

MapReduce y procesamiento batch

Profesores:

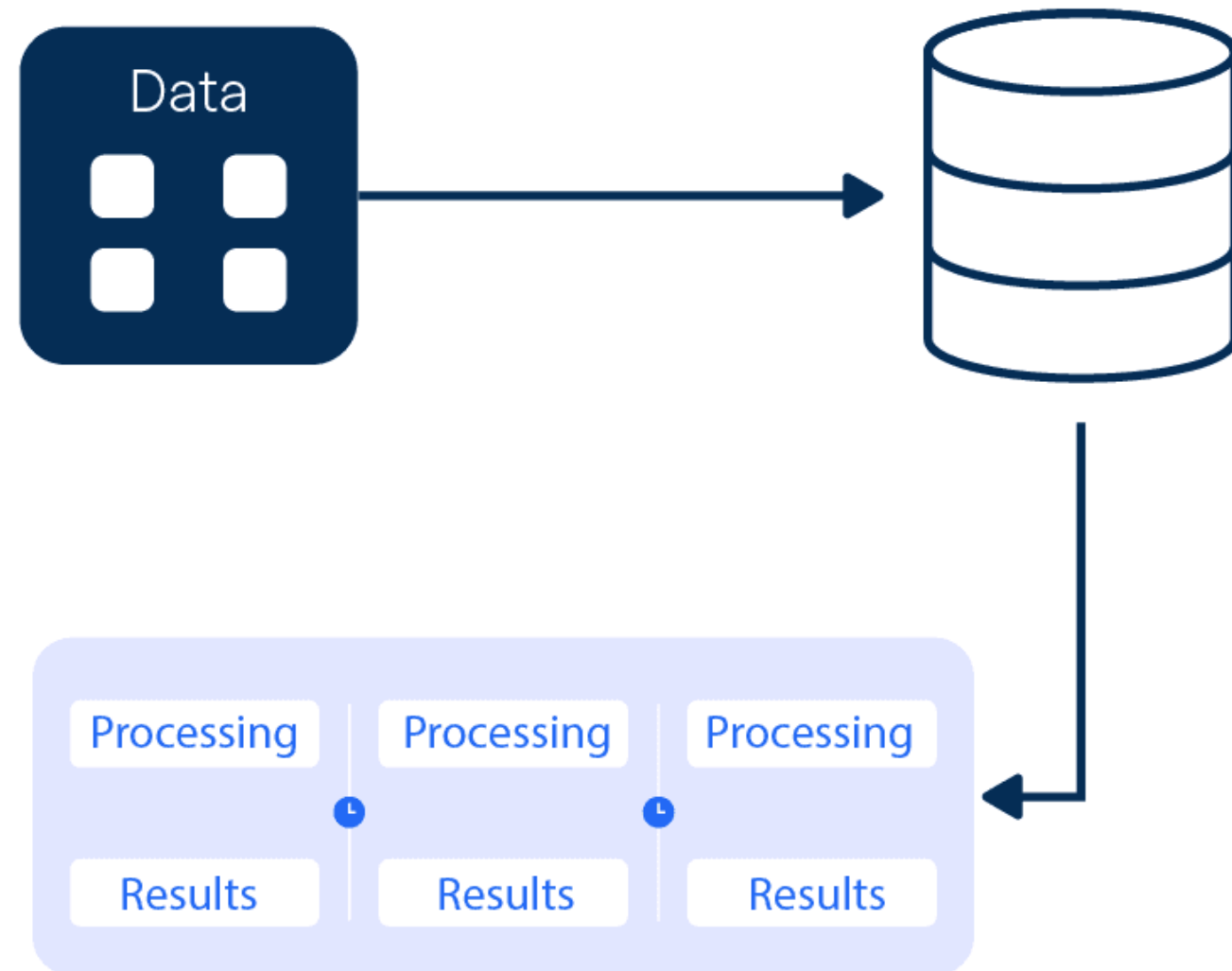
Miguel Angel Sánchez Hernández
Omar Mendoza González

Alumna:

Belem Anahi Mendieta Hernández

Indice

03	Introducción
04	¿Qué es?
05	Arquitectura básica
06	Flujo de ejecución
07	Ventajas y limitaciones
08	Ejemplo práctico
09	Casos de uso



¿Qué es el procesamiento batch?

