**Hadoop: La gran herramienta del Big Data**

Para nadie es un secreto que con el pasar de los años y el gran avance de la tecnología, hace que las personas estén más fácilmente conectadas unas a otras, esto conlleva a que las personas usen servicios que proveen las empresas envueltas en el mundo digital. A partir de esto, las empresas poseen mucha información de las personas y que con el pasar del tiempo va aumentando de una manera muy rápida, no solo el tamaño y la velocidad son propiedades clave de dicha información, también la variabilidad de la misma es un punto clave que se maneja en el procedimiento de esta. Con esta información, las empresas han buscado maneras de estudiarla para poder implementar cambios o estrategias nuevas en sus negocios, aquí es donde se buscó la necesidad de analizarla de manera rápida, ya que antes para analizarla se requería un poco más de tiempo y el análisis era a una proyección un poco más antigua, además eran basadas en modelos relacionales y estadísticas convencionales o paquetes de visualización. Aquí es donde aparece Big Data, donde su naturaleza principalmente en la forma no estructurada de los datos generados por las tecnologías modernas como, Facebook, Twitter, dispositivos GPS, Registros de llamadas, entre otros. Permitiendo que la información se procese de manera mucho más rápida y permita generar una nueva estrategia de negocio de la empresa al conocer las tendencias o gustos del usuario que el Big Data le permiten conocer.

Con esta nueva tecnología, se crearon muchas herramientas las cuales tratan con algunos de los tres tipos de Big Data que hay, los datos estructurados que consisten en datos que tienen bien definidos su longitud y su formato, como las fechas, cadenas de caracteres o números, estos se almacenas en base de datos relacionales o en los DataWarehouse. También están los datos no estructurados que son datos que están tal cual fueron recolectados, no poseen ningún tipo de formato, haciendo que no sea posible guardarlo en una tabla, como ejemplo de estos están los PDF, documentos multimedia, documentos de texto o correos electrónicos. Y por último están los datos semiestructurados que son aquellos datos que no se limitan a campos determinados, son datos que describen los objetos y la relación entre ellos, y pueden acabar siendo aceptados por convención. Como por ejemplo los archivos HTML, XML O JSON.

Hadoop es una de las herramientas para el Big Data, la cual es un framework de software escrito en Java bajo una licencia libre que nos permite el procesamiento distribuido de grandes cantidades de datos usando modelos de programación simple sobre un cluster. Este framework es usado por empresas conocidas como Yahoo y Facebook. Esta herramienta tiene como características, el procesamiento distribuido ya que a cada nodo le asigna un fichero, y a su vez paraleliza el tratamiento de los datos de forma que cada nodo de nuestro cluster de máquinas procese una parte de los datos, de esta manera optimizara la velocidad. También este framework, permite escalabilidad, ya que, si nuestro sistema necesita más velocidad, simplemente se agrega otro nodo a nuestro cluster de máquinas permitiendo que el sistema continúe con su funcionamiento sin sufrir cambios. Es un sistema tolerante a fallos ya que la información se replica en todos los nodos del sistema, haciendo que si se cae un nodo los demás poseen la información de este.

Un cluster Hadoop incluye un nodo maestro y multiples nodos esclavos, en el cual, la función del nodo maestro es de jobtracker, tasktracker, namenode y datanode. Un esclavo consiste en un nodo de datos y un tasktracker, utilización el protocolo SSH para la comunicación entre ellos.

La arquitectura de Hadoop se basa en 4 partes, El Distributed Computation(MapReduce), Distributed Storage(HDFS), YARN framework y el Common Utilities. El Common Utilities contiene todos los jar y librerias necesarias para ejecutar Hadoop, el YARN es el gestor de recursos de Hadoop, al ser este un sistema distribuido entre diferentes máquinas, con lo cual debe de haber algo que se encargue de la gestión en cada máquina, El HDFS es el sistema de archivo distribuido, escalable y portatil que está instalado en cada nodo de datos, diseñado para gestionar archivos muy grandes, y por ultimo tenemos el MapReduce que es un modelo de programación para dar soporte a la computación paralela sobre grande cantidades de datos en un cluster, además tiene una gran ventaja y es que es posible escoger y utilizar el lenguaje y las herramientas más adecuadas para la tarea concreta que se va a realizar.

Por lo tanto, Hadoop es un framework que se puede implementar sobre un hardware a un bajo costo, a su vez siendo totalmente gratuito su uso y permite a las empresas analizar y procesar sus millones de datos de los usuarios de manera fácil y rápida, es una opción muy preferible ante las otras opciones, aunque también es preciso recordar que al ser Open Source el mantenimiento de este se vuelve un poco tedioso, pero aun así existen distribuciones de Hadoop que tienen suporte 24/7, pero esto implica que su costo se eleve un poco, aunque a pesar esto, sigue siendo una opción bastante viable para desarrollar Big Data en las empresas que desean realizar estrategias para mejorar sus servicios.

# Bibliografía

PowerData. (5 de 9 de 2013). *PowerData Blog*. Obtenido de https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/bid/328879/c-mo-se-relacionan-big-data-y-hadoop

PowerData. (s.f.). *PowerData*. Obtenido de https://www.powerdata.es/big-data

Wikipedia. (23 de 4 de 2018). *Wikipedia*. Obtenido de https://es.wikipedia.org/wiki/Apache\_Hadoop