Manejo de RDBMS

Mtro. José Gustavo Fuentes Cabrera

Facultad de Estudios Superiores

Acatlán

Apuntes de Temas Selectos de Computación

Licenciatura en Actuaría

Índice

1. **Introducción** 3
2. **Diseños Relacionales** 3
3. **Diseños para minería de datos** 8
4. **Procesos Extracción-Transformación-Carga(ETL)** 10
5. **Cubos de procesamiento analítico en línea (Cubos OLAP)** 11

1. Introducción

Ahora centraremos nuestro estudio en la explotación y mantenimiento a sistemas gestores de bases de datos relacionales (RDBMS por sus siglas en inglés). En nuestro primer contacto con bases de datos realizábamos fundamentalmente diseño y tratábamos con pequeñas cantidades de datos. Si bien el diseño de datos es fundamental para el óptimo almacenamiento, evitar redundancias y favorecer el registro electrónico de la operación del negocio, en este curso estaremos en posibilidad de dar el siguiente paso hacia donde el futuro profesionista tendrá en sus manos la técnica para convertir los datos almacenados mediante el modelo relacional a información útil para la toma de decisiones.

2. Diseños Relacionales

El diseño relacional es el estándar en modelación clásica de datos, está basado en los trabajos de E.F. Codd en la década de 1970 cuya base es el concepto matemático conocido como relación. En términos simples, una relación (a veces llamada tabla) es una matriz donde se intersecan las filas y las columnas. A cada fila en la relación se le conoce como tupla y cada columna fungirá como un atributo de lo que estemos modelando. En 1976, Peter Chen propuso un método para representar gráficamente de entidades en una base de datos junto con sus relaciones, ésta técnica es llamada modelo entidad relación (ERM) el cual está basado en los siguientes componentes:

- Entidad: Cualquier cosa acerca de la cual se habrán de almacenar datos. Se representa mediante un rectángulo. En el modelo relacional está asociada a una tabla relacional donde cada fila será llamada instancia(ocurrencia) de entidad.
- Relación: Describen asociaciones entre los datos. Prácticamente todas las relaciones describen asociaciones entre dos entidades. Hay tres tipos de relaciones: uno a muchos(1:M), Muchos a muchos (M:N) y uno a uno (1:1). El nombre de la relación será un verbo.

Las tablas relacionales deben cumplir con las siguientes características:

- Es una estructura bidimensional compuesta de columnas y renglones.
- Cada tupla representa una ocurrencia única de la entidad dentro del conjunto de datos.
- Cada intersección de renglón-columna representa un valor único de datos.
- Cada columna representa un atributo de la entidad y posee un nombre distinto.
- Toda la columna debe poseer el mismo formato de datos
- Cada columna tendrá un intervalo específico de valores (dominio de atributo).
- El orden de los renglones y columnas carece de importancia
- Cada tabla tendrá un atributo (o conjunto de ellos) que identifique de manera única a cada renglón.

Lo anterior son conceptos previos que se revisan en la asignatura de Bases de Datos. Ilustremos con un ejemplo en una situación real como sería el modelo ER.
Suponga que usted trabaja en un Banco. Todos los bancos del mundo ofrecen el servicio de medios de pago a través de la emisión de plásticos (Tarjetas de Crédito o Débito). El medio plástico tiene como función disponer de los recursos (ya sea de una línea de crédito o de recursos propios en una cuenta corriente), dicha disposición se realiza mediante dos vías fundamentalmente: Disposición de efectivo en cajeros automáticos (ATM) o compras a comercio vía una terminal punto de venta (POS) sea física o virtual (compras por internet, móvil o teléfono). El banco donde usted trabaja posee un sistema automatizado donde todas las transacciones son registradas, tal sistema alimenta internamente a un modelo de datos relacional. ¿Cómo luciría dicho modelo relacional? Para contestar a esta pregunta requeriremos de un análisis profundo del funcionamiento del proceso así como de las entidades involucradas junto con sus relaciones. Por motivos didácticos, el proceso será simplificado. Veamos las entidades involucradas junto con sus atributos en este problema.

