Big Data y Visualización
Computación Escalable

Universidad Autónoma de Manizales
Docente: Alejandro Mora Rubio

Contenido

Computación Escalable

Hadoop

Hadoop Distributed File System (HDFS)

MapReduce

Hadoop YARN

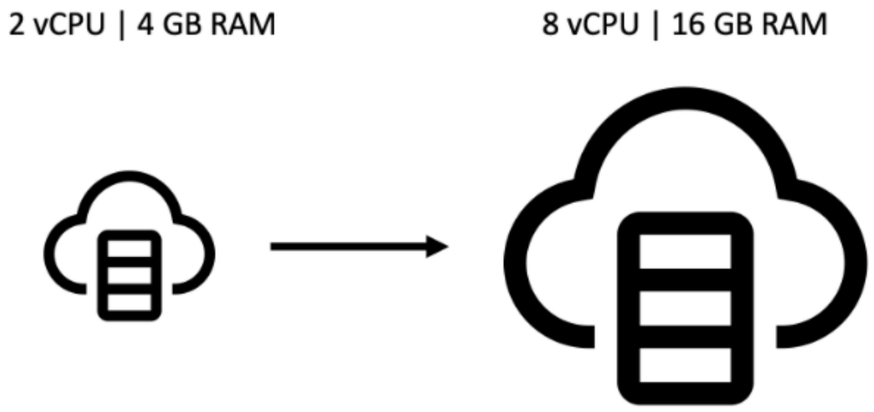
Limitaciones

Computación Escalable

Escalabilidad: capacidad de adaptación de un sistema con respecto al rendimiento del mismo a medida que aumentan de forma significativa el número de usuarios del mismo o la cantidad de datos a almacenar/procesar.

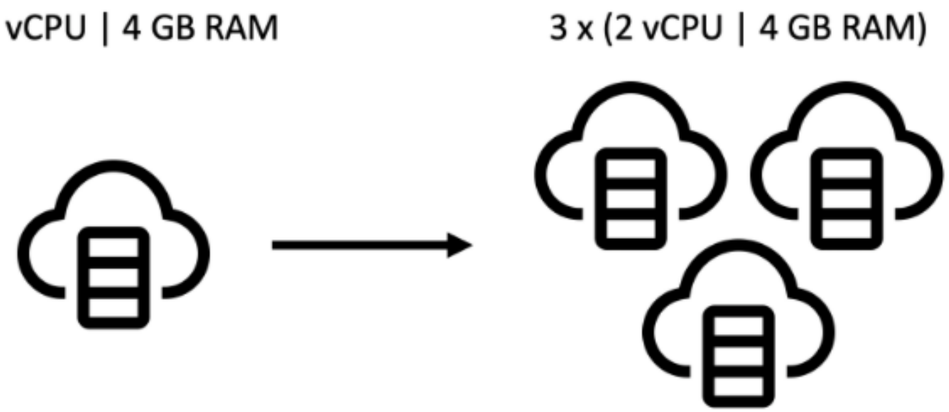
Computación Escalable

Hacia arriba (Scaling Up)



Aumentar la capacidad de un nodo de computación.