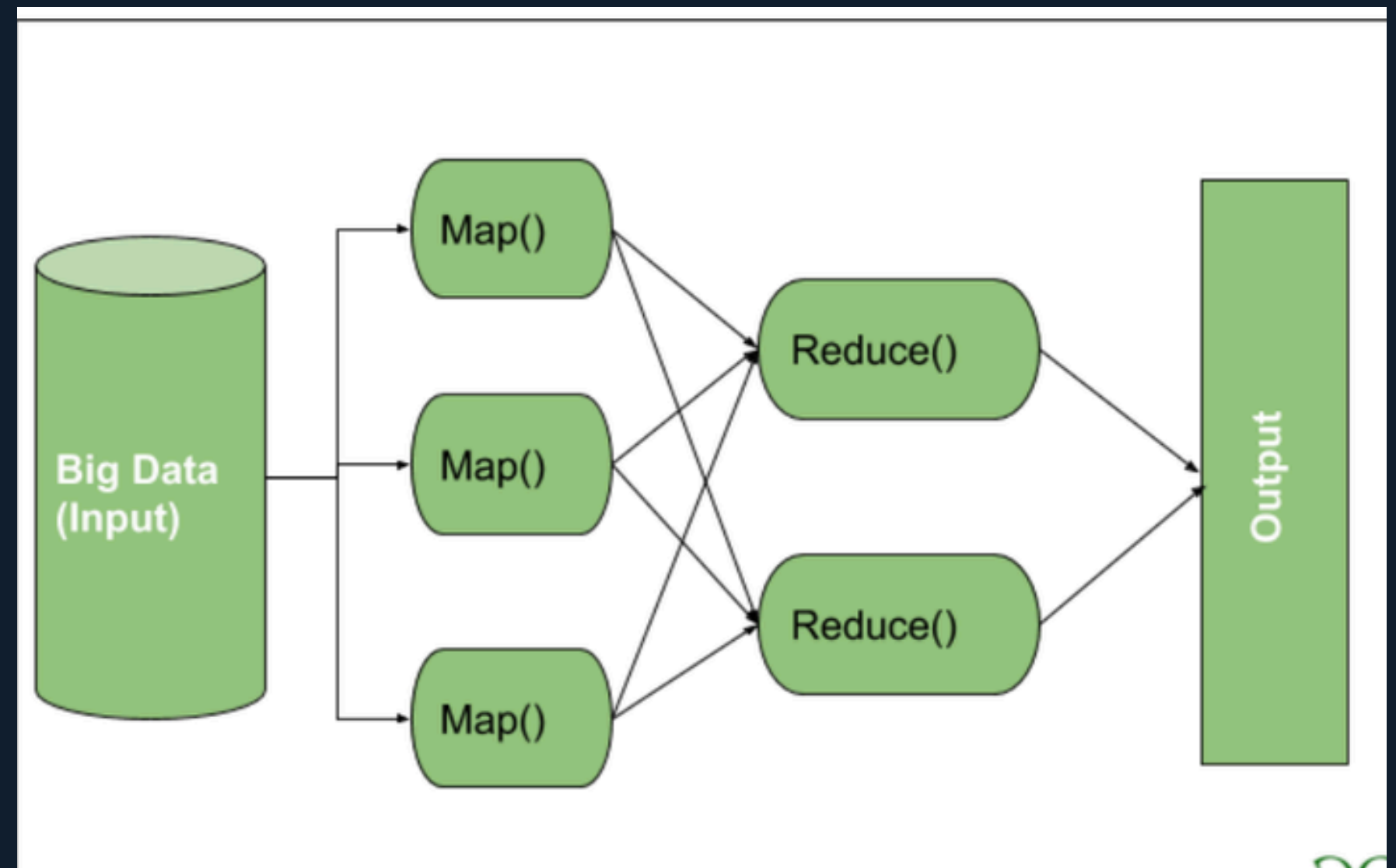
Consiste en ejecutar grandes volúmenes de datos en bloques, sin interacción del usuario y generalmente de forma programada.



