**Hadoop Distributed File System (HDFS)**

Es desarrollado por Apache Software Foundation en el 2006, está basado en GFS. Es un sistema de archivos distribuido ya que los datos se encuentran repartidos en los diferentes nodos (datanode los cuales se encuentran en diferentes maquinas) no centralizados en un mismo equipo o servidor. Los archivos son divididos en bloques de datos los cuales tienen un tamaño de 64MB ó 128 MB los cuales por lo general son copiados tres veces y distribuidos en diversos nodos los cuales no son almacenados en una misma maquina

Sus principales características son que permiten el manejo y almacena miento de grandes cantidades de información, permitiendo tasa muy grande de transferencia de datos. Su implementación no requiere de hardware especializado pues se pueden utilizar maquinas ordinarias. Es resiliente a las fallas pues toda la información se respalda por lómenos tres veces en diferentes equipos. También es fácil de escalar pues solo se requiere agregar nuevos nodos.

HDFS es bueno para:

Archivos muy grandes, Acceso a datos a través de streaming, implementarse en Hardware accesible

HDFS no es bueno para:

Low-latency data access, Archivos pequeños, Para esquemas donde se requiera que haya varias personas escribiendo o modificando los archivos

HDFS funciona a través de nodos; Datanode y NameNode.

NameNode: almacena todos los metadatos necesarios para construir el sistema de ficheros a partir de los datos que almacenan los datanodes, puede componer cada fichero a partir de sus bloques. La localización de los bloques en el clúster la almacena en memoria RAM, a partir de la información que le proporcionan los datanodes al arrancar el sistema de archivos.

Datanode: Son los equipos donde se almacenar los bloques que componen cada fichero, y los proporcionan a los namenode o a los clientes que lo solicitan.

Su modelo de coerencia consiste en que Para garantizar la tolerancia a fallos, si un datanode deja de estar disponible, el namenode lo detecta mediante un proceso de heartbeat y vuelve a replicar los bloques perdidos en otras máquinas que sí estén disponibles. Al ser el namenode fundamental para el sistema de archivos HDFS se suele implementar un segundo namenode que toma el control en caso de que el namenode primario deje de responder.

También el namenode se encarga de distribuir los bloques entre los diferentes datanodes procurando que estos estén balanceados, es decir, que no haya diferencias sustanciales en la cantidad de datos que almacena cada uno.

Una de sus aplicaciones actuales es en Apache Hadoop que es un “framework” que permite procesamiento distribuido de grandes conjuntos de datos, a través de “clusters”.