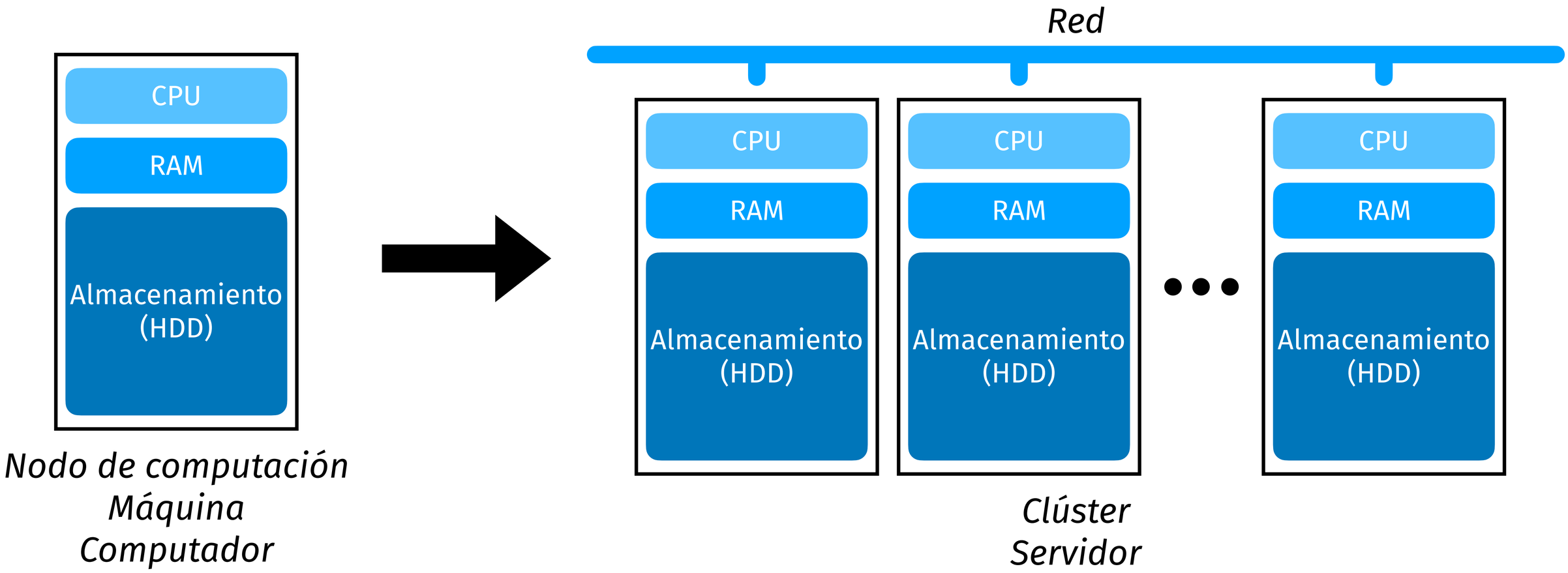
Hacia afuera (Scaling Out)



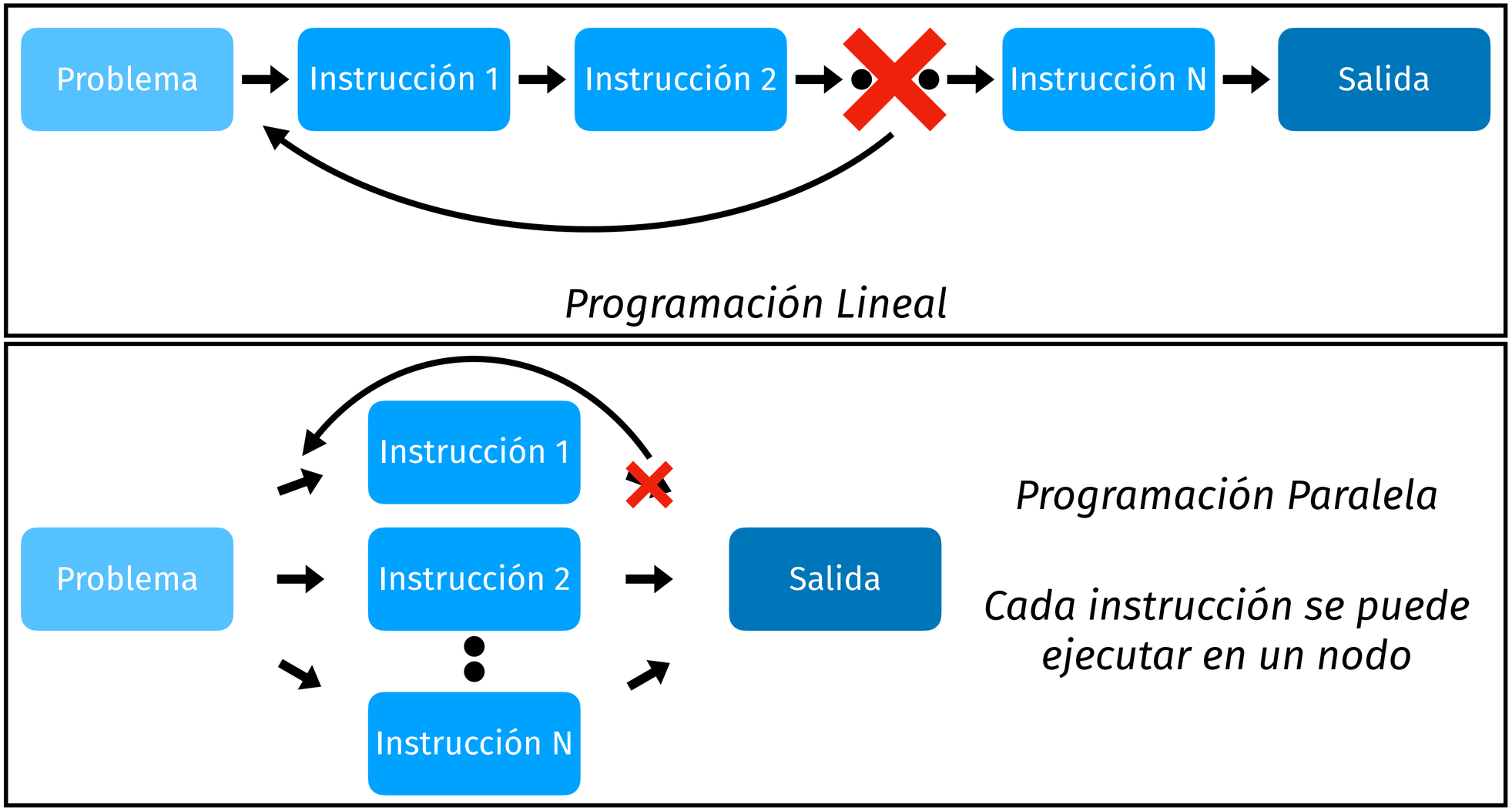
Aumentar el número de nodos de computación (Con las mismas características).

