INFORMACIÓN CORPORATIVA Y TIPOS DE BASES DE DATOS.

Si bien los manejadores de bases de datos están diseñados para satisfacer los más variados escenarios y usos posibles es importante distinguir los ambientes de bases de datos usados con mayor frecuencia, ya que de una buena determinación de estos ambientes podremos realizar las configuraciones adecuadas que nos permitan tener el mejor rendimiento de nuestro RDBMS, los ambientes más comunes son:

- OLTP
- OLAP

A continuación procederemos a hablar de cada uno de ellos:

Ambientes Transaccionales (OLTP).

Un ambiente transaccional u OLTP (OnLine Transaction Processing) es aquel dedicado a la consulta y actualizaciones concurrentes a alta escala, un ejemplo típico de este tipo de ambientes son los sistemas de cajas registradores en tiendas de autoservicio, en estos ambientes se realiza una enorme cantidad de inserciones simultaneas (un insert por cada producto registrado) y se realizan gran número de actualizaciones simultaneas (al actualizar los inventarios, por ejemplo), estos sistemas se distinguen fundamentalmente por las siguientes características:

- Gran número de actualizaciones concurrentes.
- Queries de alta selectividad, es decir, cuando se consulta la B.D. solo un número muy pequeño de registros es devuelto.

A fin de garantizar la integridad transaccional en caso de falla es altamente recomendable, por no decir necesario, que la base de datos sea transaccional, en esta situación se recomienda el uso de niveles de aislamiento en COMMITED READ.

Ambiente de Toma de Decisiones (OLAP)

Un ambiente de toma de decisiones u OLAP (OnLine Analitical Processing) es aquel dedicado a la alta gerencia y dirección a fin de poder tomar decisiones correctas, oportunas y rápidas a partir de sistemas de información gerencial. En estos ambientes se manejan grandes volúmenes de información, pero no se realizan actualizaciones en línea, en estos sistemas de información se guarda únicamente la información necesaria para la toma de decisiones y se actualiza periódicamente. Un ejemplo claro es el sistema que genera reportes cada mes para saber cuánto dinero y de qué forma a perdido una compañía por los clientes que han se han cambiado a la competencia. Sus principales características son:

- Pocas o nulas actualizaciones en línea.
- Queries que devuelven gran cantidad de información acompañada por consolidados (sumatorias) y ordenados por criterios.
- Capacidad de realizar queries no planeados por personas que no tienen necesidad conocer técnicamente el RDBMS ni su arquitectura.

En estos ambientes se trabaja fundamentalmente con tablas estáticas, es decir, que no son actualizadas en línea, generalmente la información se carga periódicamente, en esquemas por lotes podríamos decir, y lo demás es consulta por parte de la alta dirección de la empresa. Dado este esquema el uso de logical logs es por demás innecesario, por lo que es común que en estos esquemas las bases de datos aparezcan sin logging. En toda base de datos sin logging no se pueden manejar niveles de aislamiento, siendo su situación equivalente al DIRTY READ.

Cuando se usa DIRTY READ no deben hacer actualizaciones y consultas al mismo tiempo, esto se debe a que al hacer actualizaciones y consultas al mismo tiempo con una base de datos en dirty read, informix puede devolver "registros fantasmas" es decir, uno o varios registros que están en proceso de actualización o de inserción, este estado no se considera como un error de ejecución, por lo que se deberá hacer los SELECT sin que ningún UPDATE, INSERT o DELETE esté corriendo simultáneamente.

A esta categoría pertenecen los Sistemas de Información Gerencial (SIG), en especial los DatawareHouses (DWH) y DataMarts (DM).

DATAWAREHOUSING.

Elementos de un DWH.

Una definición concreta de DWH podría ser, según Bill Inmon: "Es el conjunto de bases de datos diseñadas en forma integral, orientadas al sujeto para soportar las funciones de toma de decisiones, donde cada unidad es relevante en algún momento en el tiempo", o en palabras de Ralph Kimball: "Es una copia de datos transaccionales estructurados específicamente para su consulta y análisis".

