Tipo : Lectura
Capítulo : Introducción al Big Data

I. OBJETIVO

Ampliar sus conocimientos sobre Big Data

II. LECTURAS COMPLEMENTARIAS

Se sugiere revisar los siguientes enlaces para profundizar en los conceptos tratados en el presente capítulo.

c) Tecnologías y herramientas para el almacenamiento y tratamiento de datos de Big Data

¿Cómo se relacionan Big Data y Hadoop?

(Fuente: PowerData)

Historia y Business Drivers que lo impulsan a aparecer

La historia de Hadoop está necesariamente unida a la de **Google.** De hecho, podría decirse que Hadoop nace en el momento en que **Google** precisa urgentemente de una solución que le permita **continuar procesando datos** al ritmo que necesita, en una proporción que repentinamente ha crecido de forma exponencial.

Google se ve incapaz de poder **indexar la web al nivel que exige el mercado** y por ello decide buscar una solución, que se basa en un **sistema de archivos distribuidos**, haciendo suyo el lema “divide y vencerás”.

Esta solución, que posteriormente se denominará **Hadoop**, se basa en un gran número de pequeños ordenadores, cada uno de los cuales se encarga de procesar una porción de información. La grandiosidad del sistema es que, **a pesar de que cada uno de ellos funciona de forma independiente y autónoma, todos actúan en conjunto**, como si fueran un solo ordenador de dimensiones increíbles.

En 2006, **Google** publica todos los detalles acerca de su nuevo descubrimiento, compartiendo su conocimiento y experiencia con todos los usuarios que anhelaban acceder a esta información. Entre el conjunto de beneficiarios, destaca el interés de la **comunidad Open Source** que apasionados por la idea y el nuevo horizonte que se abre frente a ellos, explotan sus posibilidades desarrollando una implementación a la que denominan **Hadoop**.

A partir de ese momento, **es Yahoo quien toma el relevo** impulsando su expansión, para lograr alcanzar a grandes e icónicas empresas en el mundo de la informática, como **Facebook,** que empiezan a incorporarlo a sus rutinas, a disfrutar de su uso y a **participar en su desarrollo**, junto con la comunidad Open Source.

¿Qué es Hadoop?

Hadoop es un **sistema de código abierto** que se utiliza para **almacenar, procesar y analizar grandes volúmenes de datos**. Sus ventajas son muchas:

- Aísla a los desarrolladores de todas las dificultades presentes en la **programación paralela.**
- Cuenta con un ecosistema que sirve de gran ayuda al usuario, ya que permite **distribuir el fichero en nodos**, que no son otra cosa que ordenadores con commodity-hardware.
- Es capaz de ejecutar procesos en paralelo en todo momento.
- Dispone de **módulos de control** para la monitorización de los datos.
- Presenta una opción que permite **realizar consultas**.
- También potencia la **aparición de distintos add- ons**, que facilitan el trabajo, manipulación y seguimiento de toda la información que en él se almacena.

Los **componentes básicos de Hadoop** son los siguientes:

HDFS

Consiste en un **sistema de archivo distribuido**, que permite que el fichero de datos no se guarde en una única máquina, sino que sea capaz de **distribuir la información a distintos dispositivos.**

MAPREDUCE

Se trata de un framework de trabajo que hace posible aislar al programador de todas las tareas propias de la **programación en paralelo.** Es decir, permite que un programa que ha sido escrito en los **lenguajes de programación** más comunes, se pueda **ejecutar en un clúster de Hadoop.**

La gran ventaja es que hace posible escoger y utilizar el lenguaje y las herramientas más adecuadas para la tarea concreta que se va a realizar.

¿Para qué sirven Big Data y Hadoop?

Hadoop es un sistema que se puede **implementar sobre hardware a un costo relativamente bajo**, siendo a su vez **totalmente gratuito para software.**

Esta circunstancia comporta que, aquella información que antes las empresas no podían procesar debido a los límites de la tecnología existente o a barreras de tipo económico, que se hacían insalvables en muchos casos; hoy pueda ser almacenada, gestionada y analizada, gracias a **Hadoop.**

Cualquier organización que utilice **Hadoop** puede obtener información nueva, al mismo tiempo que descubre y aplica cualquier otro tipo de **análisis a sus datos,** como por ejemplo una regresión lineal sobre millones de registros de su histórico.

Es precisamente por ello que se está expandiendo tanto su uso entre las empresas que se benefician de:

- El costo relativamente bajo que implica.
- El **rápido retorno de la inversión** que proporciona.
- La **posibilidad de afrontar nuevos retos** y dar solución a problemáticas que antes no podían asumir, o que quedaban sin respuesta.

A su vez, para **minimizar los riesgos de su aplicación**, existen en el mercado distintas **distribuciones de Hadoop** con soporte 24/7, de esta forma ya no es necesario depender de la **comunidad Open Source** para solucionar este tipo de cuestiones; lo que ha contribuido a impulsar en gran medida su adopción en entornos productivos.