Imágenes de IBM

Computación Escalable



Computación Escalable



Computación Escalable

Sistemas de Archivos Distribuidos (Distributed File Systems, DFS)

Paralelización

Modelos de Programación

Hadoop

Framework *Open Source* usado para procesar grandes conjuntos de datos.

Diseñado para manejar volúmenes de datos > TB.

Permite trabajar con datos estructurados y no estructurados.

Administra trabajos/procesos paralelos.

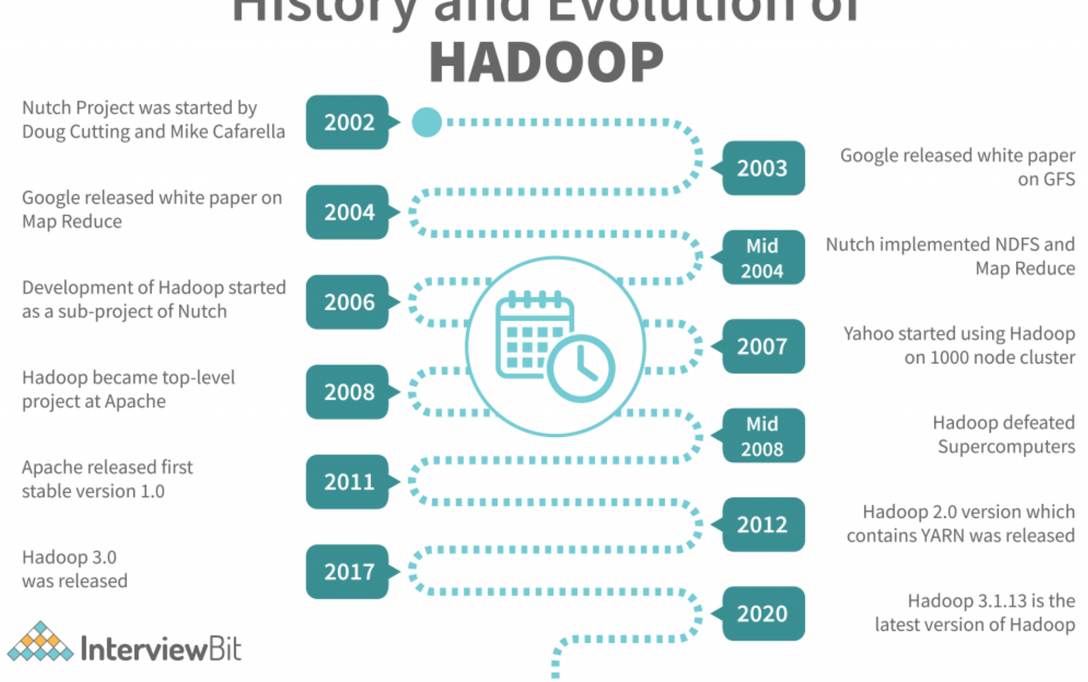


Imagen de InterviewBit

HDFS

Componente de almacenamiento.

