Introducción © EDICIONES ROBLE, S.L.

Indice

Introducción

Introducción. Antonio Sarasa

Descarga: máquina virtual

Importante: para este módulo, se les facilita a los alumnos una máquina virtual con los programas necesarios ya instalados, así como con un documento en su interior con instrucciones de uso. La máquina virtual se puede descargar en el siguiente enlace:

https://imfformacion-my.sharepoint.com/:u:/g/personal/masterbigdata_imf_com/ESGjzsqM3ptJrerSoHnQiNEBnhqw3r_I5qFWeTkWlfj7fQ?e=a29Rcg

Una de las características más representativas del Big Data tiene que ver con los datos que hay que gestionar. En las últimas décadas, el almacenamiento y gestión de la información se ha realizado mayoritariamente utilizando bases de datos relacionales. Este tipo de sistemas almacena datos que se corresponden con tipos básicos —no son estructuras complejas de datos—, la información se estructura de una manera fija, la cual viene definida de manera genérica, en forma de esquemas que son utilizados para generar las tablas de la base de datos, y el procesamiento de la información se realiza en ambientes centralizados.

Sin embargo, la información que se necesita procesar en la actualidad, y concretamente en el ámbito del denominado Big Data, no cumple las premisas comentadas. En primer lugar, el tipo de datos que se necesitan gestionar son, en muchos casos, semiestructurados o sin estructura. Esto se debe a que proceden de muy diversos dispositivos: sensores meteorológicos, wearables, teléfonos móviles… En este sentido, en caso de que los datos tuvieran estructura, esta será distinta en cada caso y puede corresponderse con estructuras complejas. Asimismo, no se puede garantizar que los datos que se vayan a almacenar puedan contener datos erróneos o vacíos. Es por ello que estas condiciones hacen inviable definir esquemas fijos para la información que se va a almacenar en una base de datos.

Por otro lado, la cantidad de datos que se necesita gestionar es muy elevada, y la velocidad de generación de estos es muy rápida. Estas condiciones hacen que sea necesario tener una capacidad de procesamiento muy potente, rápida y que sea fácilmente escalable, para poder dar servicio a un crecimiento exponencial de la información. En sistemas centralizados, esta necesidad se traduce en disponer de máquinas con unas altas prestaciones, las cuales se quedarán obsoletas en poco tiempo, debido al crecimiento exponencial en la generación de los datos. En este sentido, las soluciones que permiten escalar la potencia de cálculo mediante un entorno distribuido o clúster de máquinas son más adecuadas y más baratas que los sistemas centralizados.

Por tanto, estas nuevas condiciones plantean problemas para seguir utilizando las bases de datos relacionales como sistema de persistencia para este nuevo escenario, aunque fueran el origen de la aparición de estos nuevos sistemas surgidos con el Big Data.

En este módulo se van a estudiar los fundamentos conceptuales de las bases de datos NoSQL y las familias que existen y, a continuación, se analizarán los principales modelos de bases de datos NoSQL existentes.

1

En la unidad 1 del módulo, se realiza una introducción a las **bases de datos NoSQL**. El objetivo de esta unidad es plantear al estudiante cuáles son las nuevas necesidades de almacenamiento y procesamiento de información que surgen en el contexto del Big Data, qué limitaciones presentan las bases de datos relacionales frente a estas necesidades que motivan la aparición de otros modelos de persistencia y, por último, se describen las principales familias de bases de datos NoSQL. En este sentido, se espera que el estudiante asimile las razones que motivan la aparición de las bases de datos NoSQL en el contexto del Big Data, que sepa valorar en qué escenarios son interesantes y en cuáles no, así como que obtenga una idea general de la variedad de sistemas que han aparecido y que se encuadran como bases de datos NoSQL.

2

En la unidad 2 de este módulo, se estudia el primer modelo de bases de datos NoSQL. Se trata de las **bases de datos orientadas a documentos**, uno de los modelos más utilizados dada la flexibilidad que ofrece. En este modelo, se utiliza como unidad de almacenamiento una estructura de datos que se denomina genéricamente documento. Esta estructura de datos representa datos complejos que están compuestos por otros datos y que podrían estar formados a su vez por otros documentos. Este modelo presenta dos ventajas importantes: por una parte, la facilidad para modular la información del mundo real mediante los documentos y, por otra parte, la correspondencia natural que existe entre los documentos y otras estructuras de datos que aparecen en los lenguajes de programación, tales como los registros, lo cual evita realizar transformaciones de la información cuando se recupera de la base de datos y se quiere procesar con un lenguaje de programación. El modelo se ilustrará con una base de datos orientada a documentos denominada MongoDB. Los objetivos principales de esta parte son entender el modelo conceptual en el que se basan, entender cuándo pueden ser interesantes de utilizar y conocer a nivel básico cómo funciona MongoDB y su lenguaje de consultas.

3

En la unidad 3 del módulo se estudia el segundo modelo de bases de datos NoSQL. Se trata de las **bases de datos orientadas a columnas**. Si se considera su uso y popularidad, es similar al de las bases de datos orientadas a documentos. Una de sus grandes ventajas es la fuerte similitud con el modelo relacional, dado que tanto la terminología como la unidad de almacenamiento hace referencia a las tablas. Mientras que en el mundo relacional se almacena la información por filas, en este modelo se almacena la información por columnas. En este sentido, todos los datos que se refieren a una misma columna se almacenan juntos, y la unidad de intercambio de información se refiere a los datos de una columna como una única unidad. Esta similitud con las bases de datos relacionales llega incluso al lenguaje de consultas, el cual es muy parecido tanto en sintaxis como en comandos al lenguaje SQL de las bases de datos relacionales. Para ilustrar este modelo, se utilizará la base de datos Cassandra. En cuanto a los objetivos principales de esta área, son entender el modelo conceptual en el que se basa; apreciar las diferencias que presenta el modelo con las bases de datos relacionales; entender cuándo es interesante utilizar este modelo frente al resto de modelos NoSQL y, en particular, frente a las bases de datos documentales (dado que comparte con estas el ser un modelo orientado hacia agregados); y conocer a nivel básico como funciona Cassandra y su lenguaje de consultas.

