Sistema de ficheros distribuido**, soporta HDFS, **Cassandra** o **Kudu**.

Spark - Arquitectura



Spark - Componentes

Spark Core: es en núcleo donde se apoya toda la arquitectura, proporciona:

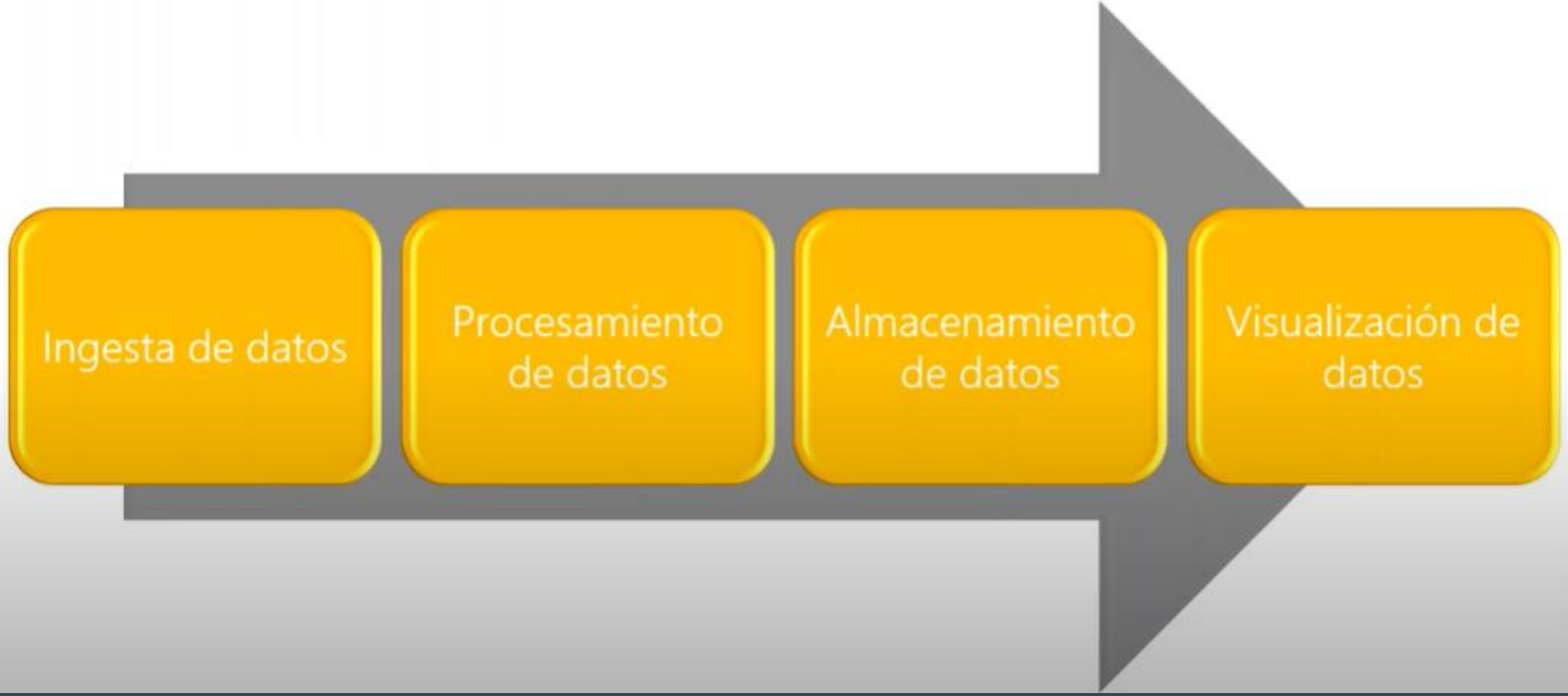
Spark SQL: introduce un concepto de abstracción de datos llamado SchemaRD, que proporciona soporte para datos estructurados y semi-estructurados.

Spark Streaming: es la capa encargada del análisis de datos en tiempo real.

MLlib: es la capa de aprendizaje automático distribuido sobre el core, que proporciona el famework de aprendizaje automático donde se pueden encontrar multitud de algoritmos de clasificación, regresión, análisis cluster, reducción de dimensionalidad y estadísticos descriptivos.

GraphX: es la capa de procesamiento gráfico distribuido sobre el core. Al basarse en RDDs inmutables, los gráficos no permiten actualizarse.

