TEMA 3: PROCESAMIENTO DE CONSULTAS Y OPTIMIZACIÓN

1.Soporte de los sistemas relacionales

Patrón de consultas estándar: Sistema ROLAP
A partir de él se obtienen informes mediante consultas:

- -Select de las dimensiones
- -Operaciones sobre los hechos
- -Join de hechos y dimensiones
- -Filtros y agrupaciones

El patrón básico sería:

- -Reducir dimensiones (condiciones de selección)
- -Join entre hechos y cada dimensión
- -Agrupar atributos de las dimensiones y agregar mediciones
- -Ordenar

El problema es que si hay muchos hechos, el JOIN genera también muchos registros.
Se presentan dos alternativas:
1) Que los hechos intervengan lo mínimo posible
2) Combinar las dimensiones reducidas y acceder sólo una vez a los hechos

Soporte de SQL: Varias versiones

- +SQL2:
  - -Incluye la orden SELECT
  - -Informes con subtotales. Por N campos se necesitan N+1 subconsultas
- +SQL3:
  - -Incluye modificadores multidimensionales.
  - -ROLLUP: subtotales con una única consulta
  - -CUBE: genera informe con todos los subtotales posibles
  - -DECODE (GROUPING): devuelve si se trata de una agrupación o NULL

2.Estándares de consulta e intercambio

En SMD el estándar de facto es MDX, similar a SQL

- -Cubo devuelto en XMLA (XML for Analysis), que es el estándar para cubos
- -SOAP (Simple Object Access Protocol) es usado para conectar clientes con servidores OLAP

3.Optimización y ajuste a nivel lógico

Agregados: Cubos obtenidos del base mediante Roll-Up para responder a consultas más rápido

- +Ventaja: Mejora en el tiempo de respuesta
- +Inconvenientes:
  - -Si cambia algo en el cubo base hay que trasladar los cambios a los agregados
  - -Ocupa más espacio
  - -Determinar número de agregados

Máximo= (Nºnodos D1 · Nº nodos D2 · … · Nº nodos DN) -1

Uso de los agregados:

- -Sistema amigable: Transparente para el usuario
- -Navegador de agregados: modifica la consulta para hacerla sobre el cubo adecuado
- -Se anota las veces que se accede a los agregados
- -Se construyen de mayor a menor tamaño
- -Se recorren desde el más pequeño para ver si responden a la consulta

4.Optimización y ajuste a nivel físico

Índice en mapa de bits

-Buenos para consultas

-Menos eficientes para modificaciones

Proceso de construcción:

-Una columna para cada valor distinto del campo a indexar

-Una fila por cada registro de la tabla a indexar

-En cada celda se pone 1 si el registro tiene el valor de esa consulta y 0 si no.

Se pueden usar tantos índices como se quiera

Permite responder consultas mediante operaciones lógicas

Índice de Join: Se materializa el JOIN entre las dimensiones y los hechos utilizando las llaves generadas correspondientes en las dimensiones afectadas.

Proceso general:

+Reducir las dimensiones:

-OR y AND necesarios para seleccionar registros que intervienen

-Las llaves generadas de los registros obtenidos se tratan como columnas del índice de JOIN entre la dimensión y los hechos.

-Las columnas anteriores se unen usando OR para conseguir una única columna para la dimensión

+Combinar las columnas de las dimensiones haciendo JOIN entre ellas

+Acceder a los hechos

Tema 4: Integración de sistemas

1.Integración de sistemas

Construcción de un almacén de datos: Integrar las fuentes en un SMD común para la empresa centrado en el foco de atención. Problema: Existencia de varios focos de atención.
Construcción de varios almacenes de datos: Se tienen varios focos de atención. Problema: Proliferación de sistemas OLAP.

Fabrica de Información Corporativa (FIC) => Creación de almacén de datos corporativo: Es una colección de datos orientados al tema, integrados, no volátiles e historiados, organizados para el apoyo de un proceso de ayuda a la decisión.

2.Componente ETL

Tareas a realizar una vez:
- -Identificar la fuente de datos
- -Identificar datos objetivo
- -Crear correspondencia fuente-objetivo
- -Definir modo de replicación de datos
- -Programar la replicación

Tareas a realizar frecuentemente:
- -Capturar datos necesarios
- -Transferir datos entre fuente y objetivo
- -Transformar datos capturados
- -Aplicar datos capturados al objetivo
- -Confirmar el éxito de la replicación
- -Documentar resultado
- -Mantener las definiciones de fuentes, objetivos y correspondencias

Extracción de datos:
Requiere:
- +Carga inicial (una vez)
- +Actualizaciones (tantas como requiera)

Enfoques:
- +Diferido: Se miran los datos cada X tiempo y se recoge ese valor. Se toma un resumen al final del periodo. Problema: Se pierde detalle.
  - -Método de Comparación de Imágenes: Se mantiene una copia de la imagen anterior y se compara con la actual.
  - -Método de Huella de Tiempo: Anota la fecha de la última modificación y registra los cambios en la fecha actual respecto a la anotada.
- +Inmediato: Registra cada cambio producido en los datos. Se ven todos los movimientos.
  - -Método Registrar Movimientos: La aplicación que maneja los datos guarda los movimientos cuando se producen.
  - -Método Registrar Movimientos con Disparadores: En un SGBD, que los disparadores guarden los movimientos cuando se producen.
  - -Método Registrar Movimientos con Archivos Log: En un SGBD, usar los log para obtener todos los movimientos producidos. Es un enfoque inmediato, pero los movimientos tienen mucho ruido.

Transformación e Integración (T):
- +Adaptar datos al modelo del Data Warehouse
  - -Unificar formatos
- +Comprobar:
  - -Datos incompletos
  - -Datos duplicados
  - -Datos erróneos o inconsistentes
  - -Datos vacíos o ilegibles
  - -Diferencias de codificación
  - -Agregaciones necesarias
- +Notificar errores a las fuentes

Carga (L):
- +Incorporar datos al sistema tras su transformación
- +¿Cuándo?
  - -Inicial: Todos los datos de las fuentes
  - -Actualizaciones periódicas: con datos modificados
  - -Agregaciones: Actualizar los agregados

3.Metadatos

-Datos sobre datos
-Facilitan la ETL
-Se encuentran en las fuentes de datos, en el programa de ETL, en el SMD...

4.Definición de proyectos de integración de sistemas

Componentes:
- +Obtención => Fuentes de datos
- +Almacenamiento => Almacén de datos corporativo
- +Acceso => Almacenes de datos corporativos

Enfoque orientado por requisitos:
- +Solo se miran los requisitos de los usuarios
- +Se hace el diseño y la implementación
- +Se simulan los datos
- Problema: en el sistema real no funciona

Enfoque orientado por datos:
- +Se cogen los datos de las fuentes
- +Se vuelcan en el SMD
- Problema: No tiene en cuenta las necesidades de los usuarios

Enfoque Mixto: Operaciones
- -Continuas
- -Basadas en valor: Teniendo en cuenta el beneficio que se va a sacar
- -Autónomas
- -Contando con la infraestructura operacional y física
- -No es importante el orden en que se tratan las partes