Instituto Politécnico Nacional
Escuela Superior de Cómputo

INGESTA DE DATOS

De Luna Ocampo Yanina
Medina Barreras Daniel Iván

Contenidos

01
Principios de Minería de Datos

02
Principios de ETL

03
Hadoop (HDFS)

04
Map Reduce

01

Principios de Minería de Datos

¿Qué es? ¿Cuáles son?

¿Qué es la minería de datos?

Introduction to data mining:
Es el proceso de descubrir automáticamente información útil en grandes repositorios de datos. Sus técnicas se implementan con el fin de encontrar patrones nuevos.

Principios fundamentales

Comprensión del negocio

Selección y preparación de datos

Exploración de datos

Modelado

Evaluación de modelos

Implementación

Comprensión del negocio

Entender el contexto empresarial y los objetos que se buscan alcanzar.

Selección y preparación de datos

Identificamos los datos que son relevantes para el análisis, los limpiamos y transformamos a un formato adecuado.

Exploración de datos

Identifica patrones y tendencias en los datos. Se pueden utilizar técnicas de visualización.

Modelado

Implica técnicas de análisis estadístico para crear modelos predictivos y descriptivos.

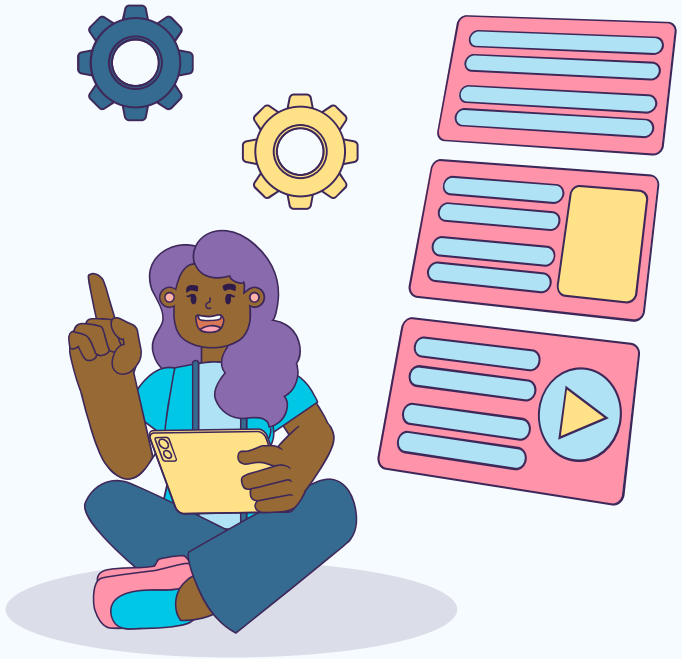
Evaluación de modelos

Importante para determinar su precisión y utilidad para determinar si son precisos y confiables.

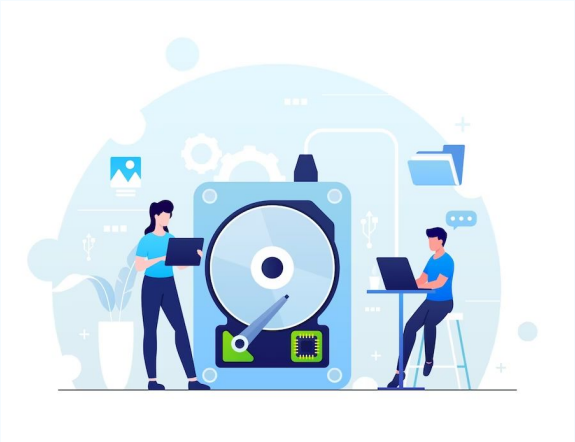
Implementación

Integrar el modelo al proceso de negocio e informar y mejorar la toma de decisiones empresariales.

EXTRA



Mantenimiento



Monitorización y actualización de los modelos a medida que cambian los datos y condiciones de negocio.

02

Principios de ETL

¿Cuáles son?

¿Qué es ETL?

Extracción, transformación y carga (ETL) es el proceso consistente en combinar datos de diferentes orígenes un gran repositorio central llamado almacenamiento de datos.

ETL utiliza un conjunto de reglas comerciales para limpiar y organizar datos en bruto y prepararlos para el almacenamiento, el análisis de datos y el machine learning (ML).

Los tres pasos distintos de ETL

Identificación de datos

Extracción

Almacén de datos de destino

Carga

Transformación

Asignar, transformar validar, autenticar, duplicar

¿Por qué es importante ETL?

Las organizaciones de hoy tienen muchos datos de varias fuentes, Al aplicar un proceso ETL, los conjuntos de datos en bruto se preparan en un formato y una estructura que son más consumibles para fines analíticos

¿Cómo beneficia ETL a la inteligencia empresarial?

Contexto histórico	Vista a largo plazo de los datos
Vista de datos consolidada	Facilita el análisis, la visualización y el sentido de los datos.
Análisis de datos preciso	Asegurando que los datos sean confiables.
Automatización de tareas	Dedicar más tiempo a innovar

¿Cómo ha evolucionado el ETL?