Herramientas Big Data



Herramientas de Ingesta de datos

Ingesta - Sqoop

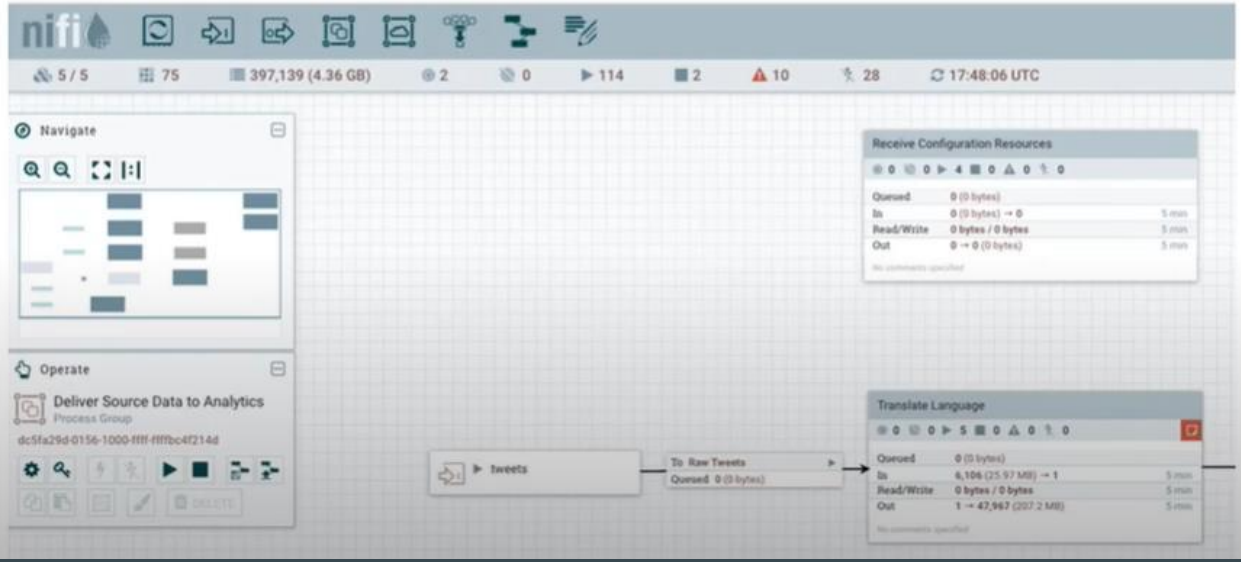
Herramienta de línea de comandos desarrollada para transferir grandes volúmenes de datos de bases de datos relacionarles a Hadoop, de ahí su nombre que viene de la fusión de SQL y Hadoop. Concretamente transforma datos relacionarles en Hive o HBase en una dirección y en la otra de HDFS a datos relacionares como MySQL, Oracle, Postgress o a un data warehouse.

Ejemplo

```
$ sqoop import
  --connect jdbc:mysql://localhost/mibbdd
  --username=root -P
  --table=mitabla
  --driver=com.mysql.jdbc.Driver
  --target-dir=/mitabla_hdfs
  --fields-terminated-by=','
  --lines-terminated-by '\n'
```

Ingesta - Nifi

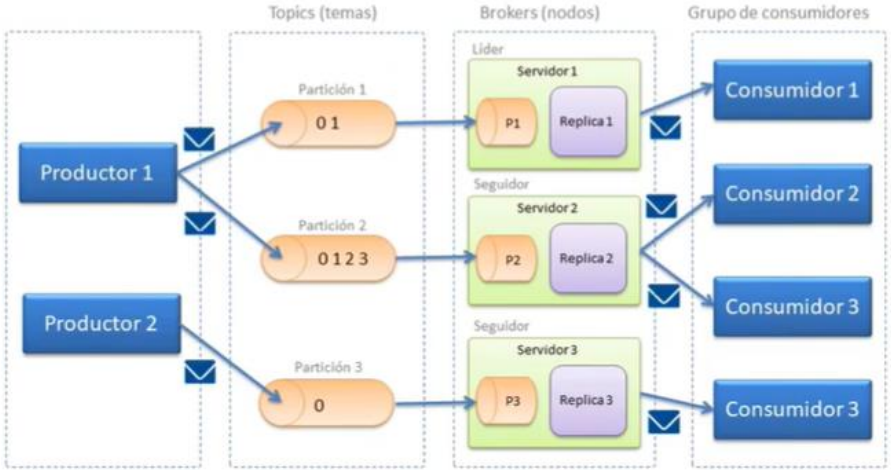
Plataforma integrada de **procesamiento y logística** de datos en **tiempo real**, para **automatizar el movimiento de datos** entre diferentes sistemas de forma rápida, fácil y segura.



Ingesta - Kafka

Sistema de intermediación de mensajes basado en el modelo publicador/suscriptor.

Se considera un sistema persistente, escalable, replicado y tolerante a fallos. A estas características se añade la velocidad de lecturas y escrituras que lo convierten en una herramienta excelente para comunicaciones en tiempo real (streaming).



