Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

NoSQL reúne conceptualmente a todas las bases de datos que no siguen los principios y conceptos de las bases relacionales. Es decir, NoSQL es un conjunto de tecnologías que representa diferentes principios para el almacenamiento y manipulación masiva de datos. Desde los años 80, la mayoría de los sistemas de información almacenan sus datos en gestores relacionales. Básicamente, estos recogen procesos transaccionales que gestionan datos simples y estructurados, por ejemplo, transacciones bancarias, compras, ventas, etcétera. La introducción de NoSQL en el mercado fue una respuesta para llenar los espacios donde los sistemas gestores de base de datos relacionales ya no podían ser útiles por la cantidad enorme de datos y la velocidad requerida de almacenamiento y recuperación. Paralelamente, los avances en virtualización condujeron a la construcción de nodos de computación en la nube (Cloud Computing), que a su vez requerían nodos de almacenamiento (Cloud Storage) para procesar de forma paralela cantidades masivas de datos. Google creó un sistema de archivos distribuidos, una base de datos orientada a columnas, un sistema gestor distribuido y un algoritmo llamado MapReduce, que da soporte al procesamiento paralelo de datos masivos. Este algoritmo junta todos los pares con la misma clave de todas las listas y los agrupa creando un grupo por cada una de las diferentes claves generadas, para luego transformar la lista de pares clave-valor en una lista de valores.

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

El término Big Data se refiere a esa cantidad masiva de datos. El concepto de Big Data alude a conjuntos de datos cuyo tamaño excede la capacidad de los típicos sistemas gestores de base de datos para capturarlos, almacenarlos, gestionarlos y analizarlos. El motor de un avión, por ejemplo, es capaz de generar más de 10 TB de datos en solo 30 minutos de vuelo. Netflix utiliza modelos predictivos para informar de forma personalizada las novedades y recomendaciones, clasificando los datos del pasado y los programas vistos o marcados como favoritos, los cuales representan una cantidad también enorme de datos. Hay cada vez más dispositivos conectados a internet recopilando datos sobre patrones de uso de clientes, redes sociales y eficacia de productos. ¿Por qué no usarlos para que las computadoras aprendan por sí solas? Así nació el aprendizaje automático o Machine Learning. El objetivo que persiguen las tecnologías Big Data es obtener información que guardan enormes cantidades de datos; esto debe lograrse a través del análisis profundo de esos datos.

Gráfico

Descripción generada automáticamente

En Big Data hay tres tipos de datos: estructurados, semiestructurados y no estructurados. En cada uno de estos tipos hay mucha información útil que se puede extraer para usarla en distintos proyectos. Los datos estructurados tienen un formato fijo y, a menudo, son numéricos, así que en muchos casos los gestionan las computadoras y no los humanos. Este tipo de datos se encuentra ordenado en bases de datos de tipo SQL y almacenes de datos (Data Warehouse). Los datos no estructurados son información desorganizada que no está en un formato predeterminado, como los datos recopilados de redes sociales, que pueden convertirse en archivos de documentos de texto almacenados en sistemas NoSQL. Los datos semiestructurados pueden contener ambas formas de datos, como registros de servidores web o datos de sensores. Big Data siempre incluye múltiples fuentes y, la mayoría de las veces, de distintos tipos, por lo que no siempre es fácil saber cómo integrar todas las herramientas para trabajar con distintos tipos de datos. La necesidad de gestionar datos también requiere una infraestructura estable y bien organizada. Será necesario procesar rápidamente grandes volúmenes y distintos tipos de datos, lo que puede sobrecargar un único servidor.

Diagrama

Descripción generada automáticamente con confianza media

Big Data se caracteriza por las llamadas 6V: Volumen (Google procesa diariamente 20 petabytes), Velocidad (los datos deben ser almacenados y procesados casi en tiempo real), Variedad (la mayoría de las veces son datos no estructurados, como multimedia y metadatos), Veracidad (es importante la exactitud de los datos para evitar análisis imprecisos y decisiones equivocadas), Valor (es necesario establecer una forma de valorar y limpiar los datos, aceptando solo los relevantes) y Variabilidad (se refiere a datos cuyo valor cambia constantemente dependiendo de su propósito).

Según el teorema de Brewer, no es posible que un sistema distribuido garantice simultáneamente tres características: consistencia, disponibilidad y tolerancia a particiones. Consistencia significa que, después de una operación de escritura, todas las operaciones de lectura posteriores pueden ver las actualizaciones. Disponibilidad implica que el sistema puede seguir operando en lecturas y escrituras incluso si un nodo falla. Tolerancia a particiones es la habilidad de un sistema para funcionar aunque una parte de la red quede inaccesible. En un sistema distribuido, la caída de un nodo es probable, por lo que el sistema deberá elegir entre mantener consistencia o disponibilidad.

Diagrama

Descripción generada automáticamente

En cuanto a la arquitectura de bases de datos NoSQL, generalmente ofrecen una consistencia débil o eventual. Emplean una arquitectura distribuida donde los datos están almacenados de forma redundante en varios servidores. Ofrecen estructuras de datos simples como arreglos asociativos o estructuras llave-valor. Las consultas se realizan exclusivamente por llaves o índices, y las consultas complejas requieren una infraestructura de procesamiento externo como MapReduce.

Texto

Descripción generada automáticamente

Las bases de datos relacionales usan transacciones ACID (atomicidad, consistencia, aislamiento y durabilidad). Sin embargo, las bases de datos NoSQL, que no se rigen por este principio, adoptan el modelo BASE, que abarca conceptos como disponibilidad básica, estado eventual y consistencia eventual. En el modelo BASE, el sistema se volverá eventualmente consistente una vez que deje de recibir información.

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Dentro de NoSQL, existen varios modelos, como las bases documentales y las bases columnares.

Diagrama

Descripción generada automáticamente En las bases documentales, se estructuran datos en documentos (ej. XML, JSON, Word, PDF). Estos permiten almacenar textos complejos y consultarlos de forma eficiente.

Diagrama

Descripción generada automáticamente con confianza mediaLas bases columnares, por otro lado, organizan los datos en pares de llave-valor, almacenados como diccionarios. Estas bases de datos son más eficientes en accesos a disco y se usan frecuentemente en ambientes de Data Warehouse para análisis OLAP.

Interfaz de usuario gráfica, Diagrama

Descripción generada automáticamente

Texto, Carta

Descripción generada automáticamente

Algunos casos de uso conocidos de bases de datos NoSQL incluyen Cassandra (Facebook, Netflix), Couchbase (PayPal, eBay), Dynamo (Amazon), HBase (Pinterest, Yahoo), y Redis (Twitter, GitHub). Estas bases son adecuadas para aplicaciones con grandes volúmenes de datos, alta demanda de accesos de lectura y escritura, y cambios frecuentes en los esquemas de datos.

Imagen de la pantalla de un celular con texto

Descripción generada automáticamente con confianza media

Tabla

Descripción generada automáticamente

Las bases de datos NoSQL tienen algunas desventajas, como una capacidad de consulta limitada y un modelo no estandarizado.

Imagen de la pantalla de un celular con texto

Descripción generada automáticamente con confianza media

Eso trae problemas de portabilidad entre diferentes entornos el desarrollo de aplicaciones clientes es más complejo y carecen de una herramienta de control de acceso las cuestiones de seguridad son débiles