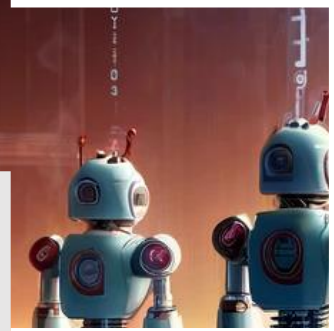
THE BRIDGE

Big Data (III)
Frameworks: Hadoop



Hadoop

Definición

Hadoop

Definición

*"**Apache Hadoop** es un framework de código abierto que permite el almacenamiento distribuido y el procesamiento de grandes conjuntos de datos en base a un hardware comercial"*

Hadoop

Definición

*"**Apache Hadoop** es un framework de código abierto que permite el almacenamiento distribuido y el procesamiento de grandes conjuntos de datos en base a un hardware comercial"*

Volumen

Sirve para almacenar grandes volúmenes de información

Hadoop

Definición

*"**Apache Hadoop** es un framework de código abierto que permite el almacenamiento distribuido y el procesamiento de grandes conjuntos de datos en base a un hardware comercial"*

Volumen

Sirve para almacenar grandes volúmenes de información

Backups

Guarda copias de la información en diferentes nodos

Hadoop

Definición

*"**Apache Hadoop** es un framework de código abierto que permite el almacenamiento distribuido y el procesamiento de grandes conjuntos de datos en base a un hardware comercial"*

Volumen
Sirve para almacenar grandes volúmenes de información

Backups
Guarda copias de la información en diferentes nodos

Tolerancia a fallos
En caso de que se caiga un nodo, cuenta con otros para mantener el servicio

Hadoop

Definición

*"**Apache Hadoop** es un framework de código abierto que permite el almacenamiento distribuido y el procesamiento de grandes conjuntos de datos en base a un hardware comercial"*

Volumen
Sirve para almacenar grandes volúmenes de información

Backups
Guarda copias de la información en diferentes nodos

Tolerancia a fallos
En caso de que se caiga un nodo, cuenta con otros para mantener el servicio

YARN
Gestor de recursos de Hadoop

Hadoop

Definición

*"**Apache Hadoop** es un framework de código abierto que permite el almacenamiento distribuido y el procesamiento de grandes conjuntos de datos en base a un hardware comercial"*

Volumen
Sirve para almacenar grandes volúmenes de información

Backups
Guarda copias de la información en diferentes nodos

Tolerancia a fallos
En caso de que se caiga un nodo, cuenta con otros para mantener el servicio

YARN
Gestor de recursos de Hadoop

Escalabilidad
Es cuestión de añadir nuevos nodos, de hardware económico

Hadoop

Definición

*"**Apache Hadoop** es un framework de código abierto que permite el almacenamiento distribuido y el procesamiento de grandes conjuntos de datos en base a un hardware comercial"*

Volumen
Sirve para almacenar grandes volúmenes de información

Backups
Guarda copias de la información en diferentes nodos

Tolerancia a fallos
En caso de que se caiga un nodo, cuenta con otros para mantener el servicio

