DIPLOMADO “CIENCIA DE DATOS”

Módulo 3

ANÁLISIS Y MANIPULACIÓN DE BASE DE DATOS

1. Introducción

1.1 Conceptos básicos

Una base de datos es una colección organizada de información estructurada que se almacena en una computadora o sistema de almacenamiento. Permite el almacenamiento, la gestión y la recuperación eficiente de datos. Algunos conceptos de base de datos son:

a) Sistema de gestión de bases de datos (SGBD): Es un software que permite crear, manipular y administrar bases de datos. Proporciona una interfaz para interactuar con la base de datos, realizar consultas, insertar, actualizar y eliminar datos, entre otras tareas.
b) Tabla: Es la estructura fundamental de una base de datos relacional. Se compone de filas y columnas. Cada fila representa una entidad o registro, y cada columna contiene un atributo o característica de ese registro. Por ejemplo, una tabla de empleados puede tener columnas como "nombre", "edad" y "salario".
c) Registro: También conocido como fila, es una instancia o entidad específica en una tabla. Cada registro contiene valores o información relacionada con los atributos definidos en las columnas de la tabla.
d) Columna: También llamada campo, es una estructura que define el tipo de datos que se almacenará en ella. Representa una característica o atributo específico de un registro. Por ejemplo, en una tabla de clientes, una columna podría ser "nombre" o "dirección".
e) Clave primaria: Es un campo o conjunto de campos que identifica de manera única cada registro en una tabla. No puede contener valores duplicados ni valores nulos. La clave primaria se utiliza para garantizar la integridad de los datos y para establecer relaciones con otras tablas.
f) Clave foránea: Es un campo en una tabla que establece una relación con la clave primaria de otra tabla. La clave foránea permite conectar dos tablas y definir la relación entre ellas. Ayuda a mantener la integridad referencial y permite realizar consultas que involucran datos de ambas tablas relacionadas.
g) Consulta: Es una solicitud o instrucción enviada a una base de datos para recuperar información específica. Las consultas permiten filtrar, ordenar y manipular los datos almacenados en la base de datos de acuerdo con determinados criterios.
h) Índice: Es una estructura de datos utilizada para mejorar la velocidad de las operaciones de búsqueda y recuperación de datos en una base de datos. Un índice se crea en una o más columnas de una tabla y permite acceder

rápidamente a los registros que coinciden con un valor o conjunto de valores específicos

Estos son solo algunos de los conceptos básicos de las bases de datos. Hay muchos otros términos y conceptos más avanzados relacionados con la modelización de datos, normalización, transacciones, etc.

1.2 Objetivo de la base de datos

El objetivo principal de una base de datos es proporcionar un medio eficiente y estructurado para almacenar, gestionar y recuperar datos. Algunos de los objetivos más importantes de las bases de datos son:

a) **Almacenamiento estructurado**: Una base de datos proporciona una estructura organizada para almacenar datos de manera que se puedan acceder y manipular fácilmente. Permite definir tablas, relaciones y restricciones para garantizar la integridad de los datos.
b) **Acceso y recuperación eficiente de datos**: Una base de datos permite buscar y recuperar datos de manera rápida y eficiente. Utilizando índices, optimización de consultas y algoritmos de búsqueda, se puede acceder a los datos de manera más eficiente que en los enfoques de almacenamiento de datos no estructurados.
c) **Compartir datos**: Las bases de datos permiten que múltiples usuarios accedan y compartan los datos de manera simultánea. Esto facilita la colaboración y la gestión de datos en entornos donde varias personas necesitan acceder a la información.
d) **Mantenimiento de la integridad de los datos**: Las bases de datos permiten aplicar restricciones y reglas para garantizar la integridad de los datos almacenados. Esto incluye la validación de datos, la aplicación de reglas de negocio y el mantenimiento de la consistencia de los datos.
e) **Seguridad de los datos**: Las bases de datos ofrecen mecanismos de seguridad para proteger los datos almacenados. Esto incluye la autenticación de usuarios, la autorización de accesos y la encriptación de datos sensibles.
f) **Respaldo y recuperación**: Las bases de datos suelen proporcionar mecanismos para realizar copias de seguridad de los datos y recuperarlos en caso de fallos o errores. Esto garantiza la disponibilidad y la continuidad de los datos en situaciones adversas.