Es decir, un datawarehouse es, en el sentido más amplio del término, un sistema de toma de decisiones con algunas características especiales tales como:

- Las bases de datos son creadas específicamente para la toma de decisiones, a diferencia de los OLTP. Las bases de datos son configuradas en sus propios ambientes los cuales son optimizados para obtener el mejor performance.
- La Data se extrae de uno o más sistemas fuentes y se integra en una fuente única. Como parte de esta integración, los datos son estandarizados, se eliminan posibles inconsistencias (data cleaning) y se realizan tareas de sumarización (Sumatorias, promedios, ordenamientos), a fin de que puedan ser explotados.
- Normalmente estos datos se usan para representar el estado de la corporación, así como para ser analizados para detectar patrones, tendencias, e incluso predicciones.

En un DWH se busca, a su vez, concentrar un repositorio de información al nivel de la compañía, a fin de que los resultados que afectan a toda la compañía sean precisos y la alta dirección tenga fuentes de información consistentes y adecuadas para la toma de decisiones.

Complemento lo anterior, y hablando de un modo más operativo, debemos contemplar el hecho de que nuestros gerentes y directores deben tener interfaces de consulta fáciles de manejar, que no requieran un conocimiento técnico elevado tales como: manejo de SQL o la arquitectura de base de datos, y que contemplen consultas no planeadas.

Para poder conseguir los puntos anteriores es necesario contemplar algunos componentes necesarios para lograr este objetivo:

- Tabla de hechos: Es la información consolidada sujeto de las consultas para la toma de decisiones.

- Dimensiones: Son básicamente los catálogos de la tablas de hechos, cabe mencionar que la normalización no es prioritaria en este esquema y, en ocasiones, es inclusive sacrificable en aras del performance, siendo necesario solamente el tener a la base de datos en primera forma normal.

Las herramientas mediante las cuales se puebla este datawarehouse, se les llama ETLs, siglas en inglés de Extract-Transform-Loading. Estas herramientas cumplen, como su nombre lo indica, tres funciones principales:

- Extraer los datos desde bases de datos, que pueden ser heterogéneas, esta actividad que puede antojarse simple, de hecho no lo es, especialmente cuando los repositorios son varios heterogéneos, ya que hay que estandarizar los datos de los diferentes repositorios, en los tipos de datos, columnas y discriminar el "ruido" de los datos realmente significativos. Así mismo en esta etapa se incluye lo que se conoce como "la limpieza de datos" o "Data Cleaning" que consiste en detectar y, en su caso, corregir diferentes inconsistencias que se pueden presentar entre los datos extraídos (por ejemplo, identificar que la compañía "Vázquez y Vázquez, S.A.", es la misma que "VyVSA", por su R.F.C., o por otras referencias indirectas), en el entendido de que estas inconsistencias no siempre pueden ser resueltas de manera automática.
- Transformar los datos para almacenar cierto grado de información, en lugar de almacenar los datos en bruto, por ejemplo, con promedios, modas, máximos y mínimos, agrupando estos datos por ciertos criterios (dimensiones).
- Por último, que no menos importante, está el proceso de carga, que puede estar basado en cargas incrementales, actualizaciones, o incluso poblar la tabla de hechos desde cero.

Existen diferentes herramientas ETL en el mercado para realizar estas tareas, en las que se pueden diseñar proyectos para realizar estas funciones, siendo extremadamente raro que haya desarrollos en casa para estas funciones.

Diferencia entre Datawarehouse y Datamart.

Quizá una de las principales confusiones sea la de distinguir un datawarehouse de un datamart, un datamart es, fundamentalmente, un subconjunto de un datawarehouse, mismo que está orientado a un departamento o unidad específica de negocios.

En un datamart la data tiene un alto grado de sumarización a nivel mensual, bimestral, trimestral, semestral o anual. No se actualiza con frecuencia pero la información se reemplaza a intervalos regulares a fin de mantener la precisión.

Estos datamart se usan normalmente para hacer decisiones a nivel de departamento o para analizar algún aspecto específico de la unidad de negocios más que una estrategia a nivel empresarial.

Conceptos vistos:

OLTP
OLAP
Dataware House
Datmart
Metadata
Tabla de hechos
Dimensiones
ETL