ETL tradicional

Almacen	Escalabilidad	Tipos de datos y fuentes de datos
Bases de datos transaccionales	Muy poca	Transaccionales

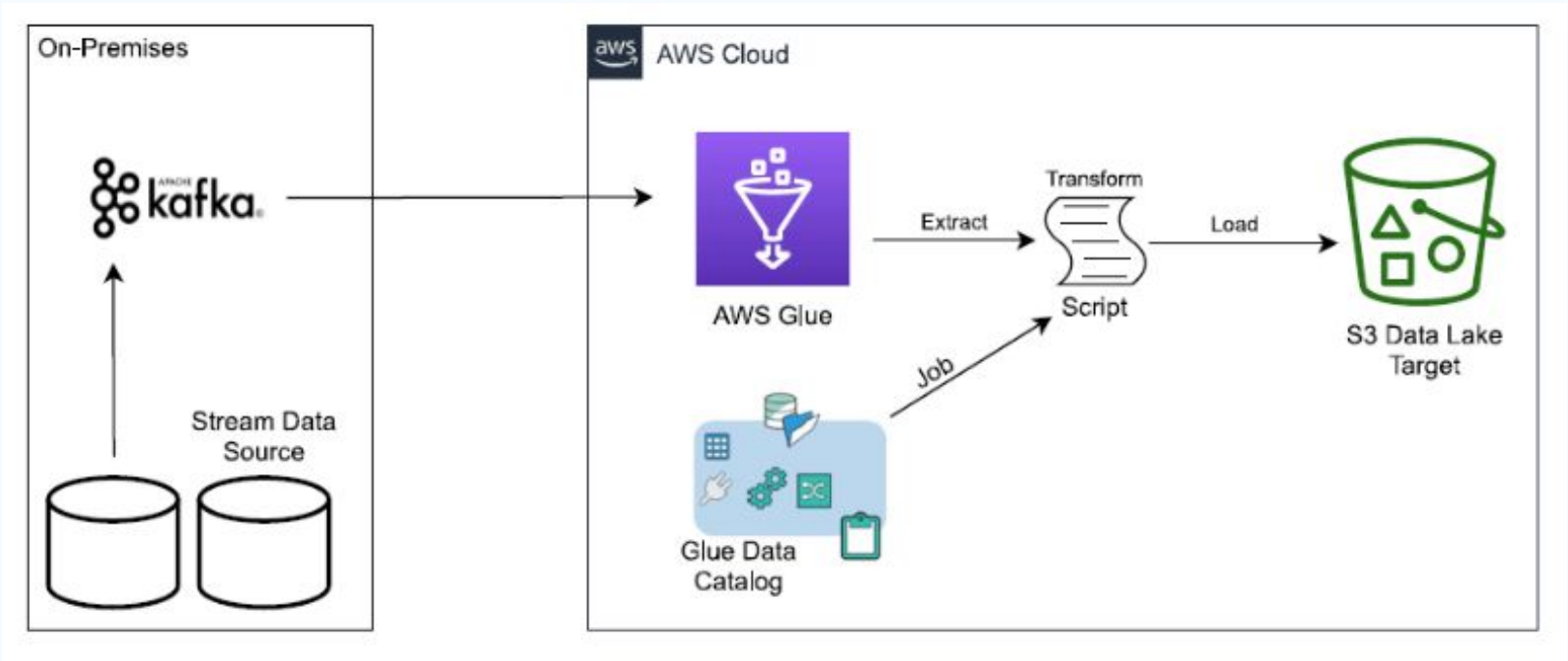
ETL moderno

Almacen	Escalabilidad	Tipos de datos y fuentes de datos
Cloud	Casi ilimitada	Todo

¿Cómo funciona la ETL?

La extracción, transformación y carga (ETL) funciona moviendo datos del sistema de origen al sistema de destino a intervalos periódicos. El proceso ETL funciona en tres pasos:

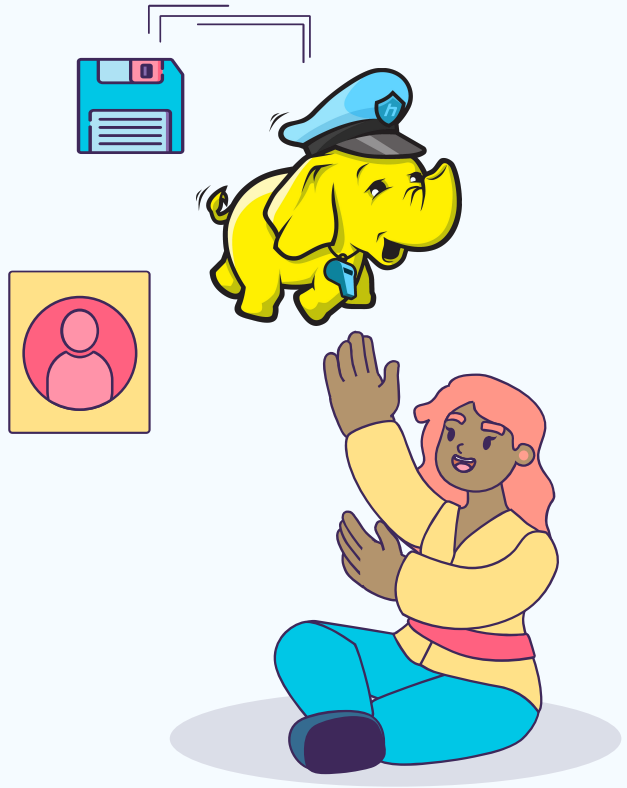
1. Extracción de los datos relevantes de la base de datos de origen.
2. Transformación de los datos para que sean más adecuados para el análisis.
3. Carga de los datos en la base de datos de destino.