En resumen, el objetivo principal de una base de datos es proporcionar un entorno seguro y eficiente para el almacenamiento, gestión y recuperación de datos, permitiendo un acceso rápido, compartido y confiable a la información. Esto ayuda a las organizaciones a tomar decisiones informadas, gestionar sus datos de manera efectiva y optimizar sus operaciones.

El objetivo de una base de datos en ciencia de datos es proporcionar una plataforma eficiente y estructurada para almacenar, gestionar y acceder a grandes volúmenes de datos de manera sistemática. La base de datos es una parte fundamental de la infraestructura en ciencia de datos, y tiene varios objetivos clave:

a) **Almacenamiento y organización de datos**: El objetivo principal de una base de datos es proporcionar un medio para almacenar datos de manera organizada y estructurada. Permite guardar grandes cantidades de información de manera segura y accesible.
b) **Acceso eficiente a los datos**: Una base de datos bien diseñada permite acceder rápidamente a los datos necesarios para análisis y consultas. Utilizando índices y técnicas de optimización, se pueden recuperar datos de manera eficiente, lo que es esencial cuando se trabaja con grandes conjuntos de datos.
c) **Integración de datos**: En ciencia de datos, a menudo se trabaja con datos provenientes de múltiples fuentes, como bases de datos, archivos, servicios web, sensores, redes sociales, entre otros. La base de datos proporciona un lugar centralizado para integrar estos datos y facilitar su análisis.
d) **Pre procesamiento de datos**: Antes de realizar análisis o modelado, es común que los datos requieran pre procesamiento, como limpieza, imputación de valores faltantes, transformación y filtrado. La base de datos facilita estas tareas de pre procesamiento para garantizar la calidad de los datos.
e) **Seguridad de los datos**: La base de datos debe asegurar que los datos estén protegidos y accesibles solo para usuarios autorizados. La seguridad es fundamental para garantizar la privacidad y confidencialidad de la información.
f) **Apoyo al análisis y modelado**: La base de datos proporciona la fuente de datos para el análisis estadístico, la minería de datos y el aprendizaje automático. Los modelos y algoritmos utilizados en ciencia de datos requieren datos de alta calidad y confiables, que la base de datos proporciona.
g) **Gestión de la información histórica**: La base de datos puede mantener registros históricos de los datos, lo que es útil para el análisis de tendencias a lo largo del tiempo y para la toma de decisiones basada en datos históricos.

En resumen, el objetivo de una base de datos en ciencia de datos es proporcionar un entorno seguro, eficiente y organizado para almacenar y gestionar grandes volúmenes de datos. Esto es esencial para realizar análisis significativos, obtener conocimientos valiosos y tomar decisiones informadas basadas en datos. La base de datos es una pieza fundamental en la infraestructura de ciencia de datos que permite que el proceso de análisis sea más efectivo y preciso.

1.3 Modelos de base de datos

Un modelo de datos es una representación estructurada y visual de la organización lógica de una base de datos. Es una descripción abstracta de cómo se organizarán y relacionarán los datos en una base de datos determinada. El modelo de datos define los tipos de datos que se utilizarán, las relaciones entre los datos y las reglas o restricciones que se aplicarán a los mismos.

Existen varios tipos de modelos de datos, pero los más comunes son:

a) **Modelo de datos relacional**: Este es el modelo más ampliamente utilizado en la actualidad. Se basa en el concepto de tablas, donde cada tabla representa una entidad o relación y las filas de la tabla representan los registros. Las relaciones entre las tablas se establecen mediante claves primarias y claves foráneas.
b) **Modelo de datos jerárquico**: Este modelo organiza los datos en una estructura jerárquica similar a un árbol, donde cada registro tiene una relación padre-hijo. Los datos se organizan en niveles, con un registro raíz en la parte superior y registros secundarios debajo de él. Este modelo fue popular en los primeros sistemas de bases de datos, pero ha sido reemplazado en gran medida por el modelo relacional.
c) **Modelo de datos de red**: Es similar al modelo jerárquico, pero permite relaciones más complejas entre los datos. En este modelo, los registros pueden tener múltiples relaciones y pueden accederse a través de diferentes caminos. Aunque fue popular en el pasado, también ha sido reemplazado por el modelo relacional en gran medida.
d) **Modelo de datos orientado a objetos**: Este modelo se basa en el paradigma de programación orientada a objetos, donde los datos se representan como objetos con propiedades (atributos) y comportamiento (métodos). Permite la encapsulación de datos y soporta herencia y polimorfismo. Es utilizado en sistemas que requieren un enfoque más orientado a objetos, como aplicaciones multimedia y sistemas de información geográfica.
e) **Modelo de datos NoSQL**: Este modelo se refiere a una variedad de enfoques de bases de datos que no siguen la estructura rígida del modelo relacional. Las bases de datos NoSQL (Not Only SQL) están diseñadas para manejar grandes volúmenes de datos no estructurados o semiestructurados, como documentos, gráficos o datos en clave-valor.

Estos son solo algunos ejemplos de modelos de datos, cada uno con sus propias características y casos de uso. El modelo de datos elegido dependerá de los requisitos y la naturaleza de los datos que se deben almacenar y gestionar en la base de datos.

1.4 Clasificación de base de dato

Las bases de datos se pueden clasificar en diferentes categorías según diversos criterios. Algunas de las clasificaciones más comunes son:

a) Según el modelo de datos:
   - **Bases de datos relacionales**: Utilizan el modelo relacional, que organiza los datos en tablas con filas y columnas, estableciendo relaciones entre ellas mediante claves primarias y foráneas.
   - **Bases de datos jerárquicas**: Organizan los datos en una estructura jerárquica de registros padre-hijo.
   - **Bases de datos de red**: Permiten relaciones complejas entre los datos, con registros que pueden tener múltiples relaciones.

b) Según el propósito y la aplicación:
   - **Bases de datos transaccionales**: Diseñadas para gestionar transacciones en entornos de negocios, donde se requiere un alto rendimiento y la integridad de los datos.
   - **Bases de datos analíticas**: Optimizadas para consultas y análisis complejos de grandes volúmenes de datos. Se utilizan para la toma de decisiones y el descubrimiento de información.
   - **Bases de datos de búsqueda**: Se centran en la búsqueda y recuperación eficiente de información, como en motores de búsqueda o sistemas de recuperación de información.

c) Según la ubicación y el acceso a los datos:
   - **Bases de datos locales**: Se ejecutan y almacenan en una única máquina o servidor.
   - **Bases de datos distribuidas**: Los datos se almacenan y se distribuyen en varios servidores interconectados, lo que permite un acceso y una disponibilidad distribuida.
   - **Bases de datos en la nube**: Los datos se almacenan y se accede a ellos a través de servicios en la nube, lo que ofrece escalabilidad, flexibilidad y disponibilidad bajo demanda.

d) Según el modelo de consistencia y replicación:
   - **Bases de datos SQL**: Garantizan una consistencia estricta de los datos y utilizan el lenguaje de consulta SQL (Structured Query Language) para manipular y acceder a los datos.
   - **Bases de datos NoSQL**: Permiten una mayor escalabilidad y flexibilidad, relajando la consistencia de los datos y proporcionando diferentes modelos de datos, como documentos, gráficos o clave-valor.

Las bases de datos relacionales se remontan a la década de 1970, cuando Edgar F. Codd, un científico de IBM, publicó un influyente artículo titulado "A Relational Model of Data for Large Shared Data Banks" en 1970. En este artículo, Codd presentó

el concepto de un modelo de datos relacional que se basaba en la teoría de conjuntos y la lógica matemática.

La propuesta de Codd proponía un enfoque revolucionario para la organización y el almacenamiento de datos, donde los datos se estructuraban en tablas bidimensionales con filas y columnas, y las relaciones entre las tablas se establecían mediante claves primarias y claves foráneas. Esta novedosa idea permitía un alto grado de flexibilidad y eficiencia en la gestión de datos, y se considera la base de las bases de datos relacionales modernas.