Administra y almacena grandes volúmenes de datos en *commodity hardware*.

División de los archivos en ***bloques***, se replican y almacenan en diferentes máquinas.

Permite escalar hasta miles de nodos.

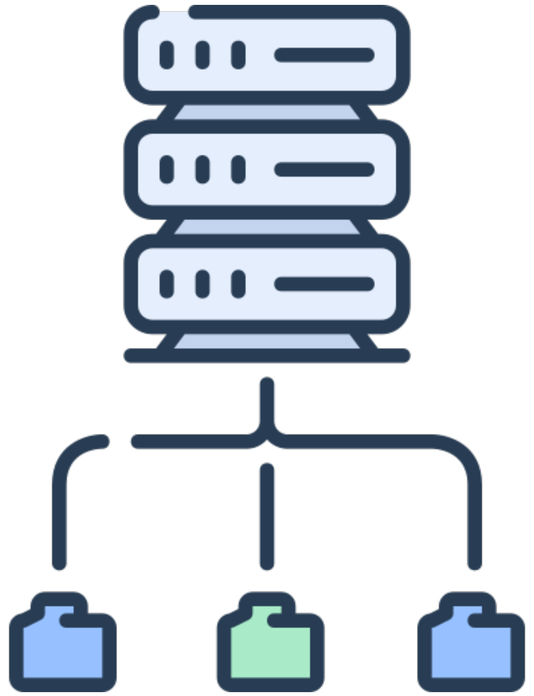


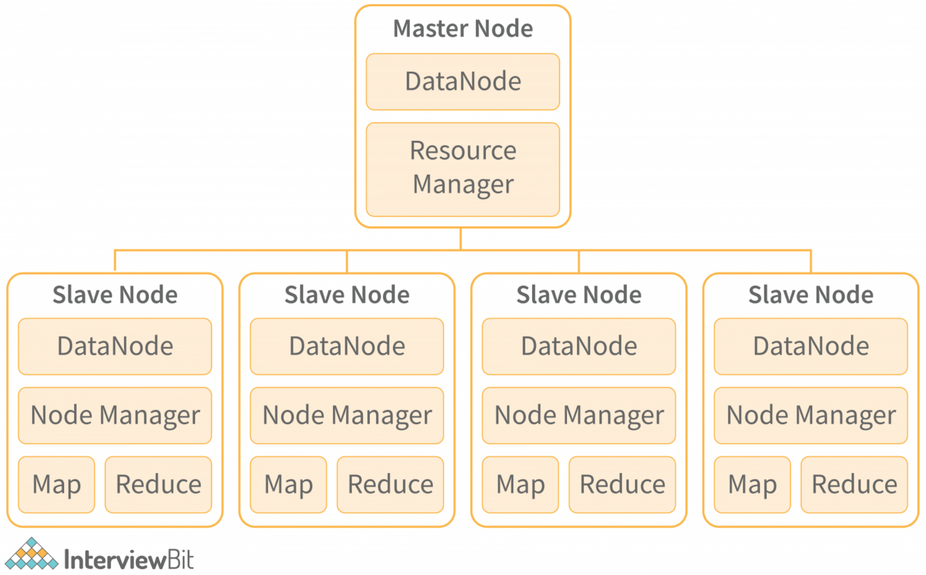
Imagen de Flaticon

HDFS

Nodo primario (NameNode) y nodos secundarios (DataNode).

Bloque mínima cantidad de datos que se pueden leer/escribir.

Los bloques se replican en distintas máquinas.