- **CLIENTE**
  - Nombre
  - ID
  - Entidad de la república
  - Sexo
  - Fecha de nacimento
  - Sucursal de adscripción
- **ESTADO**
  - Nombre del estado
  - Clave del Estado
  - ID
- **SUCURSAL**
  - Clave de sucursal
  - Entidad de la república
- **CUENTA**
  - Número de cuenta
  - Id Cliente
  - Tipo de cuenta
  - Id Producto
- **PRODUCTO**
  - Clave del producto
  - Nombre
- **PLASTICO**
  - Cuenta
  - PAN

- **TRANSACCION**
    - Entrada
    - Clave de autorización
    - Fecha
    - Hora
    - Monto
    - Estatus
    - Id plástico

El diagrama ER se muestra a continuación:

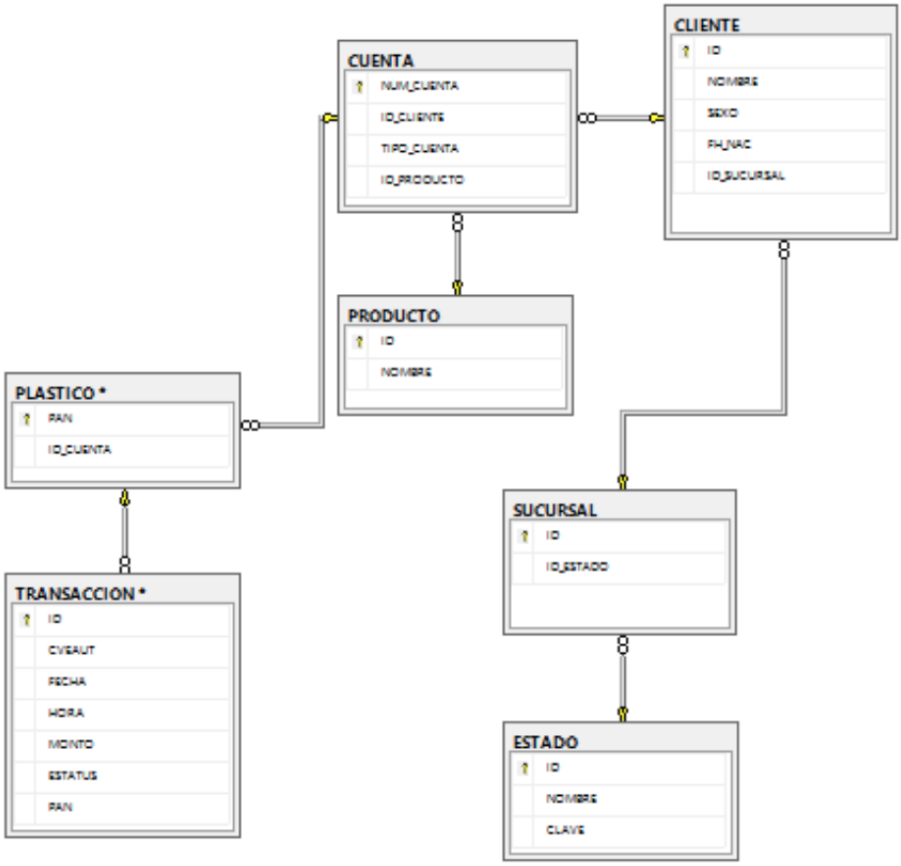

Figura 2.1: Diagrama E-R

3. Diseños para minería de datos

El diseño anterior nos permite almacenar de manera muy eficiente los datos, pero, ¿Qué pasa si queremos explotarlos? Desde un punto de vista técnico, implicaría un esfuerzo considerable en la construcción de consultas complejas que en grandes volúmenes de datos se vuelven poco eficientes si queremos dar respuestas de negocio rápidamente. Para un ejercicio tan sencillo como presentar todos los datos de un solo cliente tendríamos que ejecutar el siguiente query:

```
SELECT * FROM PLASTICO A INNER JOIN CUENTA B
ON A.ID_CUENTA=B.NUM_CUENTA
INNER JOIN CLIENTE C ON B.ID_CLIENTE=C.ID
INNER JOIN SUCURSAL D ON C.ID_SUCURSAL=D.ID
INNER JOIN ESTADO E ON D.ID_ESTADO=E.ID
INNER JOIN PRODUCTO F ON B.ID_PRODUCTO=F.ID
WHERE ID_CUENTA=2301
```

El query anterior implica 6 cruces, si bien para un solo registro esto no tendría importancia en tiempo de cómputo, en la industria puede significar horas de trabajo que derivarán en falta de productividad. Es por ello que existen diseños especiales que facilitan la explotación de datos.
Uno de los conceptos fundamentales es el de almacén de datos (datawarehouse); un Almacén de Datos es un conjunto de datos históricos, internos o externos y descriptivos en un contexto, que están integrados y organizados de tal manera que permitan aplicar de manera eficiente herramientas para resumir, describir y analizar los datos para soportar la toma de decisiones.
Antes de entrar en detalle, debemos revisar dos conceptos fundamentales:

- **OLTP (On-Line Transactional Processing)**: Procesamiento transaccional en tiempo real, es el trabajo principal de un sistema de información consistente en permitir las transacciones y dar soporte funcional a las aplicaciones de la organización. Véase como el trabajo diario para el cual se ha diseñado la base de datos.
- **OLAP(On-Line Analytical Processing)**: Procesamiento analítico en tiempo real. Es aquel que engloba un conjunto de operaciones, ex-

clusivamente de consulta, en donde se requiere agregar, cruzar e integrar grandes cantidades de información. Su propósito es el de generar desde resúmenes e informes simples hasta tableros de control muy sofisticados y modelos predictivos.

