1. Big data stack

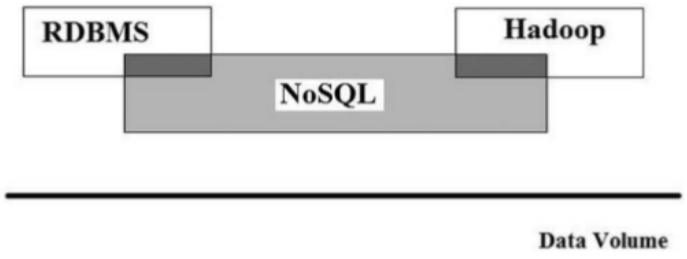
Stack → grupo de herramientas e infraestructuras integradas.
El entorno debe estar formado por:

- Consideraciones de hardware.
- software de infraestructura.
- software operativo.
- software de administración.
- interfaces de programación de apps (API) bien definidas e incluso herramientas de desarrollo de software.

La arquitectura tiene que ser capaz de realizar todos los **requisitos fundamentales:**

- Ingerir
- Integrar
- Organizar y almacenar
- Análisis
- Actuar o entregar o visualizar.
- Asegurar el sistema.

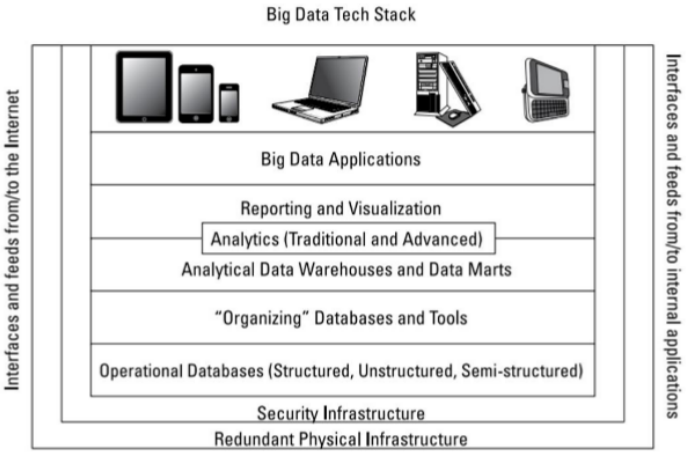
El big data produce algunos problemas de escala específicos para esta infraestructura.
¿Donde se encuentran las bases de datos relacionales, las bases de datos NoSQL y el sistema de Big data Hadoop en la escala de datos?



Se muestra el volumen de datos.

RESUMEN BIG DATA STACK

Capa 0 → infraestructura física; hardware, red…
Capa 1 → requisitos de seguridad y privacidad.
Capa 2 → bases de datos operativos
Capa 3 → organización de los servicios y herramientas de captura de datos. En esta capa se incluyen estas tecnologías.



- Un sistema de archivos distribuido: adapta la descomposición de los flujos de datos y para proporcionar escala y capacidad de almacenamiento.
- Servicios de serialización: convierte un objeto de datos en una serie de bytes que pueden ser fácilmente transmitidos a otro destino. Esto es útil para transferir objetos de datos entre diferentes sistemas o aplicaciones.
- Servicios de coordinación: para crear aplicaciones distribuidas.
- Herramientas de extracción, transformación y carga (ETL): para datos estructurados y no estructurados en Hadoop.
- Servicio de flujo de trabajo: para programar trabajos y proporcionar una estructura para sincronizar elementos de proceso entre capas.

Capa 4 → los almacenes de datos analíticos y data marts.
Data mart: parte de un almacén de datos que se enfoca en una línea de negocio particular. Estos data marts contienen información resumida y específica para un área de la organización.
Análisis de Big Data: los algoritmos tienen que ser capaces de trabajar con muchos datos en tiempo real y con gran variedad. 3 clases de herramientas:

- Informes y paneles
- Visualización: evolución de los cuadros de mando.
- Analítica avanzada: predicciones, recomendaciones, inferencias, tendencias…

Aplicaciones del Big Data:

- horizontales, tratan problemas comunes en todas las industrias.
- verticales, ayudan a resolver problemas específicos.