On-Premises
kafka
Stream Data Source
AWS Cloud
AWS Glue
Extract
Transform
Script
Load
S3 Data Lake Target
Job
Glue Data Catalog

03

Hadoop Distributed File System

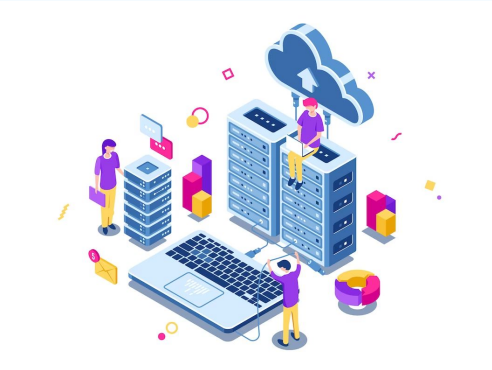
¿Qué es?

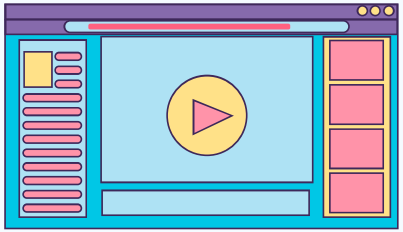


Hadoop

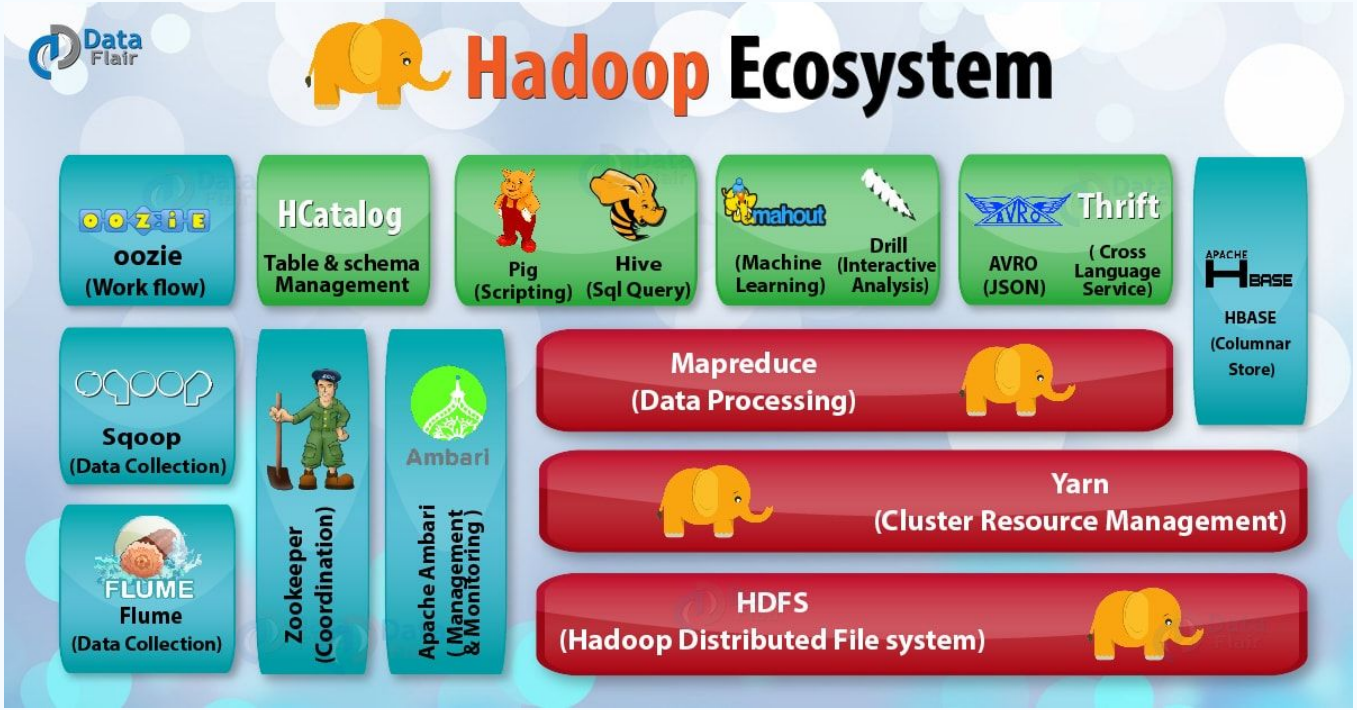
Es una infraestructura de código abierto que reúne todos los componentes necesarios para almacenar y analizar grandes cantidades de datos.