En la siguiente tabla podemos ver un resumen de las diferencias entre ambos enfoques de explotación de datos.

Cuadro 1: Diferencias OLTP-OLAP

Característica	**Base de datos Transaccional**	**Almacén de Datos**
Propósito	Operación diaria. Soporte a aplicaciones corporativas.	Recuperación de información, minería de datos.
Modelo de datos	Datos Normalizados	Estrella, copo de nieve, parcialmente desnormalizados, multidimensionales
Número de usuarios	Cientos/Miles	Pocas decenas
Tipos de usuarios	Operaciones, desarrolladores, DBA	Directores, Ejecutivos, Mineros, Científicos de datos
Acceso	SQL Lectura-Escritura	SQL (slice & dice, roll, drill, pivot) Lectura

El modelo conceptual más extendido en la arquitectura de los almacenes de datos es el llamado modelo multidimensional. En este modelo, los datos se organizan en dos grandes categorías: hechos y dimensiones. Los hechos responden a la pregunta "¿cuánto?"mientras que las dimensiones responderán al "¿cuándo?", "¿cómo?", "¿dónde?", "¿quién?", etc. ¿puede el lector determinar un ejemplo de lo anterior con los datos de nuestro modelo relacional?
Las dimensiones se agregan, por ejemplo la dimensión tiempo se puede organizar con la siguiente jerarquía:

Año→Trimestre→Mes→Día→hora
Semana→Día→hora

En este ejemplo, la dimensión tiempo posee varios caminos de agregación, esto se conoce como "copo de nieve"mientras que en caso contrario sería llamado "estrella". Debe recalcarse que toda dimensión será agregada en torno a un hecho ya que sin éste el almacén carecería de sentido. Ahora que sabemos el propósito de un almacén de datos es importante saber que no será posible organizar la totalidad de la información de una organización en una sola Estrella. Para resolver el problema, cada uno de los ámbitos específicos de la organización puede modelarse con su propia estrella, a esto se le llamará *datamart*. Un Almacén de Datos estará formado por un conjunto de datamarts.

4. Procesos Extracción-Transformación-Carga(ETL)

Una vez que hemos diseñado nuestra estructura para el almacén de datos, será necesario alimentarlo y darle mantenimiento al mismo mediante un sistema. Dicho sistema no es software específico, es un conjunto de procedimientos que se encarga de las siguientes tareas:

- Lectura de datos transaccionales
- Incorporación de datos externos
- Creación de claves
- Integración de datos
- Agregaciones
- Higiene de datos
- Mantenimiento y creación de metadatos
- Identificación de cambios
- Planificación de carga y mantenimiento
- Indización
- Pruebas de calidad

En la práctica muchas de estas labores son desempeñadas por software o personas independientes entre sí, incluso pudiese presentarse el caso en el

que se omitan una o varias partes, asimismo, mucha gente realiza la labor de forma "pasiva", es decir, extrae, transforma y carga datos, pero no es consciente de lo que implica la correcta e importante realización de esta labor.

5. Cubos de procesamiento analítico en línea (Cubos OLAP)

Con todo el antecedente con respecto a hechos, dimensiones, almacenes de datos y datamarts, estamos en posibilidad de comprender un cubo OLAP. Un cubo OLAP se define como una estructura de datos multidimensional; con mayor precisión, un cubo OLAP debiera ser referido como "hipercubo"para dimensiones mayores a 3. Esta estructura nos permite ver de manera intuitiva la agregación (menor detalle) y la disgregación (mayor detalle). Como ejemplo, en nuestro problema de transacciones en medios de pago tenemos como hecho "el 23 de enero del 2014 se tuvo un monto de $45,160 en transacciones rechazadas". Lo anterior puede ser visto como la intersección de un cubo con las dimensiones tiempo (23-enero-2014) y estatus (rechazado). Para completar nuestro estudio, revisaremos los operadores OLAP (recordemos que matemáticamente, un modelo de datos se compone de un conjunto de estructuras y lo operadores de las mismas):

- **Drill**: Funge como operador para disgregar datos (mayor detalle, menos sumarización).
- **Roll**: Funge como operador para agregar datos (menor detalle, más sumarización).
- **Slice & Dice**: Selección y proyección de datos.
- **Pivot**: Reorientación de dimensiones.