YARN
Gestor de recursos de Hadoop

Escalabilidad
Es cuestión de añadir nuevos nodos, de hardware económico

HDFS

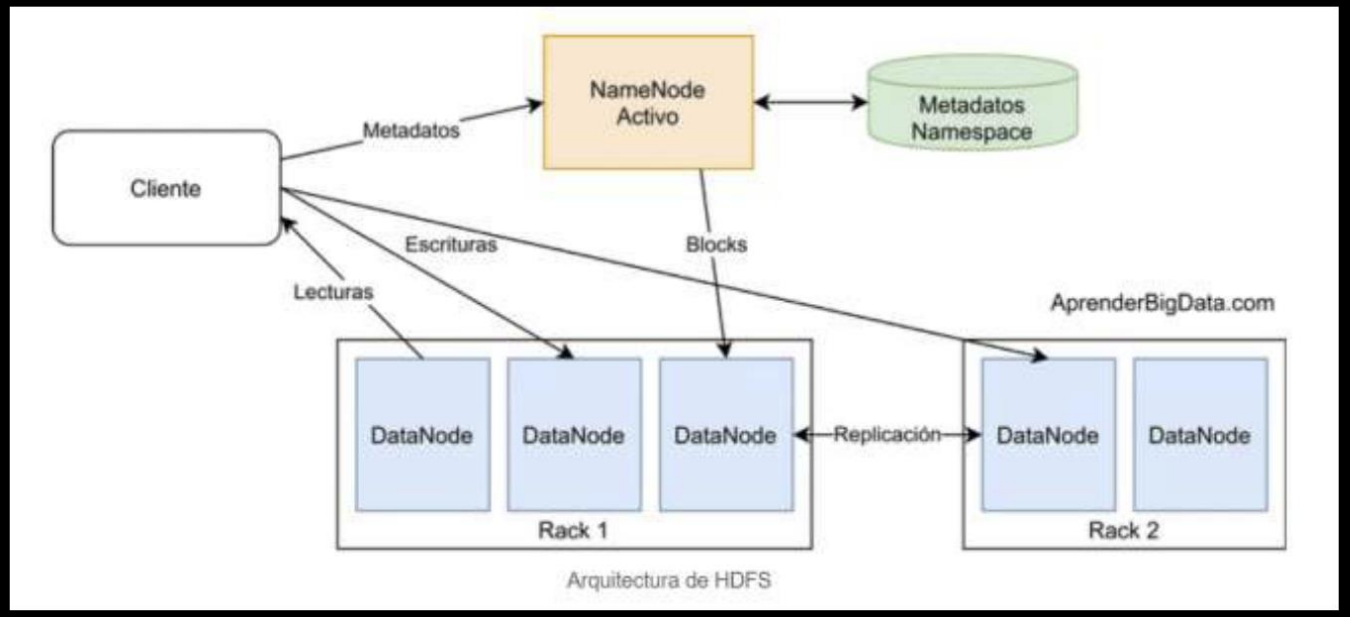
High Distributed File System

“Sistema de ficheros distribuidos de Hadoop. Sirve para el almacenamiento masivo de información, tanto para datos estructurados, semi-estructurados y no estructurados.”

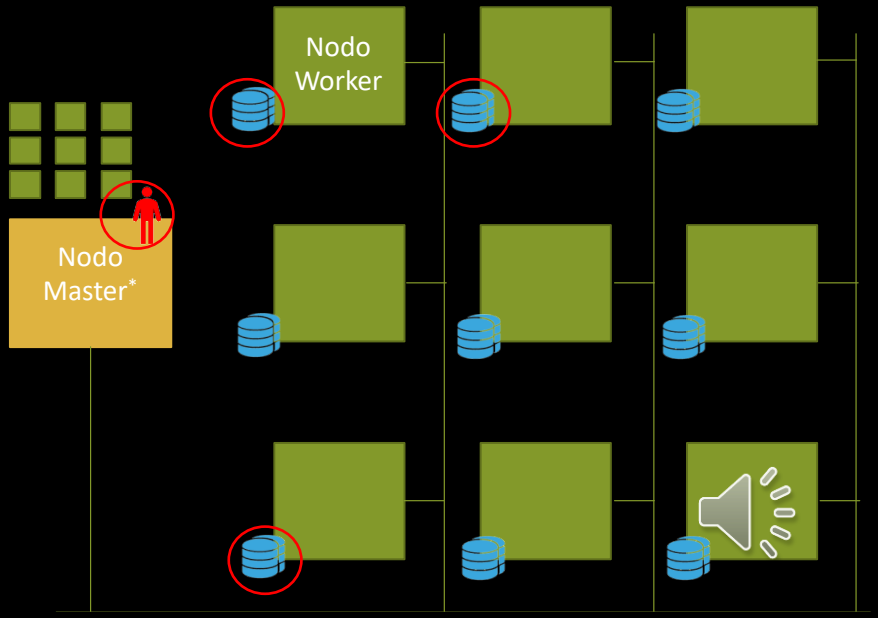
HDFS

High Distributed File System

"Sistema de ficheros distribuidos de Hadoop. Sirve para el almacenamiento masivo de información, tanto para datos estructurados, semi-estructurados y no estructurados."