- Es de bajo costo inicial
- Capacidad de analizar datos a medida que reciben (Big Data)



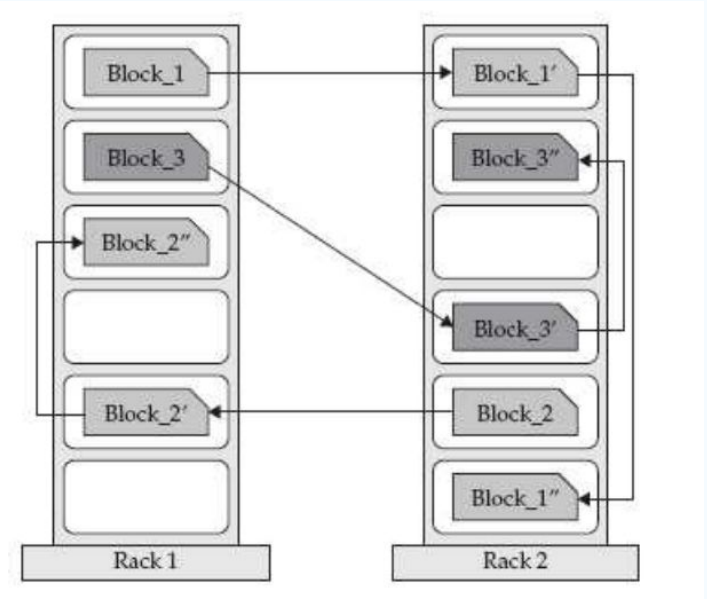


Ecosistema

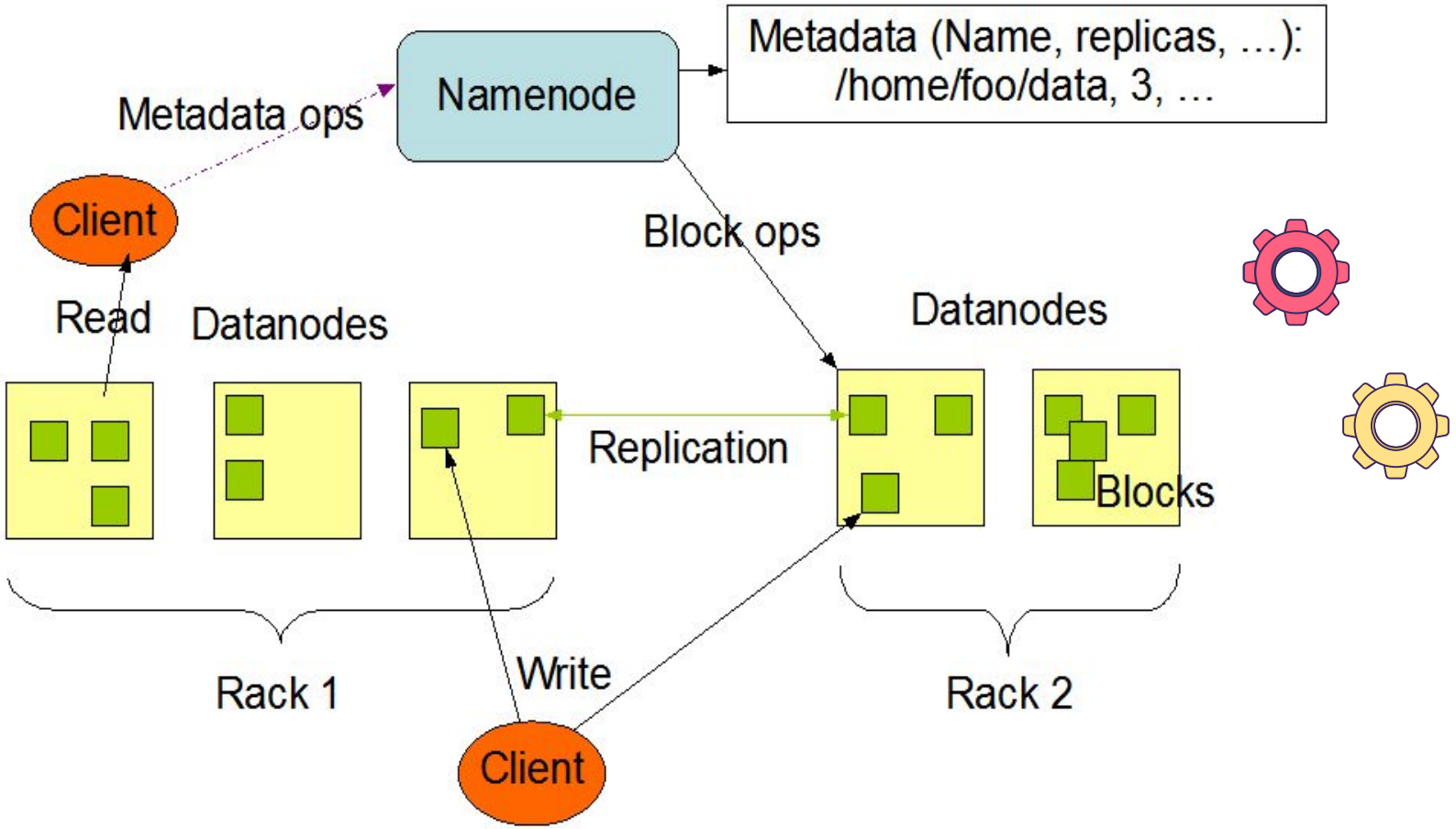

Data Flair
Hadoop Ecosystem
oozie
(Work flow)
HCatalog
Table & schema Management
Pig
(Scripting)
Hive
(Sql Query)
mahout
(Machine Learning)
Drill
(Interactive Analysis)
AVRO
(JSON)
Thrift
(Cross Language Service)
APACHE HBASE
HBASE
(Columnar Store)
Sqoop
(Data Collection)
FLUME
Flume
(Data Collection)
Zookeeper
(Coordination)
Ambari
Apache Ambari
(Management & Monitoring)
Mapreduce
(Data Processing)
Yarn
(Cluster Resource Management)
HDFS
(Hadoop Distributed File system)

¿Qué es HDFS?

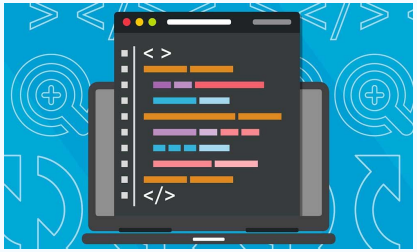
Es el sistema de almacenamiento principal de Hadoop, que se utiliza para almacenar y procesar grandes conjuntos de datos en clústeres de servidores.