MapReduce

Es un modelo de programación creado por Google que permite el procesamiento de datos masivos mediante dos fases principales que operan con pares clave-valor:

- **Map:** Recibe y transforma los datos en pares.
- **Reduce:** Toma los pares los agrupa por clave y los combina para el resultado.



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*).



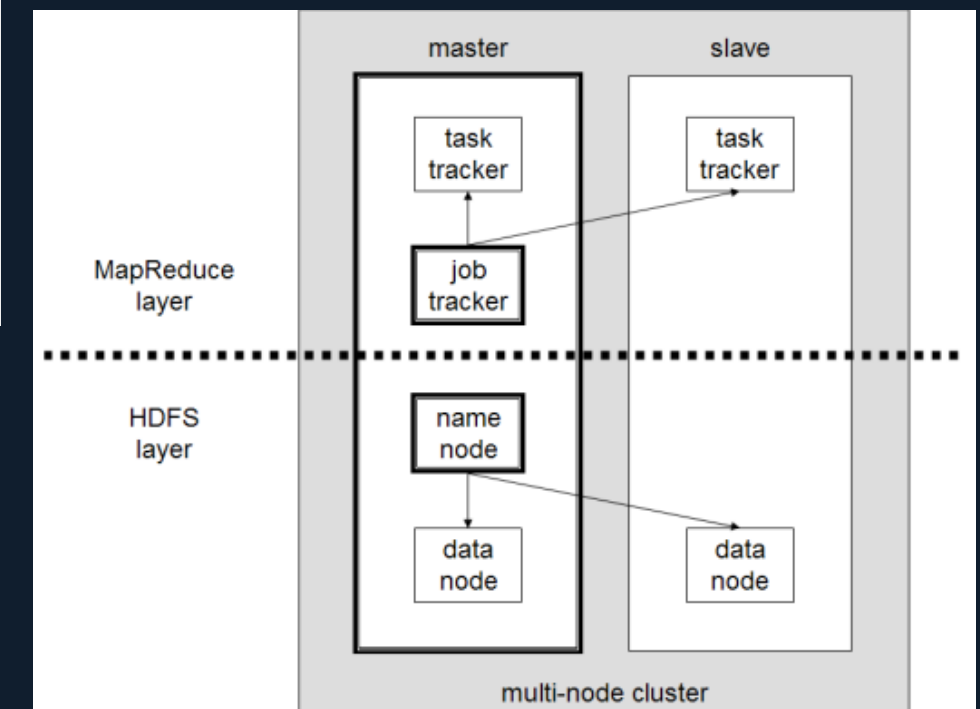
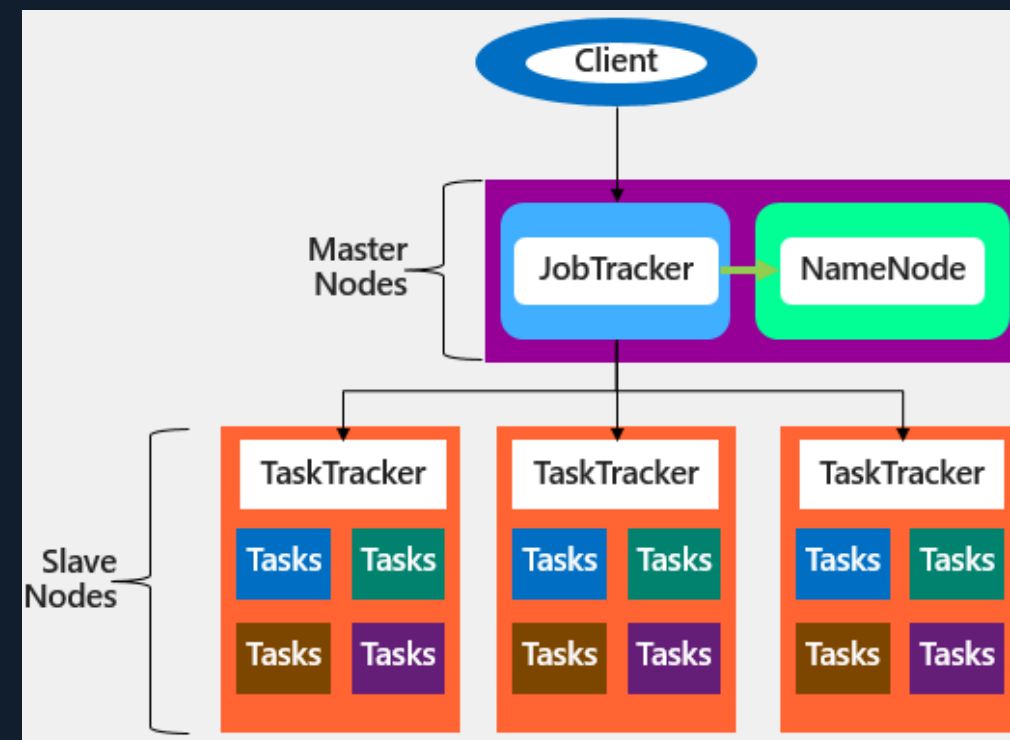
Componentes principales:

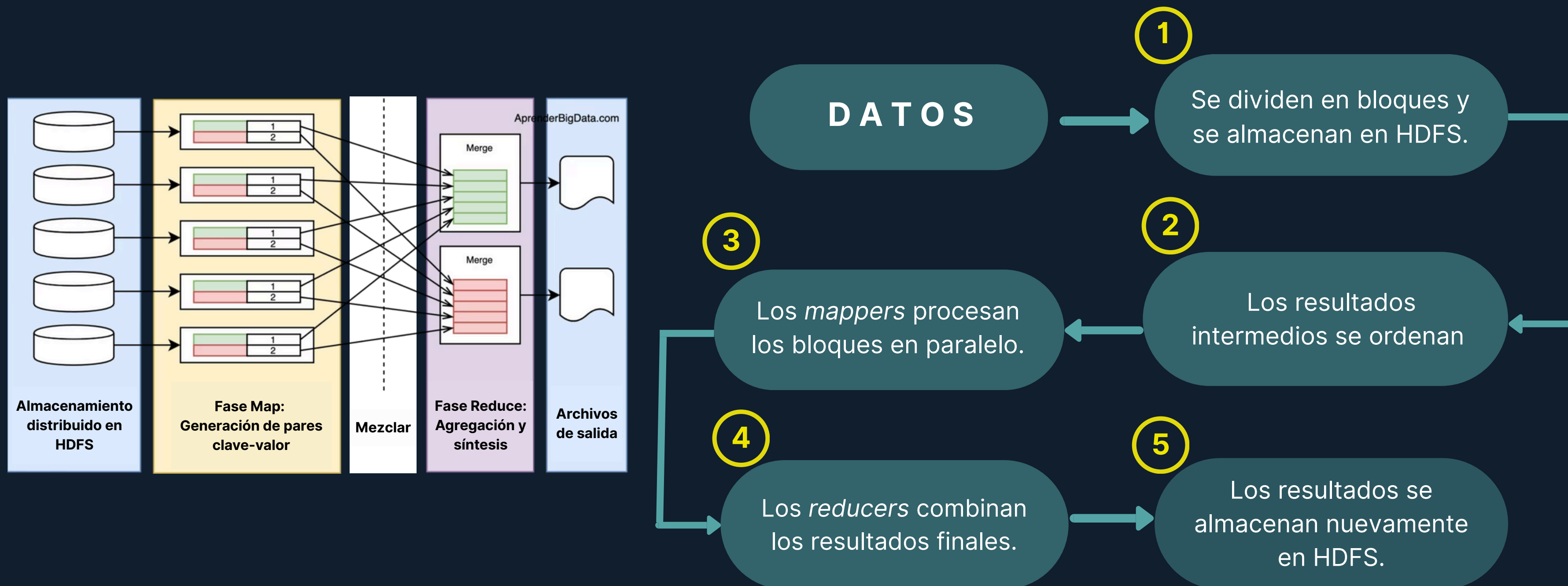
Client: Envía.



