arsys

Qué es Hadoop y cómo se utiliza

Hadoop es un marco de software de código abierto utilizado para desarrollar aplicaciones de procesamiento de datos que se ejecutan en un entorno informático distribuido. Proporciona almacenamiento masivo para cualquier tipo de datos, una gran potencia de procesamiento y tiene la capacidad teórica de procesar tareas concurrentes virtualmente ilimitada.

En Hadoop, los datos residen en un **sistema de archivos distribuido denominado Hadoop Distributed File System** —HDFS—, un sistema que tiene capacidad para almacenar los archivos en un clúster de varias máquinas, lo que es esencial para poder almacenar enormes cantidades de datos —hablamos de petabytes—.

HDFS es un sistema de ficheros diseñado para funcionar muy bien en la lectura secuencial de los datos, y el hecho de ser un sistema distribuido tiene ventajas adicionales a la de poder almacenar una cantidad potencialmente ilimitada de datos: proporciona escalabilidad, ya que basta con añadir un nodo al clúster para aumentar la capacidad de almacenamiento. Además, proporciona redundancia, lo que hace a Hadoop un sistema **tolerante a fallos.**

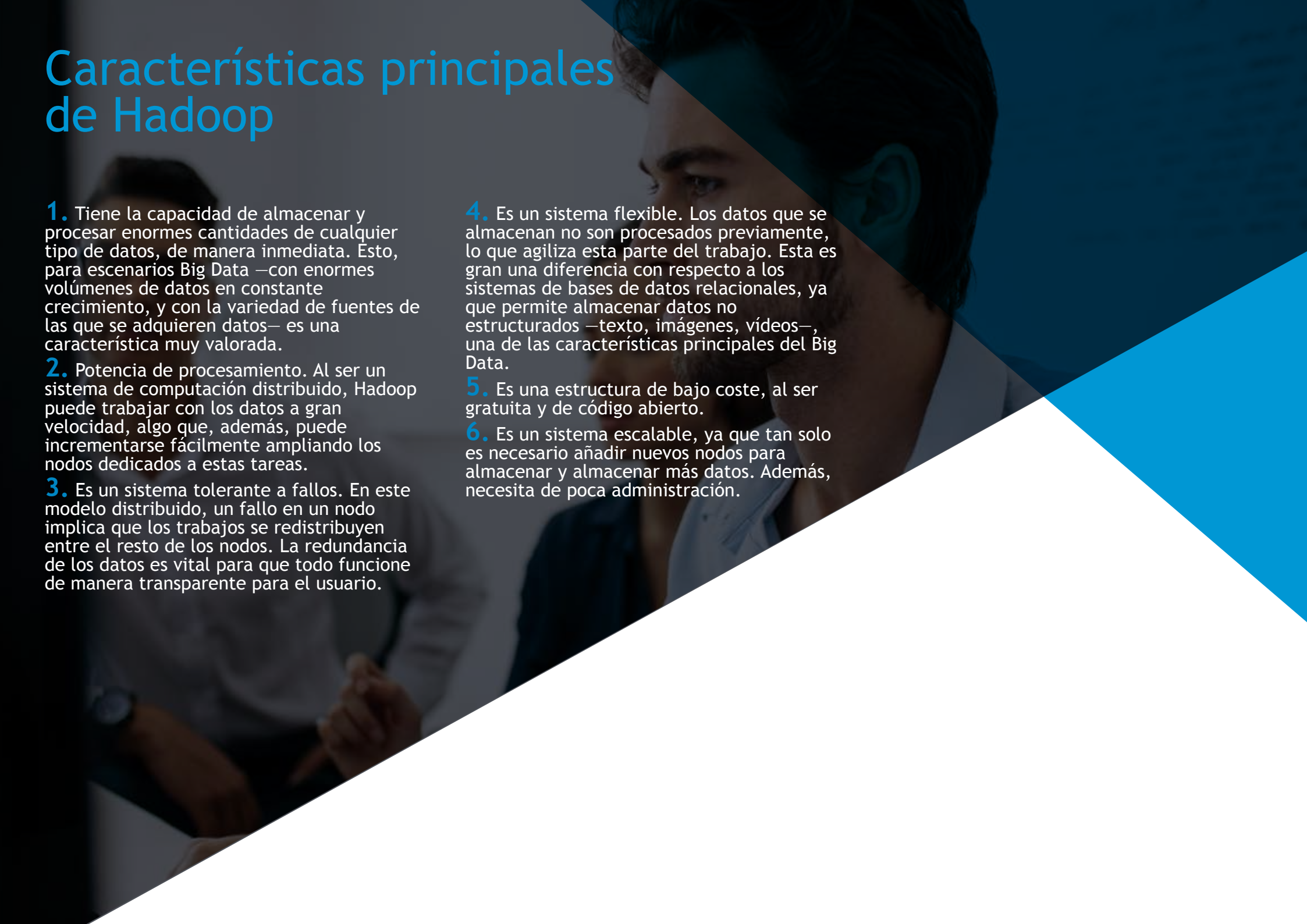
La importancia de la topología de red en Hadoop

