

BASE DE DATOS AVANZADAS DATA WAREHOUSE

Enrique Alfonso Calderón Balderas

Traslado de un modelo estrella al modelo relacional

- Cada Cubo se instrumenta con una tabla de hechos y varias tablas de dimensiones.
- Una tabla de dimensión tiene todos los atributos que la describen y el identificador generado por el sistema.
- La tabla de hechos tiene atributos para representar los indicadores y los identificadores que hereda de cada una de las dimensiones del Cubo.

Dimensiones

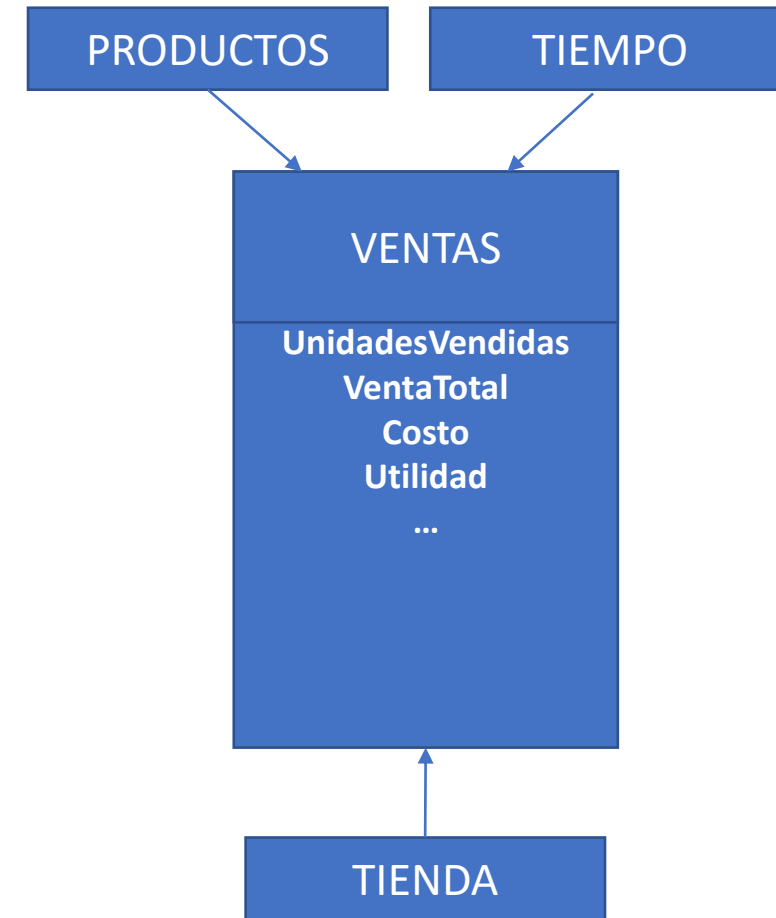
Productos(productId, codigo, descripcion, marca, modelo, ...)

Tiempo(tiempoid, anio, mes, día, semana, ...)

Tienda(tiendaid, codtienda, direccion, ciudad, estado, ...)