2. Virtualización

- Proceso de crear una representación virtual o basada en software de algo, como aplicaciones virtuales, servidores, almacenamiento y red.
- Separa los recursos y servicios del entorno de entrega físico subyacente, permite crear muchos sistemas virtuales dentro de un solo sistema físico.
- Proporciona una base para la computación en la nube.
- Al optimizar todos los elementos de su infraestructura se obtiene la eficiencia necesaria para procesar y administrar grandes volúmenes de datos estructurados y no estructurados. Big Data → distribución.

Tiene **3 características que soportan la escalabilidad y eficiencia operativa necesaria para entornos de Big Data:**

- Particionamiento: separación de los recursos disponibles.
- Aislamiento: máquina virtual separada del host, un error no afecta a todo.
- Encapsulación: máquina virtual como un solo archivo.

¿Qué podemos virtualizar?

- Servidores
- Procesador
- Infraestructura de apps .
- Redes

3. Principios de DFS

Cuando los datos superan la capacidad de almacenamiento de una sola máquina, se dividen en varias máquinas. Esto se llama un sistema de archivos distribuidos y es más complejo debido a la programación de red. Hadoop tiene un sistema de archivos distribuidos llamado HDFS.
Un sistema de archivos distribuido (DFS), da la solución para los problemas que provoca una arquitectura de este tipo.

- Primer problema a resolver → **fallo del hardware.**

Se pueden evitar la pérdida de datos mediante la replicación (copia de los datos). Permite también una alta concurrencia (acceso a la vez de muchos usuarios) y estas dos cosas provocan la falta de consistencia.

- Segundo problema → **se necesitan combinar los datos de alguna forma en las tareas de análisis.**

¿Qué es un paralelismo de datos?
Dividir los datos entre varios procesadores en sistemas de computación paralela.
Se enfoca en distribuir los datos a través de diferentes nodos de computación para que cada procesador pueda realizar la misma tarea en diferentes partes de los datos.
El paralelismo de tareas se enfoca en dividir las operaciones que se deben realizar.

4. Entorno hadoop

Hadoop

- Plataforma confiable y escalable para almacenamiento y análisis. Se ejecuta en hardware básico.
- Diseñado para procesar grandes cantidades de datos estructurados y no estructurados.
- Es autoreparable → los servidores se pueden agregar o quitar del clúster dinámicamente.
- Se desarrolló porque representaba la forma más práctica para que las empresas puedan administrar grandes cantidades de datos mediante el paralelismo de datos. Esto facilita el manejo de grandes volúmenes de información.

2 componentes de hadoop:

- Hadoop distributed file system: clúster de almacenamiento de datos de bajo coste que facilita la administración de archivos relacionados en todas las máquinas.
- Motor mapreduce: implementación de procesamiento de datos paralelo / distribuido de alto rendimiento.

5. HDFS

- Es una parte de Hadoop que se encarga de almacenar los datos en múltiples computadoras.
- Datos más seguros y fáciles de acceder.
- Se dividen los archivos grandes en piezas más pequeñas llamadas bloques y los almacena en diferentes nodos de datos.
- NameNode es el encargado de controlar todo el acceso a los archivos y asegurarse de que los datos estén en su lugar correcto. Los metadatos son información sobre los datos, como el nombre y la ubicación del archivo. HDFS los utiliza para mantener un registro de todos los archivos y directorios almacenados en el clúster.

6. MAP REDUCE

Hadoop MapReduce es un algoritmo implementado en el proyecto Apache Hadoop que funciona como un motor. Convierte la entrada en salida de forma eficiente y rápida, lo que permite obtener las respuestas necesarias. Es como un motor que necesita combustible para funcionar.
El algoritmo funciona de la siguiente manera:

- Se sube un archivo y se divide en pedazos para procesarlo.
- Se asignan los datos: asignar pares de valores clave a los elementos de las piezas.
- Corto y aleatorio: organiza las piezas haciendo posible al mismo tiempo:
  Equilibrar el número de piezas en cada ordenador.
  Piezas homogéneas en cada ordenador.
- Reducir: realizar la tarea.

7. Plataforma de BIG DATA: arquitectura.

Tiene que consumir incontables fuentes de datos de una manera rápida y económica.

Tiene que tener las siguientes capas.

- Fuentes de datos
- Capa de ingestión
- Capa de visualización
- Capa de administración de la plataforma Hadoop
- Capa de almacenamiento de Hadoop
- Capa de infraestructura de Hadoop
- Capa de seguridad
- Capa de monitoreo