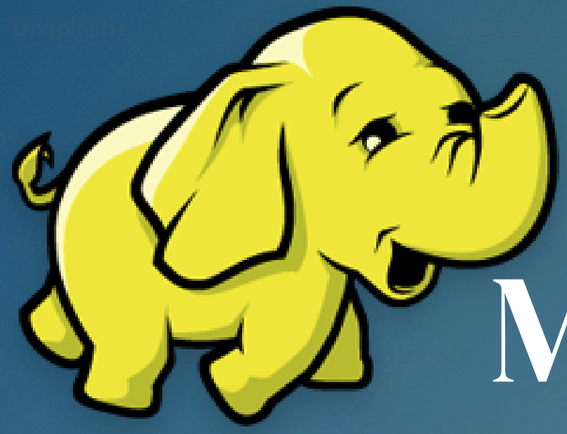


# MapReduce y procesamiento batch

*Profesores:*

**Miguel Angel Sánchez Hernández**
**Omar Mendoza González**

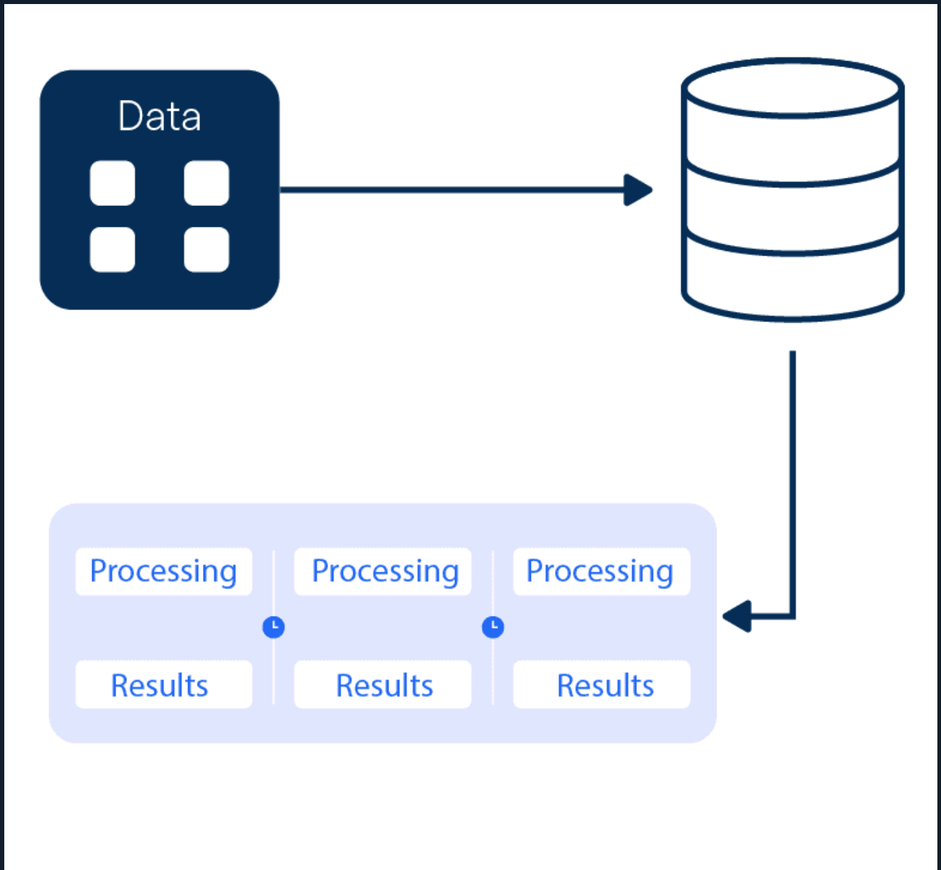
*Alumna:*

**Belem Anahi Mendieta Hernández**

# Indice

03

# INTRODUCCIÓN




## ¿Qué es el procesamiento batch?

Consiste en ejecutar grandes volúmenes de datos en bloques, sin interacción del usuario y generalmente de forma programada.