4

En la unidad 4 del módulo, se estudia el tercer modelo de bases de datos NoSQL: las **bases de datos orientadas a grafos**. Se trata de un modelo muy distinto con respecto a los modelos orientados a columnas o a documentos. Estos últimos son orientados hacia agregados, es decir, existe una unidad de almacenamiento que se gestiona como una unidad (en un caso los documentos y, en otro caso, las familias de columnas). Sin embargo, este modelo no entiende agregados y utiliza como unidad de almacenamiento una estructura de grafo. Además, difiere en otros aspectos del resto de modelos NoSQL, como por ejemplo en su modo de ejecución, ya que la mayoría de bases de datos NoSQL están preparadas y diseñadas para ejecutarse en entornos distribuidos; sin embargo, el modelo orientado a grafos está diseñado para entornos centralizados. Así pues, este modelo será muy útil cuando se estén gestionando datos jerarquizados o que presenten un número importante de relaciones entre los mismos. Para ilustrarlo, se utilizará la base de datos Neo4j. Finalmente, uno de los objetivos importantes de esta unidad es apreciar cuándo interesa utilizar este modelo en el que las relaciones son importantes frente al modelo relacional en donde también es importante las relaciones entre tablas. Asimismo, se espera que el estudiante adquiera un dominio básico del lenguaje de consultas Cypher.

5

En la unidad 5 del módulo, se estudia el último modelo de bases de datos NoSQL: las **bases de datos orientadas a clave-valor**. Estas bases de datos están emparentadas con las bases de datos orientadas a documentos y a columnas en el sentido de que están orientadas hacia agregados. En este caso, el agregado está constituido por una estructura de datos que almacena una secuencia de pares clave-valor. Sin embargo, se diferencia del resto de modelos en la opacidad de la información que se almacena, dado que en estos sistemas actúan como un repositorio de información donde la información representada no puede ser gestionada ni manipulada. El modelo se basa en un sistema de claves que son asociadas a la información que se almacena, las cuales sirven para identificar a cada elemento de información y poder así recuperarlo o eliminarlo. Sin embargo, si se quiere procesar la información representada o consultarla, habrá que hacerlo fuera de la base de datos, pues esta permanece oculta. Una ventaja de este modelo es que puede almacenar cualquier tipo de datos, pero tiene como desventaja que todo procesamiento se deberá hacer fuera del contexto de la base de datos. Para ilustrar este modelo, se utilizará la base de datos Redis. En definitiva, uno de los objetivos principales de esta unidad es que el estudiante adquiera las competencias para distinguir cuándo un sistema que aparentemente es rudimentario en cuanto a procesamiento de la información puede ser interesante como solución de persistencia. Probablemente, en estos casos, sea necesario buscar una solución combinada con otra base de datos, además de la de clave-valor.

6

Por último, en la unidad 6 se tratará el tópico de la **ingesta de datos**. En el contexto del Big Data, se genera una enorme cantidad de datos que se quieren procesar. Para que estos procesos sean más eficientes, se han creado aplicaciones específicas para gestionar los datos a procesar, de modo que estas aplicaciones quedan liberadas de realizar esta tarea. En esta unidad se estudian dos ejemplos concretos de este tipo de herramientas: Apache Kafka y Apache Flume.

Los objetivos generales del módulo que los alumnos alcanzarán tras su estudio pueden resumirse en los siguientes:

1. Comprender y conocer las limitaciones de las bases de datos relacionales para cubrir las necesidades actuales de procesamiento y almacenamiento de información.
2. Conocer los fundamentos conceptuales en los que se basan las bases de datos NoSQL.
3. Conocer los principales modelos de bases de datos NoSQL, sus características y su utilidad.
4. Saber interpretar la información almacenada en una base de datos NoSQL de los tipos estudiados.
5. Saber manipular y consultar de manera básica la información almacenada en una base de datos NoSQL de los tipos estudiados.
6. Saber diseñar una solución para un dominio concreto usando una base de datos NoSQL.

Recursos

Enlaces de Interés

https://imfformacion-my.sharepoint.com/:u:/g/personal/masterbigdata_imf_com/ESGjzsqM3ptJrerSoHnQiNEBbuB4Yqiia_t7kr8VFCfKog?e=peSsLR
https://imfformacion-my.sharepoint.com/:u:/g/personal/masterbigdata_imf_com/ESGjzsqM3ptJrerSoHnQiNEBbuB4Yqiia_t7kr8VFCfKog?e=peSsLR

Máquina virtual del módulo de almacenamiento e integración de datos

https://imfformacion-my.sharepoint.com/:u:/g/personal/masterbigdata_imf_com/ESGjzsqM3ptJrerSoHnQiNEBnhqw3r_l5qFWeTkWlfj7fQ?e=a29Rcg
https://imfformacion-my.sharepoint.com/:u:/g/personal/masterbigdata_imf_com/ESGjzsqM3ptJrerSoHnQiNEBnhqw3r_l5qFWeTkWlfj7fQ?e=a29Rcg

Para este módulo, se les facilita a los alumnos una máquina virtual con los programas necesarios ya instalados, así como con un documento en su interior con instrucciones de uso.