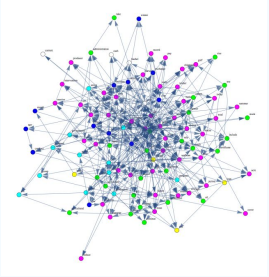


HDFS Architecture

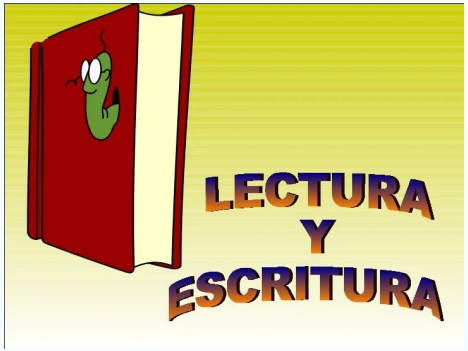


Datos estructurados y no estructurados

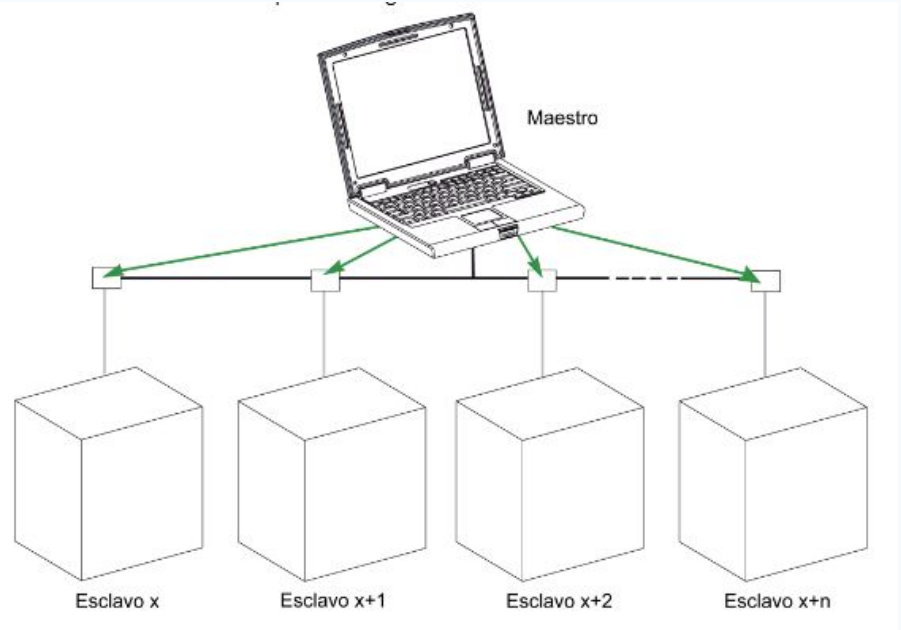




WORM

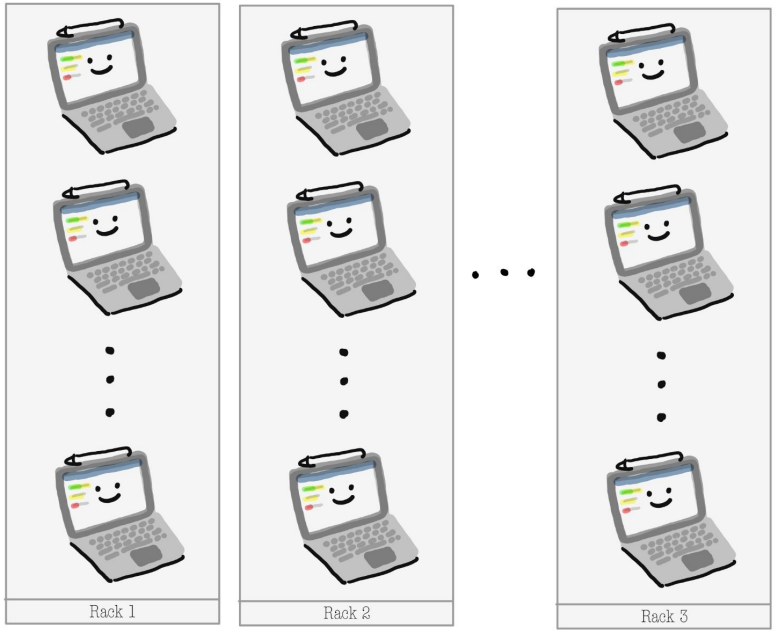


Modelo maestro / esclavo



1 máquina

= 1 nodo

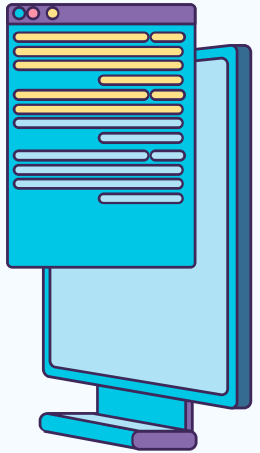


Clúster

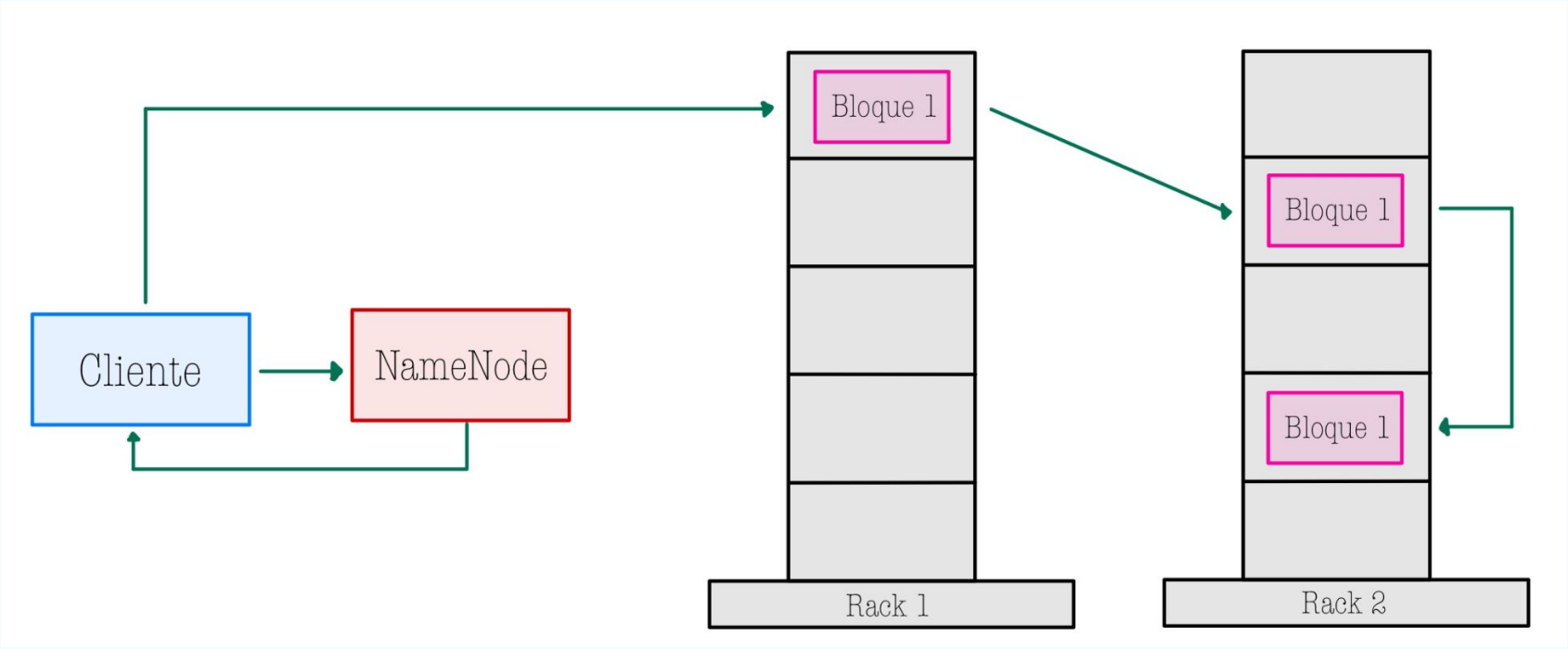
150 MB

64 MB | 64 MB | 22 MB

nodos