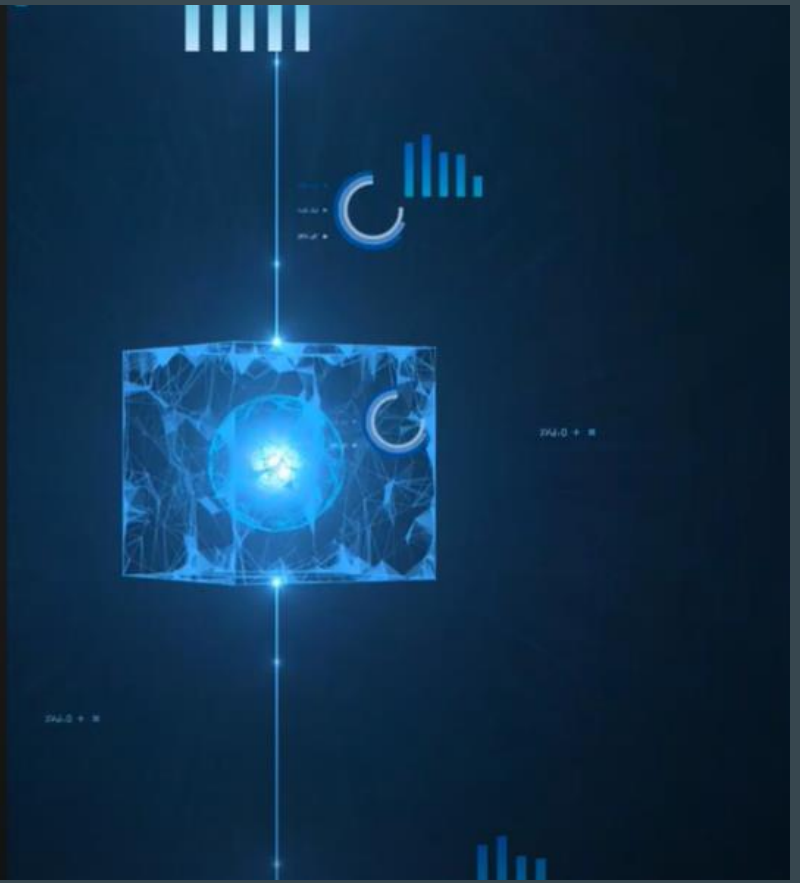
Herramientas de Procesamiento de datos

Herramientas de Almacenamiento de datos

Herramientas de Visualización de datos

Google Cloud Platform

Aplicaciones prácticas Big Data

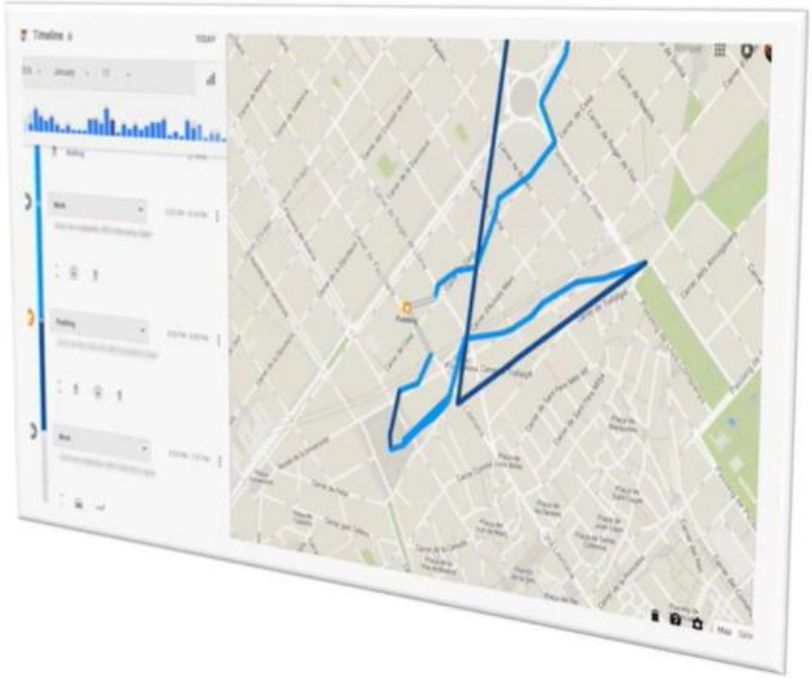


Ejemplo - Segmentación de clientes por ubicación

Una de las áreas de mayor aplicación del Big Data es la de Marketing y Ventas.

El poder proporcionar información que nos permita segmentar clientes por ciertos criterios hace que las empresas mejoren significativamente sus ventas.

Por ejemplo si somos capaces de saber la ubicación podremos tener información de sus hábitos y con ello podremos sacar perfiles de consumidor.



Ejemplo - Trading financiero

El área financiera utiliza de forma habitual el Big Data para poder analizar el mayor número de variables para poder tomar mejores decisiones en la compra venta de acciones.

Ejemplo - Coche autónomo

El área del transporte está empezando a utilizar Big Data para dar respuesta la conducción autónoma.

Ejemplo - Detección de enfermedades

El área de salud utiliza de forma habitual el Big Data para detectar enfermedades a partir del análisis masivo de cadenas de ADN.

THE END

Miguel Carrizo
mcarrizo@ies21.edu.ar