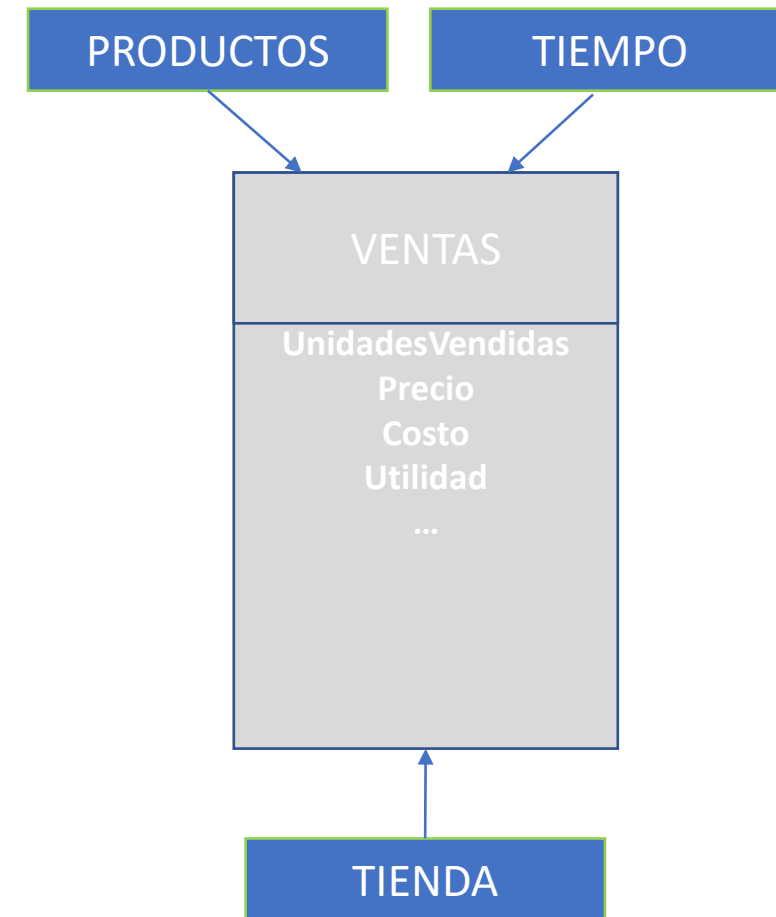
Hechos

Ventas(hechosId,productId, tiendaid, tiempoid, UnidadesVendidas, VentaTotal, Costo, Utilidad, ...)



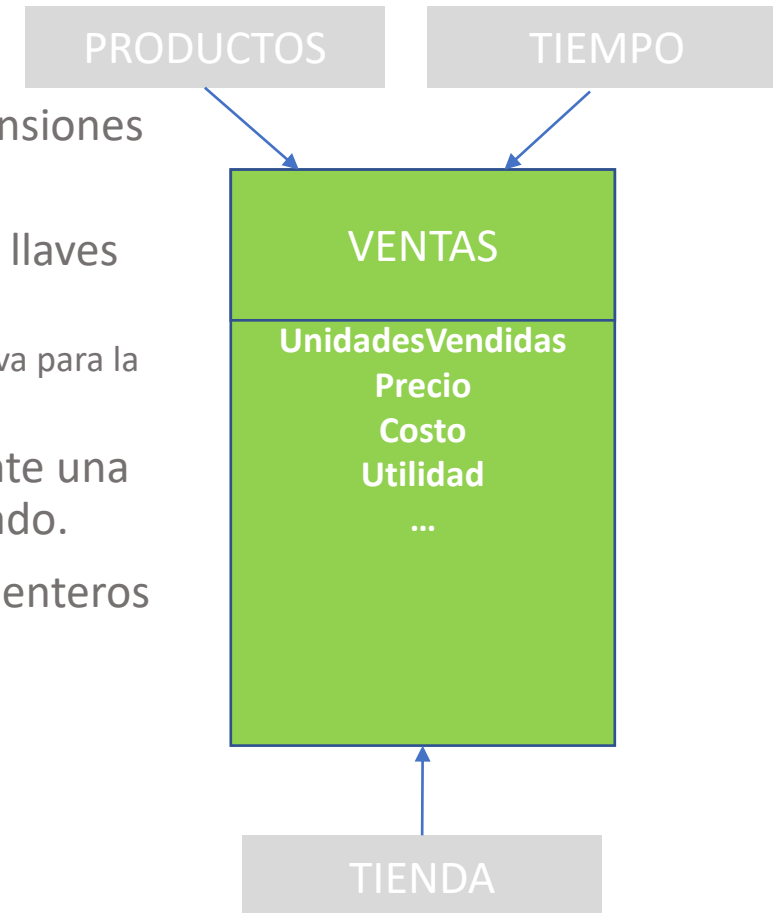
Tablas de dimensiones

- El identificador debe ser numérico para facilitar las búsquedas
- Debe contener una columna por cada atributo de la dimensión
 - **Esto es porque comúnmente es el filtro para los usuarios finales**
 - País
 - Estado
 - Día del año
 - Día de la semana
 - Día del mes
 - Marca
 - Etc.
- Ejemplo:
- PRODUCTOS (productId, codigoproducto, descripcion, marca, modelo, color, dimensiones, ...)