Funcionamiento



Caracteristicas

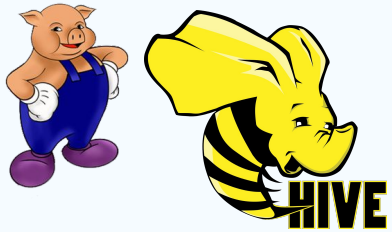
01 Escalable horizontalmente
Permite agregar más servidores a medida que aumenta el tamaño de datos.

02 Tolerancia a fallos
Manejar fallos sin pérdida de datos.

03 Bajo costo
Lo hace accesible en comparación a otras soluciones de almacenamiento.

04 Alto rendimiento
Maneja grandes cantidades de datos con velocidad de lectura y escritura.

Aplicaciones en Big Data



Procesamiento de datos

Apache Hive
Apache Pig

Análisis de datos de streaming

Apache Storm
Apache Spark Streaming

Análisis de datos de redes

Apache HBase
Apache Cassandra

04

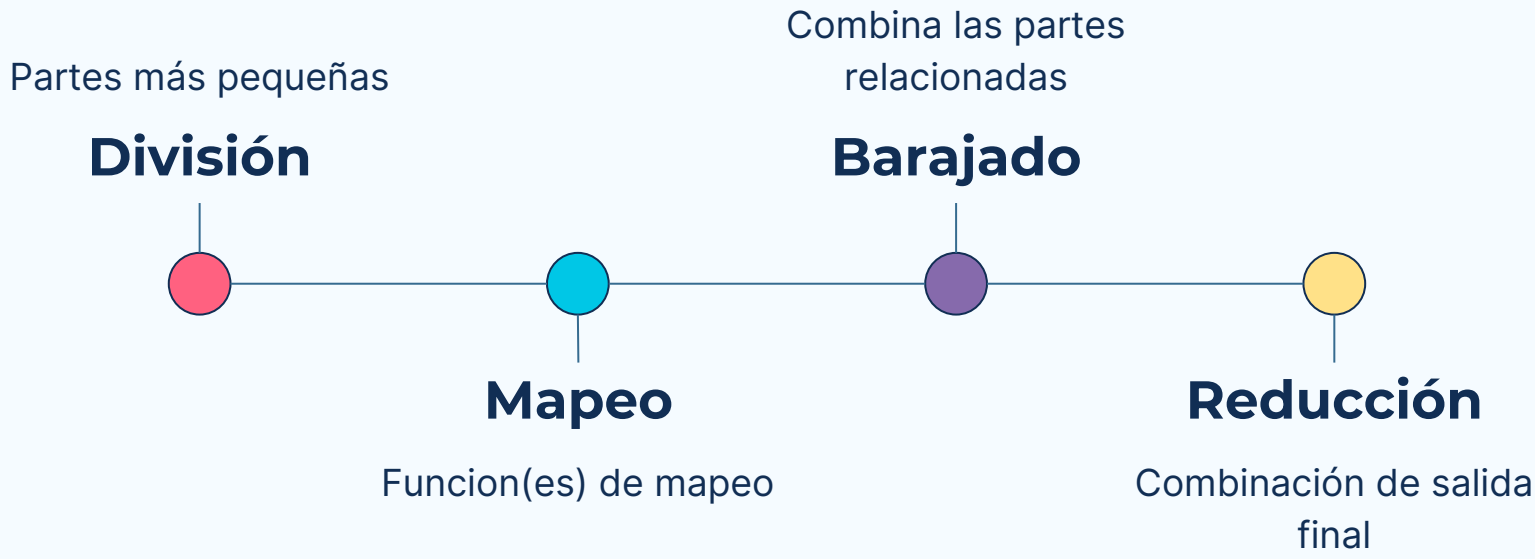
Map Reduce

Framework para escribir aplicaciones que procesan grandes cantidades de datos en paralelo en grandes clústeres de hardware de bajo coste de forma fiable y tolerante a fallos.