Para un sistema distribuido como Hadoop, la topología de red afecta directamente a su rendimiento cuando la red de nodos crece. Además, si se desea que el sistema sea tolerante a fallos y tenga una alta disponibilidad, la manera en que se organizan los nodos es vital.

En Hadoop, la red se representa como un árbol en el que la distancia entre nodos es igual a la suma de su distancia a su ancestro común más cercano.

Hadoop es una arquitectura perfecta para procesamiento de *Big Data*, y por eso su adopción es cada vez mayor. Sus principales características, como su potencia de procesamiento, tolerancia a fallos y su capacidad de almacenamiento ilimitada, así como el bajo coste de implementación son, sin duda, responsables de su éxito.

Características principales de Hadoop



1. Tiene la capacidad de almacenar y procesar enormes cantidades de cualquier tipo de datos, de manera inmediata. Esto, para escenarios Big Data —con enormes volúmenes de datos en constante crecimiento, y con la variedad de fuentes de las que se adquieren datos— es una característica muy valorada.
2. Potencia de procesamiento. Al ser un sistema de computación distribuido, Hadoop puede trabajar con los datos a gran velocidad, algo que, además, puede incrementarse fácilmente ampliando los nodos dedicados a estas tareas.
3. Es un sistema tolerante a fallos. En este modelo distribuido, un fallo en un nodo implica que los trabajos se redistribuyen entre el resto de los nodos. La redundancia de los datos es vital para que todo funcione de manera transparente para el usuario.
4. Es un sistema flexible. Los datos que se almacenan no son procesados previamente, lo que agiliza esta parte del trabajo. Esta es gran una diferencia con respecto a los sistemas de bases de datos relacionales, ya que permite almacenar datos no estructurados —texto, imágenes, vídeos—, una de las características principales del Big Data.
5. Es una estructura de bajo coste, al ser gratuita y de código abierto.
6. Es un sistema escalable, ya que tan solo es necesario añadir nuevos nodos para almacenar y almacenar más datos. Además, necesita de poca administración.

Componentes de Hadoop

El proyecto Hadoop es bien conocido, sobre todo, por dos de sus componentes principales: MapReduce y HDFS.

Hadoop **MapReduce** es un modelo computacional y un marco de software para escribir aplicaciones que se ejecutan en Hadoop, y que son capaces de procesar enormes datos en paralelo en grandes grupos de nodos de cómputo.

Por otro lado, ya hablamos de **HDFS**, el sistema de archivos distribuido de Hadoop. HDFS se encarga de la parte de almacenamiento de las aplicaciones de Hadoop, de forma que las aplicaciones MapReduce consumen datos de HDFS. Al tratarse de un sistema de archivos distribuido es posible realizar cálculos fiables y extremadamente rápidos.

Otros componentes populares de Hadoop son, en realidad, proyectos relacionados concebidos para la computación distribuida y el procesamiento de datos a gran escala. Algunos de ellos son:

- **Hive:** una infraestructura de almacenamiento de datos construida sobre Hadoop para proporcionar agrupación, consulta, y análisis de datos
- **Hbase:** una base de datos distribuida no relacional de código abierto
- **Mahout:** para producir implementaciones gratuitas de algoritmos de aprendizaje automático distribuidos o escalables enfocados principalmente en las áreas de filtrado colaborativo, agrupación y clasificación
- **Sqoop:** una aplicación con interfaz de línea de comando para transferir datos entre bases de datos relacionales y Hadoop.
- **Flume:** un servicio distribuido, fiable, y altamente disponible para recopilar, agregar, y mover eficientemente grandes cantidades de datos.
- **ZooKeeper:** que ofrece un servicio para la coordinación de procesos distribuido y altamente confiable que da soluciones a varios problemas de coordinación para grandes sistemas distribuidos.

arsys

www.arsys.es

www.facebook.com/arsys.es

twitter.com/arsys

www.linkedin.com/company/arsys-internet/