**Hadoop cluster**

ES un software de código abierto(esta basado en java )su motor de procesamiento es en paralelo. Su nombre viene de un juguete del hijo de uno de sus creadores.

Su utilidad es el de almacenar, procesar y analizar una gran cantidad de datos y tareas (hasta petabytes)de cualquier tipo de archivo( estructurados, no estructurados, semiestructurados; archivos de registro, imágenes, video, audio, comunicación) por esto es muy utilizado en la rama big data. Esto hace que sea también empleado como buscador web.

La historia del proyecto empieza cuando se comenzó a desarrollar bajo el nombre de Nutch en el 2002 sus desarolladores fueron Doug cutting(luego de se fue a Yahoo llevándose consigo el proyecto) y Mike cafarella.

En el 2008 Yahoo presento Hadoop como un proyecto de código abierto.Luego paso a la gestión de Apache Software Foundation (fundación sin fines de lucro) que sigue hasta el dia de hoy.

**Arquitectura**

Se basa en una estructura maestro esclavo con un nodo maestro y varios nodos esclavos.

El nodo maestro: procesa la información y se encarga de almacenar el metadato asociado a los esclavos dentro del rack(combinación de nodos datos , máximo 40 nodos master).Este nodo es el responsable de mantener el estado de los nodos slave, convirtiendo a uno de esos nodos en un nodo pasivo(esto sucede ya que si se bloque el nodo maste el nodo pasivo se transforma en master)

El nodo slave o esclavo almacena la información que esta procesando por el nodo master.

Proceso cliente:es un proceso que se lanza a petición de un nodo máster, ya sea para almacenamiento de archivo o recuperación de un archivo en el clúster Hadoop. El nodo máster se comunica directamente con el proceso cliente y actúa según el tipo de petición que este le realiza.

HDFS

Es el sistema de ficheros que almacena los datos en Hadoop., HDFS “rompe” estos ficheros en bloques (de tamaño configurable, comúnmente de 128MB o 256 MB), y luego los distribuye entre los distintos Data Nodes que conforman el cluster HDFS.

El servidor HDFS en un cluster se llama Name Node. El NameNode almacena los metadatos,

Mapreduce

MapReduce es un framework que proporciona un sistema de procesamiento de datos paralelo y distribuido. MapReduce está orientado a resolver problemas con conjuntos de datos de gran tamaño, por lo que utiliza el sistema de archivos distribuido HDFS,

El Framework MapReduce tiene una arquitectura maestro / esclavo. Cuenta con un servidor maestro. y varios servidores esclavos, uno por cada nodo del clúster.

El JobTracker (arquitectura maestro) es el punto de interacción entre los usuarios y el framework MapReduce.

El JobTracker gestiona la asignación de tareas y delega las tareas a los TaskTrackers(arquitectura esclavo). Los TaskTrackers ejecutan tareas bajo la orden del JobTracker y también manejan el movimiento de datos entre la fase Map y Reduce.

En conclusion basicamente el jobtracker se dedica a ordenar los datos en el orden del llegada de los usuarios y designar las diferentes tareas a realizar a los tasktrackers

(estos reciben el codigo a realizar,una vez hecho, lo reenvian al jobtracker para finalizar el proceso).

**Las aplicaciones que mas reconocidas que tienes son:**

Facebook

Debida a la gran cantidad de datos por los millones de usuario, el hadoop mas grande que tiene es de casi 2500 nucleos de cpu y 1 petabyte de disco duro(carga mas de 2501 gigabytes comprimidos por dia).

Ebay

También la utiliza para guardar datos de los usuarios, tienen un motor dre procesamiento llamado kylin con esta tecnología que utiliza el paradigma de ejecución mapReduce

Oracle

Utiliza esta tecnología ya que los datos de sus clientes aumentan un 40% cada año provocando mucho gasto monetario en almacenaje.(10.000 usd por terabyte). Con sus tecnologías logran reducirlo hasta 10 veces.

SalesForce

Es una empresa de servicios en la nube en estos servicios procesan más de mil millones de transacciones por día a las que se accede a través de múltiples canales: API, web y móvil. Por eso necesitan gran velocidad de procesamiento .

Yahoo

Como hablamos anteriormente esta tecnología para lo primero que se utilizo es para yahoo,siendo la que mas aporta a la comunidad, por esto es que es la empresa a quien le pertenecen los clústeres hadoop mas grandes del mercado