JobTracker: Coordina.

TaskTracker: Ejecuta.

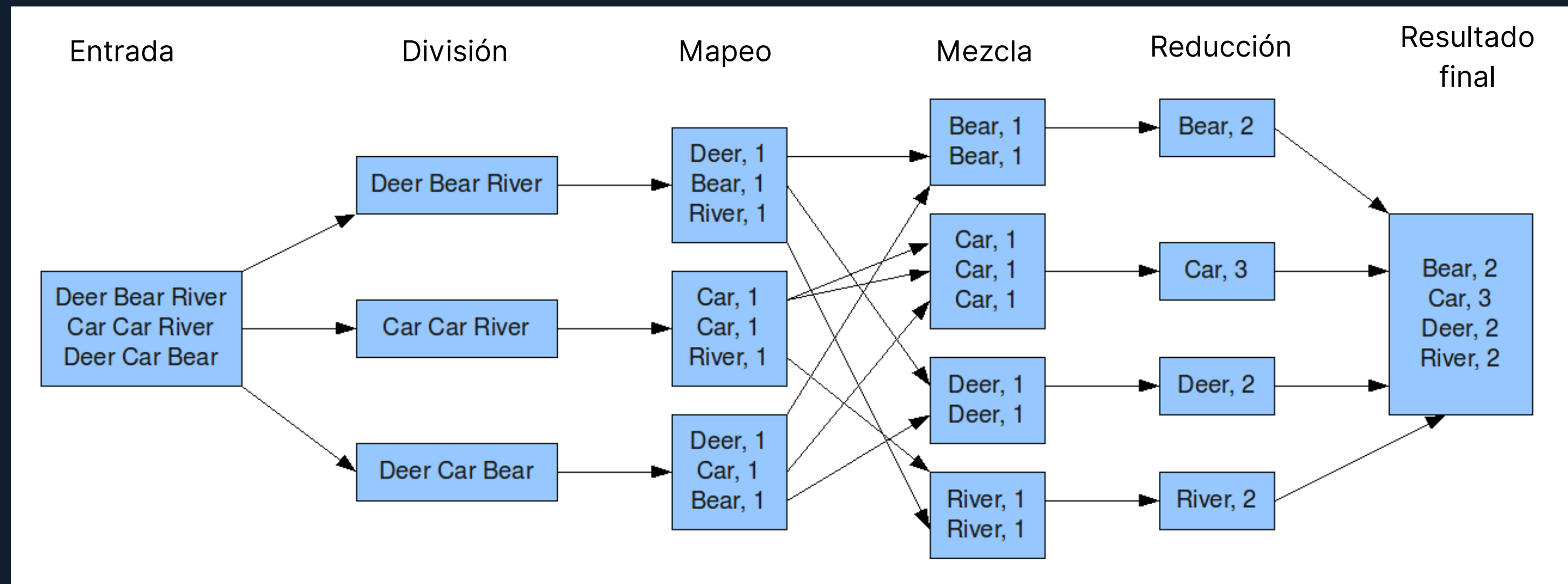
HDFS: Almacena.





 <i>Ventajas</i>	 <i>Limitaciones</i>
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

- Contador de palabras





**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**