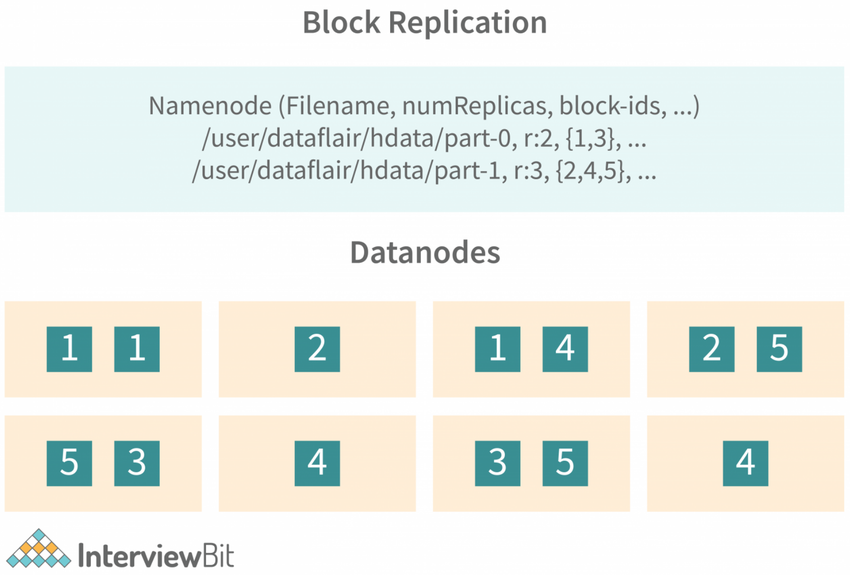


Imagen de InterviewBit

MapReduce

Modelo de programación distribuida basada en dos tareas:

Map operaciones para extraer información en el formato *llave → valor*

Reduce agrega los resultados para cada una de las llaves generando un resultado final

