# Introducción a los almacenes de datos

**Orígenes de los almacenes de datos.**

Muchas organizaciones que usan tecnología de base de datos estándar para recopilar, almacenar y procesar grandes cantidades de sus datos operativos comenzaron a ver más de cerca sus almacenes de datos actuales e históricos como fuentes de información para ayudarse a tomar mejores decisiones empresariales. Han desarrollado almacenes de datos para sistemas de apoyo de decisiones (DSS, por sus siglas en inglés) y aplicaciones similares. Decisiones como dónde abrir un nuevo almacén, a qué audiencia dirigirse para una campaña publicitaria, a cuáles clientes otorgar préstamos y cuándo ordenar artículos adicionales se pueden hacer con más confianza cuando se basan sobre un cuidadoso examen de los patrones encontrados en los datos existentes. Los proveedores de DBMS, incluidos Oracle e IBM, rápidamente agregaron características a sus líneas de productos para permitir el almacenamiento de los datos a partir de sus sistemas de base de datos estándar. Se han desarrollado nuevas y poderosas herramientas analíticas para extraer más información de los datos almacenados en tales almacenes. SQL contiene extensiones que apoyan las funciones requeridas por los almacenes de datos. Los datos en un almacén de datos con frecuencia se recopilan a partir de varios departamentos o sitios que pertenecen a una gran empresa. El término lo acuñó W. H. Inmon, quien describió un almacén de datos como “una colección de datos orientada a sujeto, integrada, no volátil, variable en el tiempo que se utiliza principalmente en toma de decisiones organizacionales” (Inmon, 2002). Un almacén de datos se establece para aplicaciones que apoyan decisiones, y no para procesamiento de transacciones ordinarias. Está optimizado para recuperación de datos, en oposición a procesamiento de transacciones.

# Bases de datos operativas y almacenes de datos.

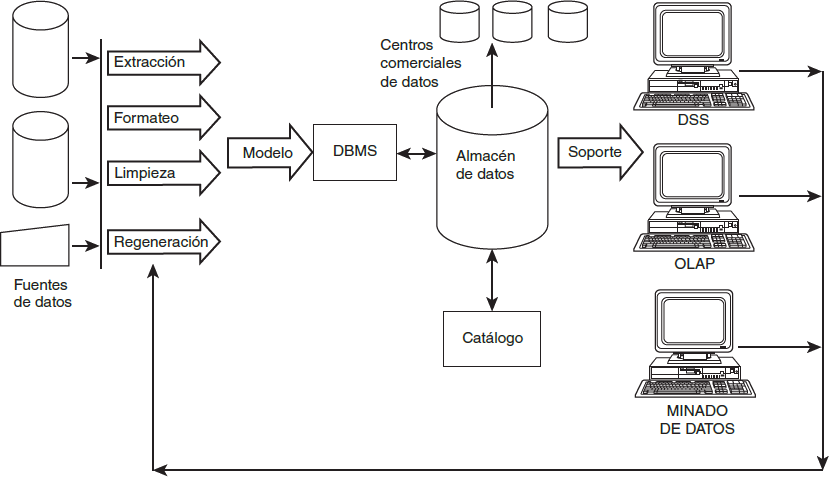
Las bases de datos operativas tradicionales soportan procesamiento de transacción en línea (OLTP, por sus siglas en inglés), que de manera característica involucran un número limitado de transacciones repetitivas, cada una de las cuales afecta algunas tuplas en un momento en una base de datos relacional. Una base de datos como ésta se desarrolla para servir a las necesidades de información de los usuarios finales, y está diseñada para soportar sus operaciones empresariales diarias. La alta disponibilidad y

desempeño eficiente son factores cruciales en el éxito de una base de datos operativa. Debe proporcionar apoyo para un gran volumen de transacciones y entregar respuestas a las consultas de los usuarios u otras operaciones en línea en un marco temporal corto. Una base de datos operativa se actualiza en tiempo real, conforme ocurren las transacciones del negocio. Actualizaciones, inserciones y borrados se deben realizar rápidamente para mantener la base de datos en un estado que refleje el entorno actual de la empresa.

En contraste, los almacenes de datos soportan OLAP (On-Line Analytical Processing: procesamiento analítico en línea), así como toma de decisiones. Los datos en un almacén de datos se pueden llevar directamente de bases de datos operativas múltiples, en diferentes periodos (datos históricos) y también pueden incluir datos de otras fuentes, datos resumidos y metadatos. Las fuentes pueden tener distintos modelos o estándares, pero el almacén de datos integra los datos de modo que los usuarios ven un modelo consistente. El almacén de datos por lo general contiene una cantidad muy grande de datos, y está optimizado para procesamiento eficiente de consultas y presentación de resultados para apoyo de decisiones. Las actualizaciones no son tan frecuentes como lo son en las bases de datos operativas, pero se realizan de manera periódica. Las aplicaciones OLAP por lo general deben pasar a través de grandes cantidades de datos para producir resultados. Los analistas examinan los datos almacenados en el almacén usando consultas complejas, que generalmente involucran operadores de agrupación y agregación. Pueden hacer análisis en series de tiempo usando datos históricos. El minado de datos es el proceso de descubrir nueva información mediante búsqueda de grandes cantidades de datos. El propósito es descubrir patrones o tendencias en los datos que serán útiles para la organización.