En 1974, IBM lanzó el primer sistema comercial basado en el modelo relacional, conocido como IBM System R. Aunque inicialmente este sistema no fue ampliamente adoptado, sentó las bases para el desarrollo posterior de sistemas de bases de datos relacionales.

En 1977, Larry Ellison, Bob Miner y Ed Oates fundaron Oracle Corporation y lanzaron Oracle V2, que se considera una de las primeras bases de datos relacionales comerciales exitosas.

En la década de 1980, el modelo relacional ganó popularidad y se convirtió en el estándar de facto para el almacenamiento y gestión de datos. Muchos otros proveedores de bases de datos, como IBM, Microsoft, Sybase e Informix, lanzaron sus propios sistemas de bases de datos relacionales en esta década.

El lenguaje de consulta estructurada (SQL) se desarrolló en los años 70 y se convirtió en el lenguaje estándar para interactuar con bases de datos relacionales. SQL facilitó la manipulación y consulta de datos en un modelo relacional, lo que contribuyó aún más a la popularidad de las bases de datos relacionales.

En las décadas siguientes, las bases de datos relacionales continuaron evolucionando y mejorando en términos de rendimiento, escalabilidad, seguridad y capacidades de análisis. Además, surgieron sistemas de bases de datos relacionales de código abierto, como MySQL y PostgreSQL, que también ganaron popularidad en la comunidad de desarrollo y ciencia de datos.

En la actualidad, las bases de datos relacionales siguen siendo ampliamente utilizadas en una variedad de aplicaciones, desde pequeñas aplicaciones empresariales hasta grandes sistemas de misión crítica. Aunque han surgido otras tecnologías de bases de datos, las bases de datos relacionales siguen siendo una opción confiable y efectiva para el almacenamiento y gestión de datos estructurados.

En ciencia de datos, una base de datos es una colección estructurada y organizada de datos que se almacenan y gestionan de manera sistemática para su análisis, procesamiento y consulta. Estas bases de datos se utilizan para almacenar grandes cantidades de información que pueden provenir de diversas fuentes, como sensores, registros transaccionales, redes sociales, encuestas, entre otras.

A continuación, se describen algunos aspectos clave de las bases de datos en **ciencia de datos**:

a) **Estructura**: Las bases de datos en ciencia de datos pueden seguir diferentes modelos de datos, siendo el modelo relacional uno de los más comunes. En el modelo relacional, los datos se organizan en tablas con filas y columnas. Cada fila representa una entidad o registro, y cada columna almacena un atributo o característica del registro.
b) **Escalabilidad**: Las bases de datos en ciencia de datos deben ser capaces de manejar grandes volúmenes de datos, ya que los proyectos de ciencia de datos a menudo involucran conjuntos de datos masivos. La escalabilidad se refiere a la capacidad de la base de datos para manejar eficientemente el crecimiento de datos sin comprometer el rendimiento.
c) **Flexibilidad**: En muchos proyectos de ciencia de datos, los datos pueden ser variados en términos de su estructura y formato. Las bases de datos utilizadas en ciencia de datos deben ser lo suficientemente flexibles como para adaptarse a diferentes tipos de datos y esquemas.
d) **Consultas y análisis**: La base de datos debe ser capaz de realizar consultas y análisis complejos para extraer información relevante de los datos almacenados. Esto se logra mediante lenguajes de consulta como SQL (Structured Query Language) o herramientas de análisis como Python o R.
e) **Integración de herramientas de ciencia de datos**: Muchas bases de datos utilizadas en ciencia de datos ofrecen integración con herramientas populares de ciencia de datos, como bibliotecas de Python para análisis de datos, bibliotecas de aprendizaje automático, etc.
f) **Seguridad**: Dado que los datos almacenados en la base de datos pueden contener información sensible, la seguridad es fundamental para proteger los datos de accesos no autorizados y garantizar la privacidad.

En general, una base de datos bien diseñada y optimizada es esencial para el éxito de cualquier proyecto de **ciencia de datos**, ya que proporciona una base sólida para el almacenamiento y análisis de datos.