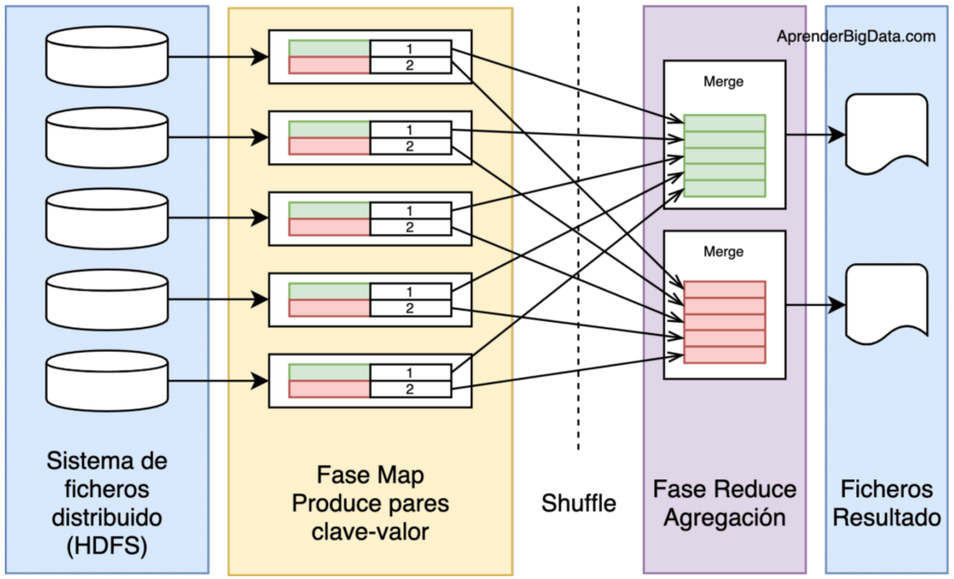


Imagen de Aprender BIG DATA

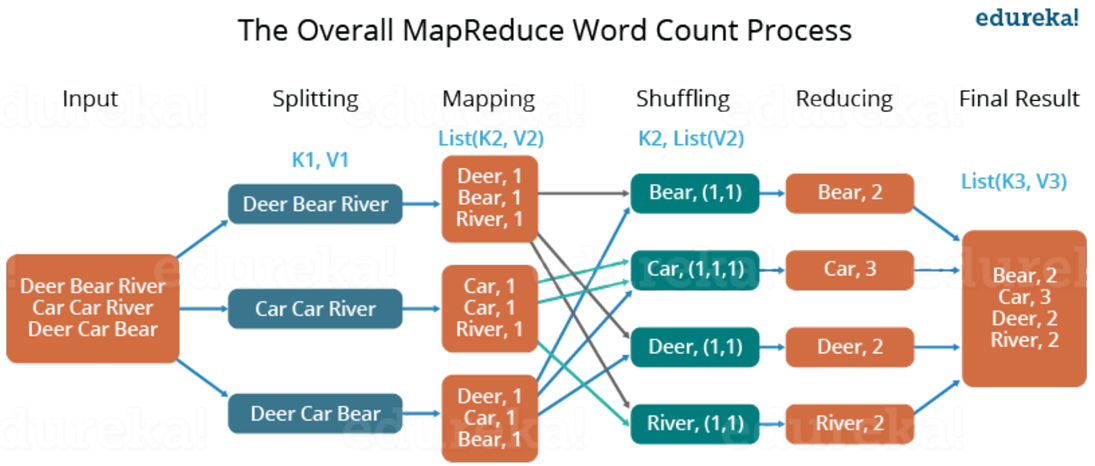


Imagen de Edureka

Hadoop YARN

Acrónimo para *Yet Another Resource Manager*

Administra los recursos disponibles para la ejecución de tareas, incluso utilizando herramientas diferentes a MapReduce.

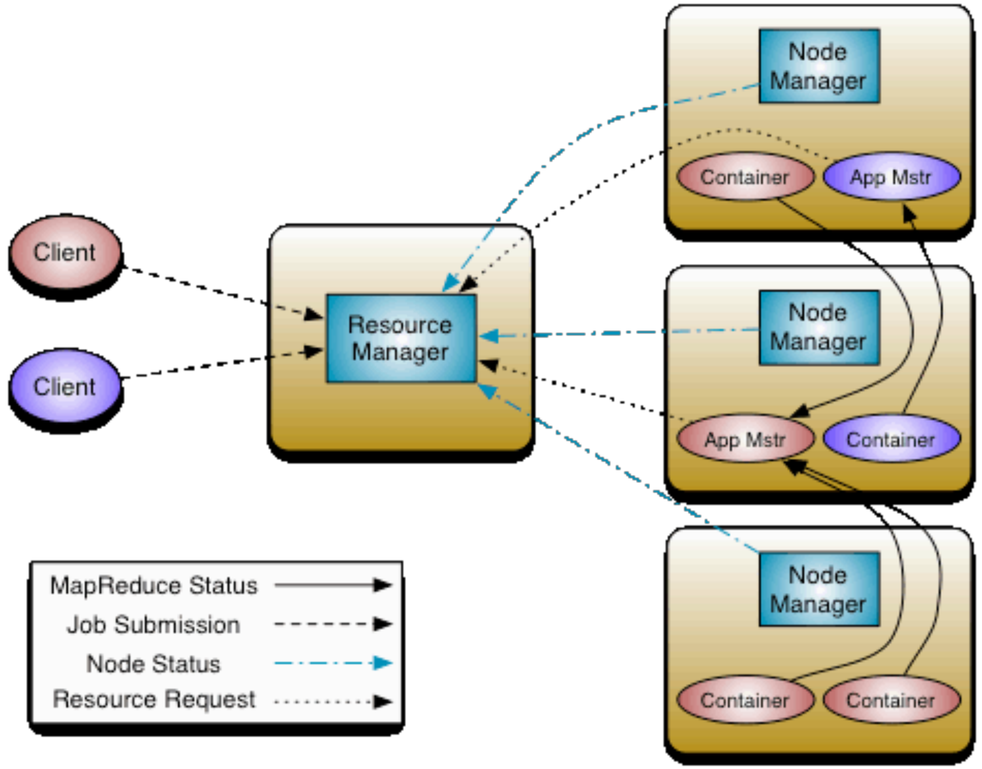


Imagen de Apache

Limitaciones

- Tareas no paralelizables
- Dependencias entre los datos (orden de procesamiento)
- Acceso de baja latencia
- Procesamiento de pequeñas bases de datos.
- Almacenamiento de datos sensibles.