04

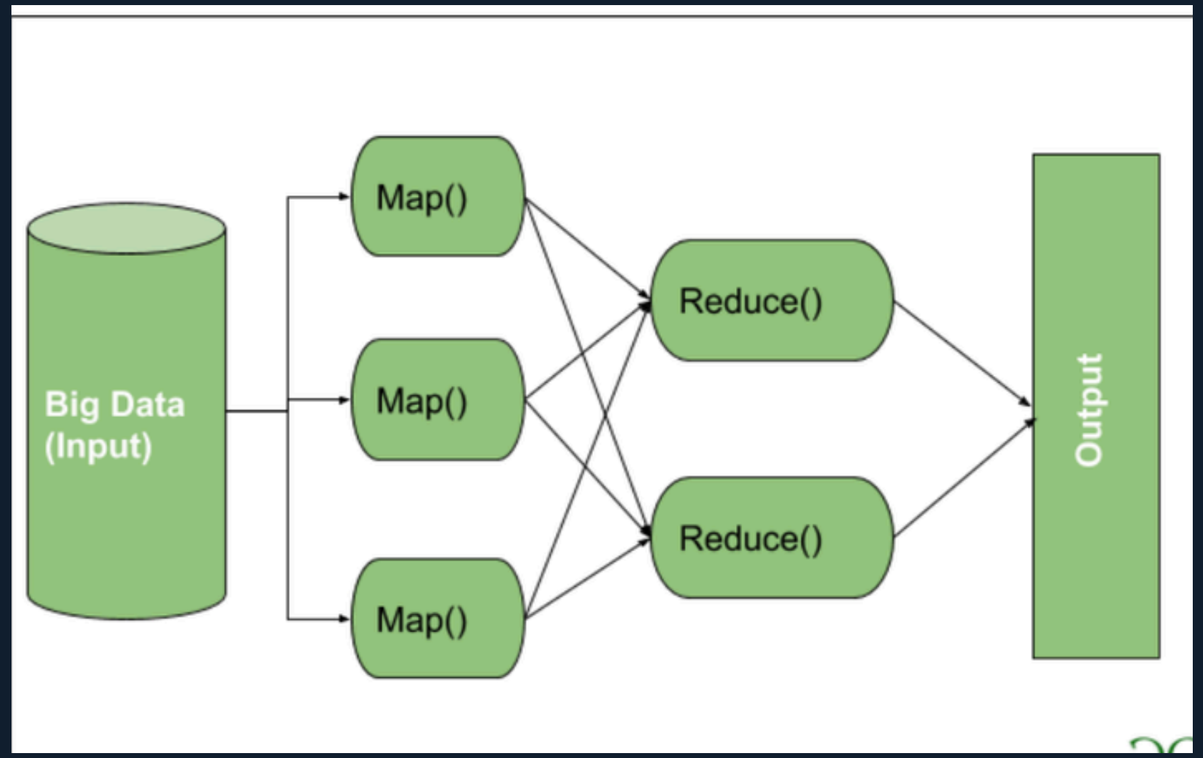
# ¿QUÉ ES?

## MapReduce

Es un modelo de programación creado por Google y adoptado por *Apache Hadoop* que permite procesar grandes volúmenes de datos en paralelo sobre un clúster distribuido.

Divide el trabajo en dos fases:

- **Map**: transforma los datos.
- **Reduce**: los agrupa y combina.



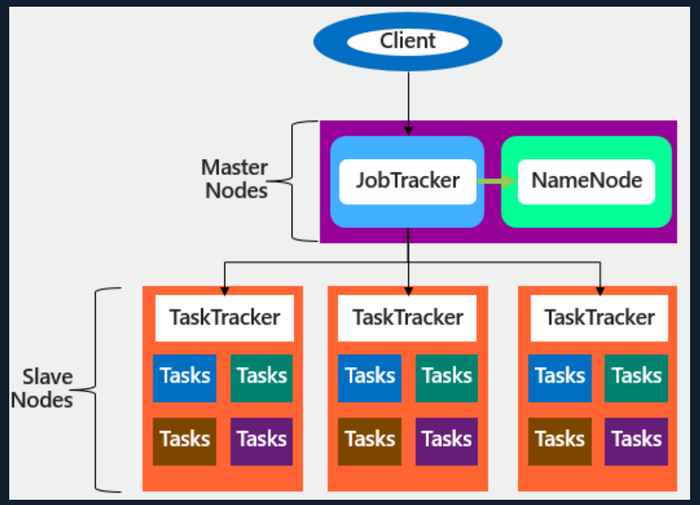
05

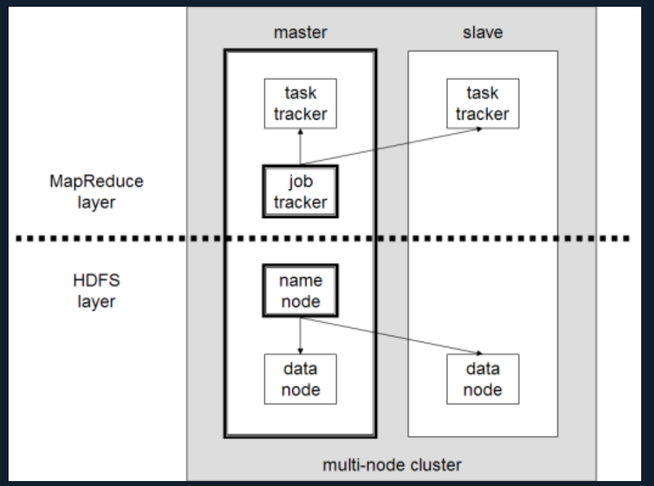
# ARQUITECTURA BÁSICA DE MAPREDUCE

En Hadoop 1.x, el sistema se basa en un *JobTracker (gestiona los trabajos)* y varios *TaskTrackers (ejecutan tareas en nodos del clúster)*. Cada nodo procesa una parte de los datos almacenados en el *HDFS (Hadoop Distributed File System)*.