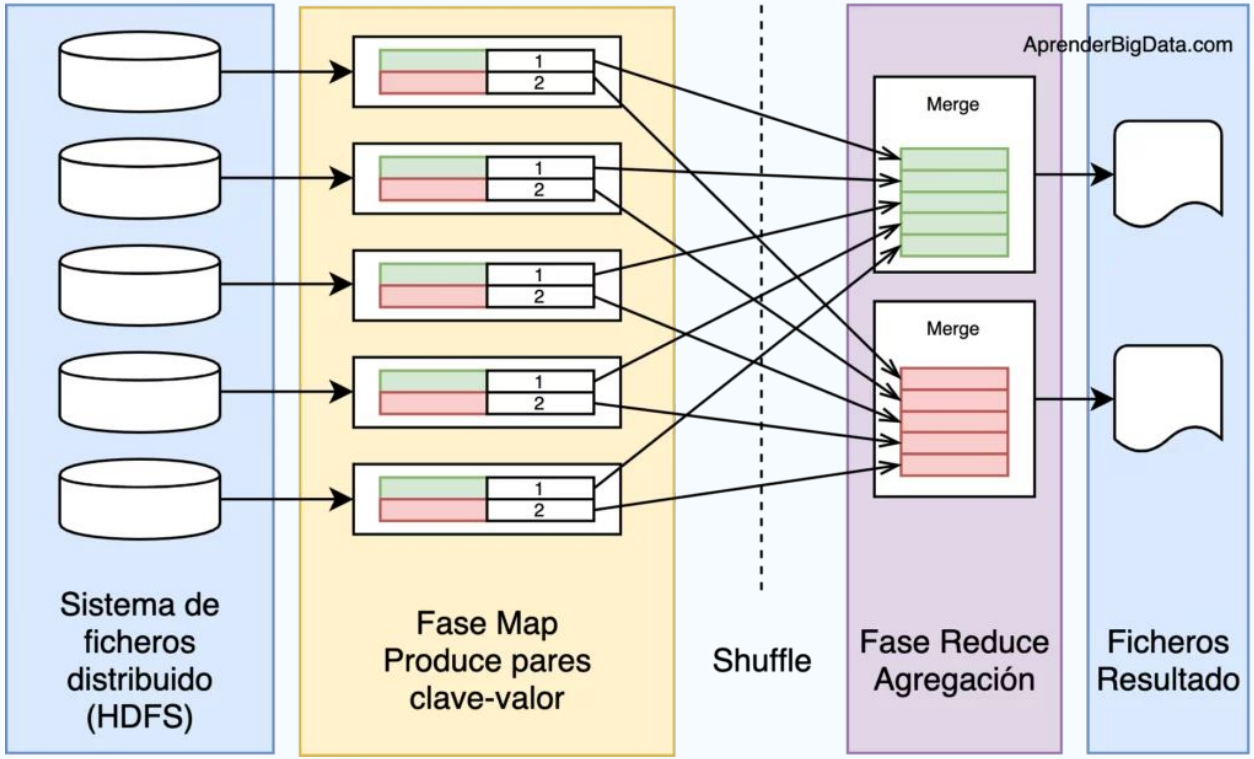
¿Qué es MapReduce?

MapReduce es un modelo de programación que simplifica la tarea de procesamiento de datos al permitir a los usuarios realizar procesamiento paralelo y distribuido en grandes volúmenes de datos.

Etapas en MapReduce



Arquitectura MapReduce



Limitaciones en Hadoop MapReduce

01 Complejidad de programación

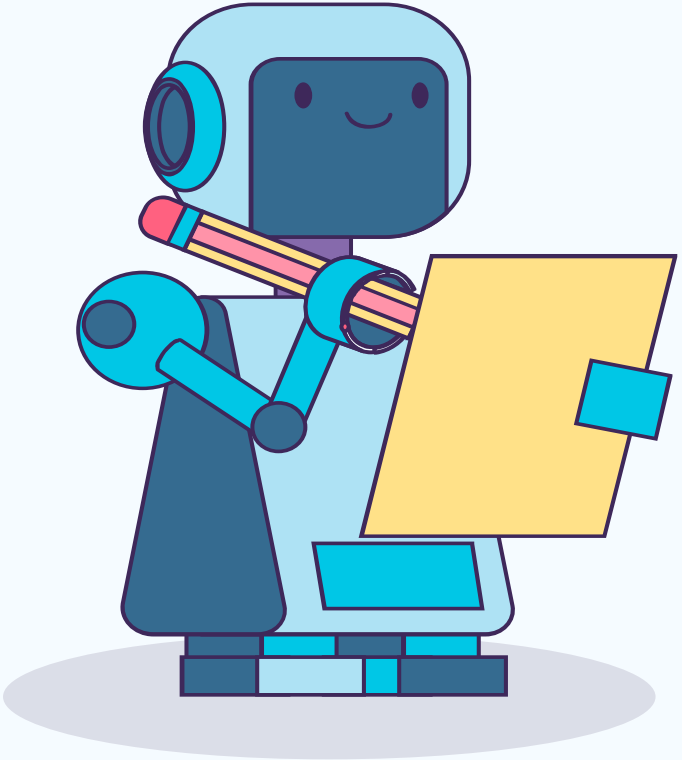
Código java complejo

02 Overhead de E/S

Lectura y escritura siempre en disco duro

03 Falta de tolerancia a fallos en tiempo real

Si una tarea de MapReduce falla, todo el proceso debe reiniciarse desde el principio



Gracias :)

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**

Referencias

- Tan, P.-N., Steinbach, M., & Kumar, V. (2013). *Introduction to Data Mining: Pearson new international edition PDF eBook*. Pearson Higher Ed.