Arquitectura de HDFS



Parquet files

*Formato de **almacenamiento de datos** orientado a **columnas**, lo que facilita el uso de una codificación y compresión eficientes para reducir tu tamaño. Es gratuito y de código abierto, característico del ecosistema Hadoop.*

Dataset	Columns	Size on Amazon S3	Data Scanned	Cost
Data stored as CSV file	4	4TB	4TB	**$20** (4TB x $5/TB)
Data stored as GZIP CSV file	4	1TB	1TB	**$5** (1TB x $5/TB)
Data stored as Parquet file	4	1TB	**.25TB**	**$1.25** (.25TB x $5/TB)

Map Reduce

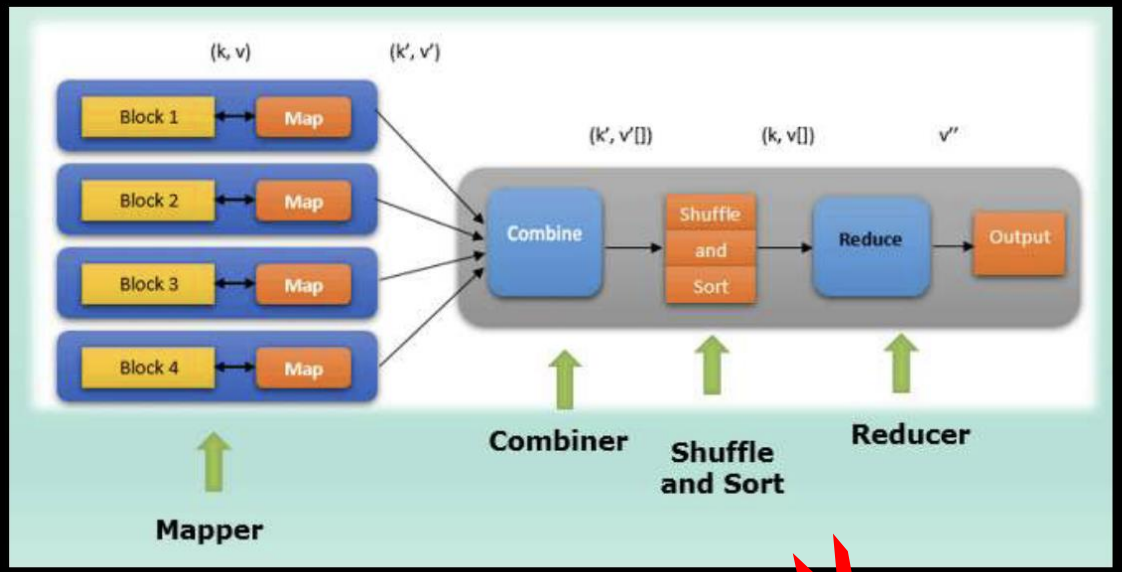
Paradigma de programación

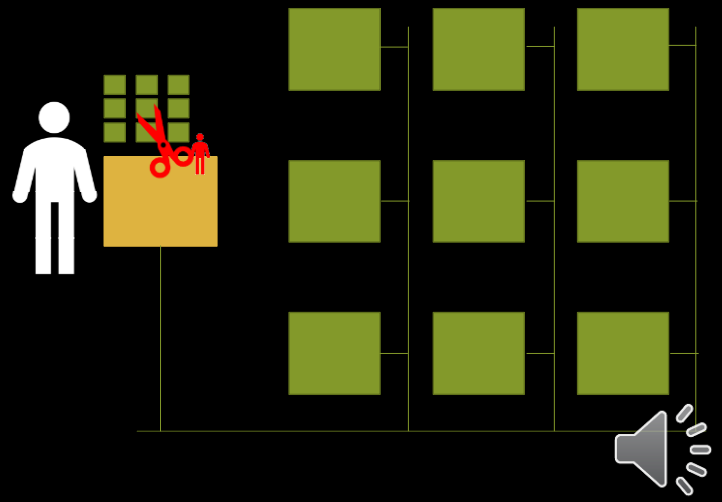
“MapReduce es una técnica de procesamiento y un programa modelo de computación distribuida basada en java. Mediante el Map se generan pares clave-valor y en el Reduce se produce la agregación.”

Map Reduce

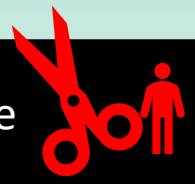
Paradigma de programación

"MapReduce es una técnica de procesamiento y un programa modelo de computación distribuida basada en java. Mediante el Map se generan pares clave-valor y en el Reduce se produce la agregación."





La "primera" versión, junto con YARN, de



Map Reduce

Paradigma de programación

Map Reduce

Paradigma de programación

Fase Map

1.División de Datos: Los datos de entrada se dividen en fragmentos (splits) que se asignan a diferentes nodos del clúster.

2.Procesamiento de Map: Cada nodo aplica una función de mapeo a su fragmento de datos. La función de mapeo transforma los datos en pares clave-valor.

Map Reduce

Paradigma de programación

Fase Map

1.División de Datos: Los datos de entrada se dividen en fragmentos (splits) que se asignan a diferentes nodos del clúster.