Tablas de hechos

- Incluye el identificador de cada una de las tablas de dimensiones con la que se relaciona.
- Comúnmente, la llave primaria es la concatenación de las llaves foráneas.
 - En algunas ocasiones es más práctico crear una llave primaria exclusiva para la tabla de hechos y usar las llaves foráneas como identificadores.
- Se incluye una columna para cada indicador que represente una medida de desempeño para el modelo de negocio analizado.
- Los indicadores son comúnmente “números” con valores enteros o reales.
- Pueden representar:
 - Cantidad de Dinero
 - Demoras o duraciones
 - Diferencias (de incrementos o decrementos)
 - Porcentajes
 - Sumatorias
 - Etc.





Integridad relacional del modelo.

- Se define un constraint de llave primaria o un índice único para el identificador de cada tabla de dimensión.
 - Evitar usar la llave del modelo transaccional o de las fuentes.
- En la tabla de hechos se puede definir como llave primaria:
 - la concatenación de todos los identificadores de las dimensiones.
 - Una llave exclusiva de la tabla de hechos y como llaves foráneas los identificadores de las dimensiones.



Diseño de índices para incrementar el desempeño de consultas.

- Todas las llaves primarias serán índices del tipo “UNIQUE” (valor por defecto en la construcción del PK)
- Para tablas con gran cantidad de registros es necesario definir índices (non-unique) para columnas con registros de múltiples valores
- Indexar aquellas columnas por las que se realizaran una gran cantidad de búsquedas.
- Particionar los índices
- Alojarse los índices en discos físico diferentes a los de datos
- Sobre-indexe (PERO NO EXAJERE)





Ejercicio

- Identifica de los esquemas cargados en tu máquina virtual de Oracle hechos que podrían definirse como parte de cubos para un DataWarehouse (esquema de la matricula)
 - Identifica al menos 4 hechos relevantes.
 - Genera preguntas que respondan esos hechos teniendo en cuenta el tiempo.
- Definir dos cubos a partir de las preguntas de indicadores. Identifica que tablas podrían usarse para generar las tablas de dimensiones y que tablas se usarían para generar los hechos.
- Define las tablas relacionales de los cubos propuestos.
- Solo una persona del grupo envia las respuestas a la cuenta de correo enrique.calderon@tec.mx a mas tardar el próximo viernes 19 de abril antes de las 7 am.

BASE DE DATOS AVANZADAS DATA WAREHOUSE-ETL



ETL

- Proceso que permite a las organizaciones mover datos de múltiples fuentes, reformatearlos y limpiarlos, y cargarlos en otro sistema para apoyar los procesos de inteligencia de negocios.
 - **Extraction** – (Extracción): Obtener información de las Bases de Datos transaccionales o fuentes externas de datos que alimentan el DWH
 - **Transform** – (Transformación): Agrupaciones, sumalizaciones, traducciones, codificaciones y decodificaciones, depuración, validación.
 - **Load** – (Carga): Llenado de las tablas de dimensiones y hechos que representan los cubos del DWH
- 
- 



Técnicas de ETL - Extracción

- Es la primera parte del proceso de ETL
- Consiste en “extraer” los datos desde los distintos sistemas de origen.
 - Bases de Datos Relacionales
 - Bases de Datos no relacionales
 - Archivos planos
 - ...
- Se extrae la información en un formato específico para realizar la transformación de la misma.
- Se debe garantizar el mínimo o nulo impacto en el sistema origen.




