

1. Big data y Almacenamiento Distribuido

1.1. Big Data

Big Data datos que no pueden ser procesados por métodos tradicionales. Se caracterizan por las 5 *V*'s:

- Volumen: cantidad de datos.
- Velocidad: velocidad a la que se generan los datos.
- Variedad: variedad de datos.
- Veracidad: calidad de los datos.
- Valencia: valor de los datos.

1.1.1. Volumen

El impedimento de las computadoras a manejar grandes cantidades de datos. Cluster: un conjunto de computadoras que trabajan en conjunto y pueden ser vistas como un sistema único.

1.1.2. Velocidad

La velocidad con la cual se generan los mismos. Para procesar los datos se necesita que el tiempo de procesamiento sea menor al tiempo de generación de los datos, sino estos se acumulan y el algoritmo puede descartar un gran volumen de datos.

2. VIDEO

Definición 2.1. (Map-Reduce) Procesamiento distribuido de datos utilizando un cluster.

- Modelo de programación para procesar grandes volúmenes de datos.
- Surge de la necesidad de procesar grandes volúmenes de datos de forma escalable.
- El usuario especifica una función **map** que procesa un par clave/valor para generar un conjunto de pares clave/valor intermedios.
- Se debe especificar una función **reduce** que combina todos los valores asociados a una misma clave intermedia.

Definición 2.2. (Map)

- Transforma nuestros datos.

- Debe ser aplicada a cada dato de nuestro ser.
- Puede ser paralelizada y distribuirse entre las distintas máquinas de un cluster.

Algunas diferencias dependientes de la implementación:

- **Hadoop:** $Map(k, v) \rightarrow list(k2, v2)$
- **spark:** $Map(r) \rightarrow list(r')$

Definición 2.3. (Reduce)

- Combina los resultados del map.
- Es necesario procesar los datos de todas las máquinas del cluster.
- Reduce locales en paralelo y reduce entre máquinas mediante esta de shuffle & sort.

Algunas diferencias dependientes de la implementación:

Hadoop: $ReduceByKey((k, v), f) \rightarrow list(k, v)$

- El sistema agrupa todos los registros para los cuales la clave es la misma.
- Requiere que todos los registros de igual clave estén en la misma máquina que ejecute el reduce: Shuffle & Sort.

Spark:

- La función reduce toma dos valores para dar como resultado la combinación de ambos.
- El resultado de un reduce entre dos registros es un input del siguiente reduce.
- Operaciones **conmutativas** y **asociativas** de modo de poder ejecutarse distribuidas.

Definición 2.4. (Cluster) Conjunto de computadoras que trabajan juntas y pueden ser vistas como un sistema único.

Definición 2.5. (Almacenamiento distribuido)

- FileSystem Distribuido (DFS)
- Encargado de gestionar cómo y dónde guardar la información en una computadora, y cómo poder consultarla.
- Almacenar grandes volúmenes de datos en múltiples equipos.
- Replicación de datos para tolerancia a fallos.
- Tolerancia a fallos.

- Alta disponibilidad (seguir funcionando aunque un nodo falle).
- Relativo a bajo costo.

Ejemplos de SD:

- GDS (Google File System)
- HDFS (Hadoop Distributed File System)
- CEPH (Ceph File System)
- S3 (Amazon Simple Storage Service)