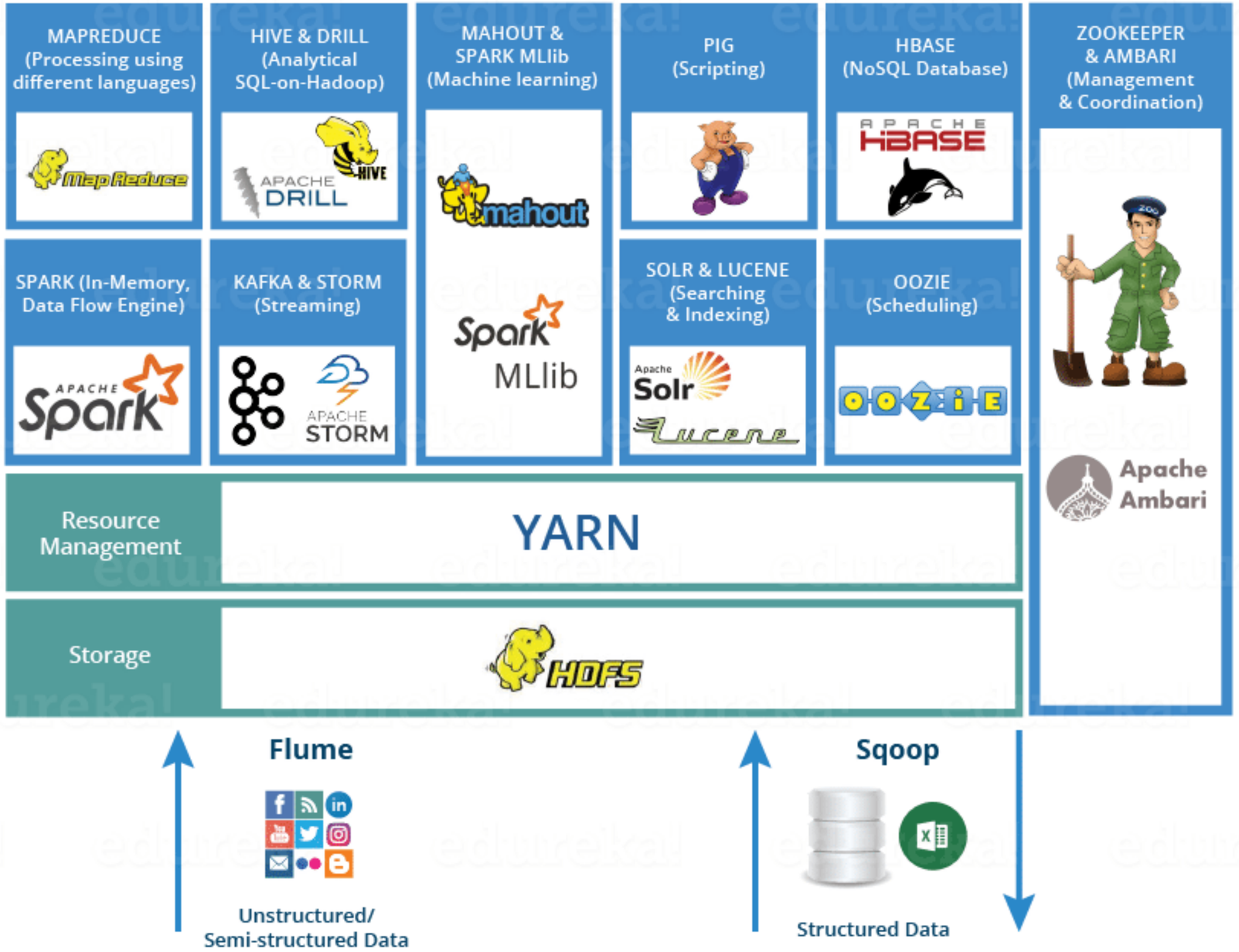


Imagen de EDUREKA

Actividad

Investigación acerca de las regulaciones actuales para la recolección y almacenamiento de datos de los usuarios, comparando la situación en Colombia y otros países.

Por ejemplo, ¿Qué implica la Ley 1581 de 2012? La cual está relacionada con la protección de datos personales en Colombia.

Entregable: reporte escrito (máx. 2 páginas, formato IEEE) donde se describan los hallazgos y se analice la responsabilidad del ingeniero/ desarrollador.

Fecha de entrega: 1 de Marzo.