Técnicas de ETL – Extracción a un archivo

- Los DBMS ofrecen mecanismos para extraer el resultado de una consulta a un archivo.
- El archivo extraído de una fuente, se trasfiere al servidor de la BD del DWH para su carga.
- Ejemplo en Oracle:
 - `set head off;`
 - `set pagesize 0;`
 - `spool nombre_archivo.dat`
 - `select ... from ... where ...;`
 - `spool off;`

Técnicas de ETL – Extracción a una tabla

- Puede usarse dentro de una misma instancia o entre instancias de DBMS homogéneos o heterogéneos, que disponen de mecanismos de vinculación entre instancias.
- Consiste en crear una nueva tabla o insertar en una tabla existente, usando como fuente de datos una consulta.
- Ejemplos:
 - **//Pueden usarse como fuentes tablas o vistas, remotas o locales**
 - Oracle
 - `create table nuevaTabla as select... from... where...;`
 - `insert into tablaExistente (campos) select... from... where...;`
 - SQL-Server
 - `select... into nuevaTabla from... where...;`
 - `insert into tablaExistente (campos) select... from... where...;`



Técnicas de ETL – Uso de cursores

- Los cursores permiten recorrer y procesar uno a uno los renglones de las tuplas del conjunto de resultados de una consulta.
- Son preferibles a la opción de crear una nueva tabla o insertar en una existente cuando el número de registros a extraer es muy grande.
- Requieren memoria de la instancia y pueden demandar bloqueos.
- Al usarse para operaciones de extracción, lo más conveniente es definirlos sólo de lectura (read-only)
- Los cursores se utilizan generalmente en procedimientos almacenados.
- Si el DBMS soporta cursores que combinen tablas locales y remotas, es posible integrar las tres operaciones de ETL.

Técnicas de ETL – Uso de cursores

- Implementación en Oracle

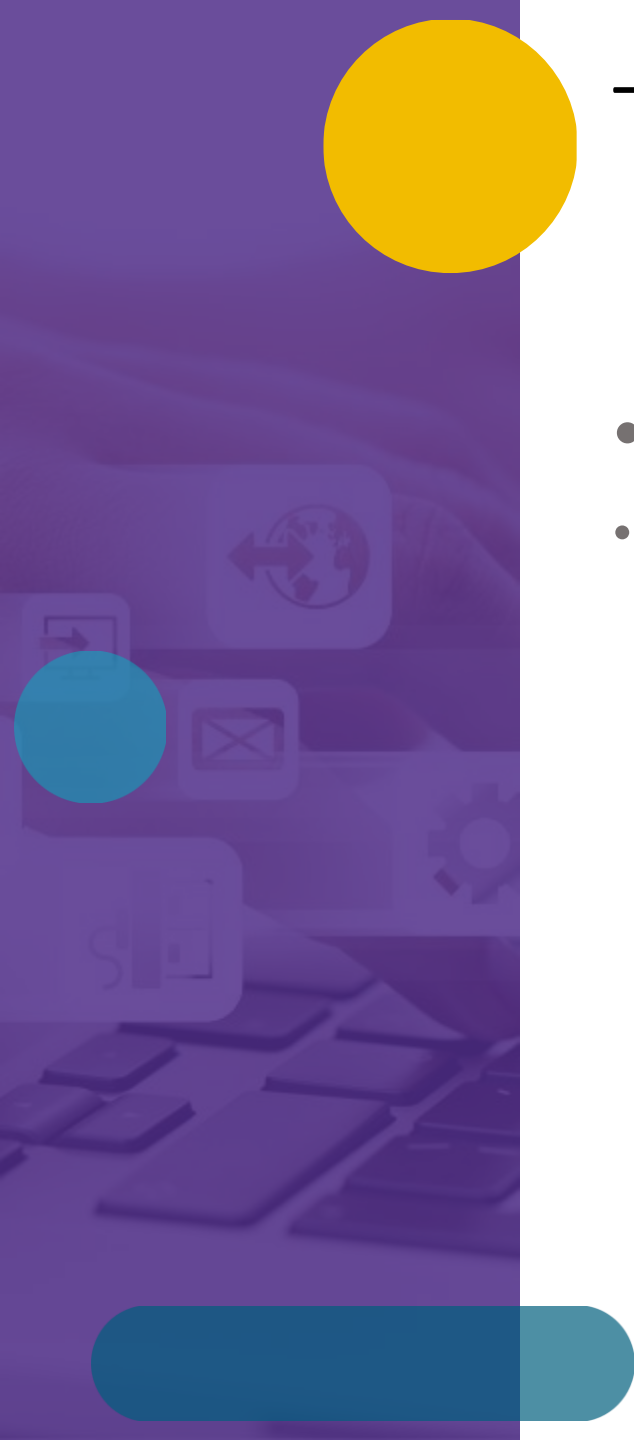
- Declaración implícita:

- **for variableRenglon in (select... from... where...) loop**
- **//acceso a los resultados mediante variableRenglon.columna**
- **end loop;**

- Declaración explícita:

- **declare cursor nombreCursor is select... from... where...;**
- **open nombreCursor;**
- **fetch nombreCursor into lista_de_variables;**
- **close nombreCursor;**

- **Atributos: nombreCursor%FOUND, nombreCursor%NOT_FOUND, nombreCursor%ISOPEN, etc.**



Técnicas de ETL – Uso de cursores

- Implementación en SQL-Server

- Declaración explícita:

- **declare nombreCursor cursor for select... from... where...**
- **open nombreCursor**
- **fetch [next from] nombreCursor into lista_de_variables;**
- **close nombreCursor;**
- **deallocate nombreCursor**

- Variable del sistema: @@fetch_status



Técnicas de ETL – Transformación

- En la fase de transformación se aplican una serie de reglas de negocio o funciones sobre los datos extraídos para convertirlos en datos “validos” a ser cargados.
- En algunas ocasiones las transformaciones son mínimas.
- En otras ocasiones las transformaciones deben ser más complejas (la mayoría de las veces)
- Transformaciones “típicas”
 - Traducir códigos (Hombre, H, Masculino en las fuentes a 1 en los destinos)
 - Operaciones de agrupación, sumarización, promedios, etc. (ventas=precio*no_productos)



Técnicas de ETL – Uso de precompiladores de C

- Si se requieren operaciones de transformación costosas, que se aplicaran a un gran número de registros, lo más eficiente es escribir un programa en C que pueda utilizar cursores para extraer los datos
- Los resultados pueden insertarse en la Base de Datos del DWH o enviarse a archivos para su carga posterior.
- Los DBMS ofrecen APIs para usar cursores desde C:
 - Oracle Pro*C: pre compilador de C
 - SQL-Server: DBLibrary (versiones 2000 y anteriores), ActiveX Data Objects (ADO)
 - DB2: DB2 Express-C



Técnicas de ETL – Carga

- En la fase de carga los datos de la fase anterior (Transformación) son cargados al sistema destino.
- Dependiendo de los requerimientos de la organización, este proceso puede abarcar una gran cantidad de acciones diferentes.
- La información nueva puede sobrescribir los datos antiguos.
- El proceso de carga interactúa directamente con la base de datos destino.
 - Al realizar la carga todas las restricciones se tendrán que cumplir
 - PK-FK, no null, rangos, triggers, etc.
 - Las restricciones (si están bien definidas) garantizan la calidad de los datos.



Técnicas de ETL – Carga masiva desde archivos.

- Los DBMS ofrecen mecanismos para la carga optimizada de datos.
- Se manejan bloques de registros y cache buffers.
- Cuentan con esquemas para establecer el formato de la entrada y relacionarla con las tablas en modo APPEND (agregar) y REPLACE (reemplazar)
- Implementaciones
 - Oracle : sqlldr
 - SQL-Server: bcopy (BULK INSERT)
 - DB2: DB2 Load