# Arquitectura de un almacén de datos.

A diferencia de una base de datos operativa, para la cual se pueden especificar requisitos por adelantado, un almacén de datos debe diseñarse para apoyar consultas ad hoc y nuevos y no anticipados tipos de análisis. En la siguiente figura se muestra una arquitectura típica.



Los datos se toman de fuentes de datos, que pueden incluir bases de datos operativas múltiples, otras entradas como archivos independientes y datos ambientales como información geográfica o datos financieros. Los datos se deben extraer de las fuentes con el empleo de herramientas externas al sistema que puedan acomodar las diferencias entre las fuentes heterogéneas. Los datos se reformatean en un formato consistente. Los datos también se pueden verificar para integridad y validez, un proceso llamado limpieza de datos, para asegurar su calidad antes de cargarlos en el almacén. Luego los datos se ponen en el modelo de datos para el almacén y carga. El proceso de carga es una transacción larga, pues por lo general está involucrado un gran volumen de datos, de modo que el sistema debe usar herramientas de gestión de transacción para garantizar recuperación adecuada en el evento de falla durante la transacción de carga. El sistema de gestión de base de datos que soporta al almacén de datos tiene un catálogo de sistema que almacena metadatos, así como otros componentes del sistema de la base de datos. Luego el almacén de datos se usa para soportar consultas para OLAP, con el fin de proporcionar información para sistemas de apoyo de decisiones que usan los administradores para toma de decisiones estratégicas, y a fin de proporcionar los datos para las herramientas de minado de datos que descubren nueva información acerca de patrones en los datos. Ciertos segmentos de los datos están organizados en subconjuntos llamados data marts (“mercado” de datos, subconjunto de información de un Dataware House), que se enfoca en sujetos específicos. Por ejemplo, un data mart

podría contener información especializada acerca de un solo departamento dentro de la organización. Todos estos usos pueden resultar en nuevo conocimiento, que luego se puede usar como una fuente de datos desde la que se pueden formatear datos y ponerlos en el almacén. Los datos de todas las fuentes se deben refrescar de manera periódica. Si hay suficiente espacio de almacenamiento, los nuevos datos simplemente se agregan al almacén existente, y los datos antiguos se mantienen en tanto sean útiles. De otro modo, los datos que ya no se usan más se purgan de manera periódica, y se agregan nuevos datos. La frecuencia y ámbito de las actualizaciones depende del entorno. Los factores que se deben considerar para decidir la política de actualización incluyen cuánto almacenamiento está disponible, si el almacén necesita datos recientes, si puede estar fuera de línea durante la regeneración, y cuánto tardará el proceso para transmitir los datos, limpieza, formateo, carga y construcción de índices. La política usual es hacer una regeneración parcial de manera periódica.

# ACTIVIDAD

Una vez que revises la lectura anterior responde

1. Qué es un Almacén de Datos?

es una colección de datos en la cual se encuentra integrada la información de la empresa u organización. Esta información es de utilidad en el proceso de toma de decisiones gerenciales.

1. Cuál es la diferencia entre un almacén de datos y una base de datos operativa?

|  |  |
| --- | --- |
| **Datos Primitivos / Operacionales** | **Datos Derivados / Data Warehouse** |
| Orientados a la aplicación | Orientados al sujeto |
| Detallados | Sumarizados, al menos refinados |
| Exacta, al momento del acceso | Representa valores a través del tiempo |
| Sirve a los oficinistas/trabajadores | Sirve a los directivos |
| Puede ser actualizada | No se actualiza |
| Se ejecuta repetitivamente | Se ejecuta heurísticamente |
| Los requerimientos para procesar se entiende a priori | Los requerimientos para procesar se entienden a posteriori |
| Compatible con entornos de desarrollo tradicionales | Incompatible con entornos de desarrollo tradicionales |
| Accesible atómicamente | Accesible por conjuntos de datos |
| Orientado a la transacción | Orientado al análisis |
| No redundante | Redundante |
| Estructura estática | Estructura dinámica |
| Pequeños volúmenes de datos | Grandes volúmenes de datos |
| Alto uso | Bajo uso |

1. Cómo funciona un almacén de datos?

* **Con una estructura básica**, sistemas operativos y archivos planos proporcionan datos en bruto que se almacenan junto con metadatos. Los usuarios finales pueden acceder a ellos para su análisis, generación de informes y minería.
* **Al añadir un área de ensayo** que se puede colocar entre las fuentes de datos y el almacén, ésta proporciona un lugar donde los datos se pueden limpiar antes de entrar en el almacén. Es posible personalizar la arquitectura del almacén para diferentes grupos dentro de la organización.
* **Se puede hacer agregando data marts**, que son sistemas diseñados para una línea de negocio en particular. Se pueden tener data marts separados para ventas, inventario y compras, por ejemplo, y los usuarios finales pueden acceder a datos de uno o de todos los data marts del departamento.