2.Procesamiento de Map: Cada nodo aplica una función de mapeo a su fragmento de datos. La función de mapeo transforma los datos en pares clave-valor.

Paso 1: División de Datos

- Doc1: "apple orange apple"
- Doc2: "orange banana banana"
- Doc3: "banana apple orange"

Estos documentos se dividen en fragmentos y se distribuyen a diferentes nodos.

Map Reduce

Paradigma de programación

Fase Map

1.División de Datos: Los datos de entrada se dividen en fragmentos (splits) que se asignan a diferentes nodos del clúster.

2.Procesamiento de Map: Cada nodo aplica una función de mapeo a su fragmento de datos. La función de mapeo transforma los datos en pares clave-valor.

Paso 1: División de Datos

•Doc1: "apple orange apple"
•Doc2: "orange banana banana"
•Doc3: "banana apple orange"
Estos documentos se dividen en fragmentos y se distribuyen a diferentes nodos.

Paso 2: Función Map

Cada nodo aplica la función de mapeo a su fragmento:
•Nodo 1 (Doc1): [("apple", 1), ("orange", 1), ("apple", 1)]
•Nodo 2 (Doc2): [("orange", 1), ("banana", 1), ("banana", 1)]
•Nodo 3 (Doc3): [("banana", 1), ("apple", 1), ("orange", 1)]

Map Reduce

Paradigma de programación

Fase Reduce

1.Agrupamiento: Todos los pares clave-valor generados por la fase de mapeo se agrupan por clave.

2.Procesamiento de Reduce: Cada grupo de claves se pasa a una función de reducción que combina los valores asociados con esa clave para producir un resultado final.

Map Reduce

Paradigma de programación

Fase Reduce

1.Agrupamiento: Todos los pares clave-valor generados por la fase de mapeo se agrupan por clave.

2.Procesamiento de Reduce: Cada grupo de claves se pasa a una función de reducción que combina los valores asociados con esa clave para producir un resultado final.

Paso 3: Agrupamiento

Los pares clave-valor se agrupan por clave:

- "apple": [1, 1, 1]
- "orange": [1, 1, 1]
- "banana": [1, 1, 1]

Map Reduce

Paradigma de programación

Fase Reduce

1.Agrupamiento: Todos los pares clave-valor generados por la fase de mapeo se agrupan por clave.

2.Procesamiento de Reduce: Cada grupo de claves se pasa a una función de reducción que combina los valores asociados con esa clave para producir un resultado final.

Paso 3: Agrupamiento

Los pares clave-valor se agrupan por clave:

- "apple": [1, 1, 1]
- "orange": [1, 1, 1]
- "banana": [1, 1, 1]

Paso 4: Función Reduce

Cada grupo se pasa a la función de reducción que suma los valores:

- "apple": 1 + 1 + 1 = 3
- "orange": 1 + 1 + 1 = 3
- "banana": 1 + 1 + 1 = 3

Map Reduce

Paradigma de programación

Fase Reduce

1. **Agrupamiento**: Todos los pares clave-valor generados por la fase de mapeo se agrupan por clave.

2. **Procesamiento de Reduce**: Cada grupo de claves se pasa a una función de reducción que combina los valores asociados con esa clave para producir un resultado final.

Paso 3: Agrupamiento

Los pares clave-valor se agrupan por clave:

- "apple": [1, 1, 1]
- "orange": [1, 1, 1]
- "banana": [1, 1, 1]

Paso 4: Función Reduce

Cada grupo se pasa a la función de reducción que suma los valores:

- "apple": 1 + 1 + 1 = 3
- "orange": 1 + 1 + 1 = 3
- "banana": 1 + 1 + 1 = 3

Resultado Final

La frecuencia de palabras es:

- "apple": 3
- "orange": 3
- "banana": 3

Map Reduce

Paradigma de programación

Fase Map

1.División de Datos: Los datos de entrada se dividen en fragmentos (splits) que se asignan a diferentes nodos del clúster.

2.Procesamiento de Map: Cada nodo aplica una función de mapeo a su fragmento de datos. La función de mapeo transforma los datos en pares clave-valor.

Fase Reduce

1.Agrupamiento: Todos los pares clave-valor generados por la fase de mapeo se agrupan por clave.

2.Procesamiento de Reduce: Cada grupo de claves se pasa a una función de reducción que combina los valores asociados con esa clave para producir un resultado final.


THE
BRIDGE