*Componentes principales:*

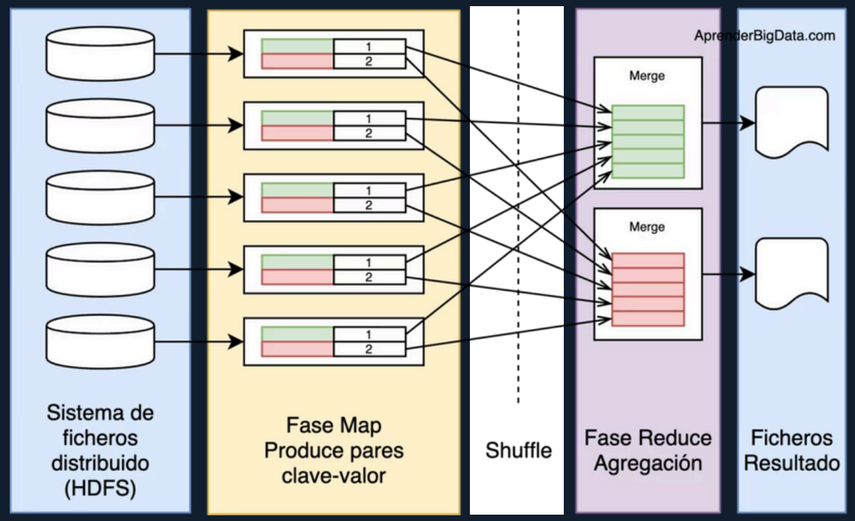
**Client**: Envía.
**JobTracker**: Coordina.
**TaskTracker**: Ejecuta.
**HDFS**: Almacena.



06

# FLUJO DE EJECUCIÓN



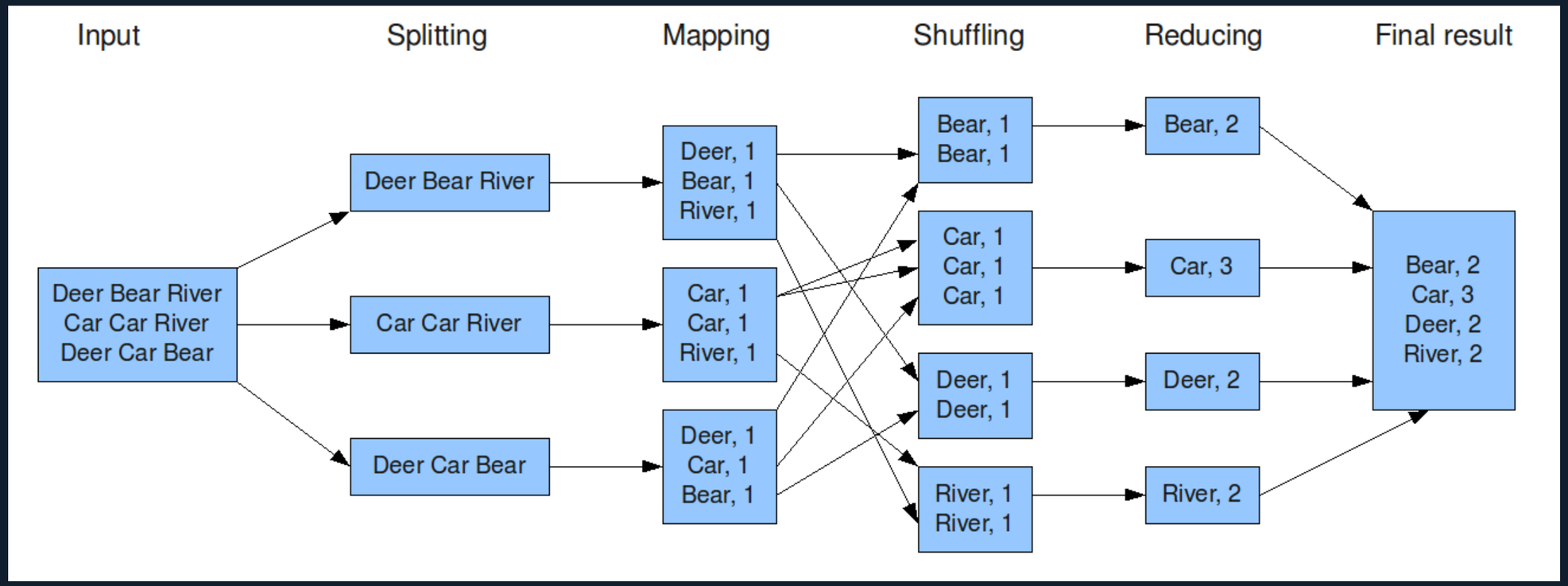

07

# VENTAJAS Y LIMITACIONES

| ***Ventajas*** | ***Limitaciones*** |
|---|---|
| Escalabilidad horizontal | Procesamiento lento |
| Tolerancia a fallos | No apto para tiempo real |
| Paralelismo automático | Complejidad en la programación |
| Integración con HDFS | Dificultad en tareas iterativas |

08

# EJEMPLO PRÁCTICO



09

# CASOS DE USO

**Análisis de logs de servidores web.**

**Procesamiento de grandes volúmenes de texto**

**Análisis de clics o comportamiento de usuarios**

**Generación de reportes de ventas históricos**