

MÓDULO XIII

INTRODUCCIÓN A LA EXPLORACIÓN DE DATOS:
CONCEPTOS
FUNDAMENTOS
CARACTERÍSTICAS

Marcela Russo
Laboratorio IV

CONCEPTOS

- **Analizar los datos** es el primer paso para explorar y visualizar datos que permite descubrir conocimientos desde el inicio o bien identificar patrones para profundizarlos.
- La exploración de datos facilita la toma de las mejores decisiones sobre dónde profundizar en los datos y tener una comprensión amplia de la organización.
- Ayuda a la familiarización con los datos, descubrir patrones y ahondar con preguntas reflexivas que estimularán a un análisis más profundo y valioso.
- El tipo de datos que se usa para fines de apoyo a la toma de decisiones es conceptualmente distinto al que se utiliza en las bases de datos de procesamiento de transacciones.
- Las bases de datos que pueden almacenar dichos datos son diferentes.
- Las arquitecturas de cómputo que pueden procesar los datos son diferentes.

FUNDAMENTOS

- El procesamiento de apoyo a las decisiones ayuda a la alta gerencia a darle dirección a una organización a mediano y largo plazos.
- El apoyo a las decisiones necesita una visión amplia que integra procesos empresariales.
- Utilizar una base de datos común para procesamiento operacional y para procesamiento de apoyo a las decisiones puede degradar en forma considerable el desempeño y hacer que sea difícil resumir la actividad en todos los procesos empresariales.

CARACTERÍSTICAS DE DATA WAREHOUSE

- El data warehouse se refiere a un depósito de datos central donde se integran, depuran y estandarizan los datos de bases de datos operacionales y otras fuentes para apoyar la toma de decisiones.
- Las actividades de transformación (depuración, integración y estandarización) son esenciales para lograr beneficios.
- Los beneficios tangibles de un data warehouse pueden incluir mayores ingresos y menores gastos permitidos por el análisis empresarial, que no eran posibles antes de la utilización de los data warehouse.
- El procesamiento de apoyo a las decisiones utiliza data warehouse con datos históricos tanto en el plano individual como en el resumido:
 - ✓ Los datos del plano individual proporcionan flexibilidad para responder a una amplia variedad de necesidades de apoyo a las decisiones.
 - ✓ Los datos resumidos dan respuesta rápida a consultas repetitivas.
- Los proyectos de data warehouse se emprenden generalmente por razones competitivas con el fin de lograr una ventaja estratégica o seguir siendo competitivos.

CARACTERÍSTICAS DE DATA WAREHOUSE

- Los requerimientos de procesamiento de las aplicaciones de apoyo a las decisiones han llevado a cuatro características distintivas para los data warehouse:
 - ✓ **Orientado a los sujetos:** Un data warehouse se organiza en torno de los principales sujetos o entidades, como clientes, pedidos y productos. Esta orientación a los sujetos contrasta con la mayor orientación hacia los procesos en el procesamiento de transacciones.
 - ✓ **Integrado:** Los datos operativos de múltiples bases de datos y fuentes de datos externas se integran en un data warehouse para proporcionar una base de datos individual y unificada para el apoyo a las decisiones. La consolidación de los datos requiere convenciones de nomenclatura consistente, formatos de datos uniformes y escalas de medición comparables en las bases de datos y en las fuentes de datos externas.

CARACTERÍSTICAS DE DATA WAREHOUSE

- ✓ Variante de tiempo: Los data warehouse usan marcas de tiempo para representar datos históricos. La dimensión del tiempo es crítica para identificar tendencias, pronosticar operaciones futuras y establecer objetivos de operación. Los data warehouse esencialmente consisten en una extensa serie de fotografías instantáneas, cada una de las cuales representa datos operativos capturados en un momento en el tiempo.
 - ✓ No volátil: Los datos nuevos en un data warehouse se anexan, en vez de reemplazarse, de modo que se preservan los datos históricos. El acto de anexar datos nuevos se conoce como refrescar el data warehouse. La falta de operaciones de actualización y eliminación garantiza que un data warehouse no contiene anomalías de actualización o eliminación. Los datos de las transacciones se transfieren a un data warehouse sólo cuando se ha completado la mayoría de las actividades de actualización.
-
- Los data warehouse mejoran la calidad de la toma de decisiones al consolidar y agregar datos de transacciones.
 - Es posible aumentar el valor de un data warehouse si se pueden descubrir patrones ocultos en los datos.
 - El modelo multidimensional de datos soporta la representación y operación de datos especialmente diseñados para el procesamiento de apoyo a las decisiones en los data warehouse

MINERÍA DE DATOS

- Se refiere al proceso de descubrimiento de patrones implícitos en datos almacenados en un data warehouse y la utilización de esos patrones como una ventaja empresarial.
- Facilita la capacidad de detectar, entender y pronosticar patrones.
- La aplicación más común de las técnicas de minería de datos es el marketing de objetivo.
- Está considerada como un anexo de un data warehouse maduro.
- Necesita datos más detallados que los que proporcionan los data warehouse tradicionales.
- Los volúmenes de datos y su dimensionalidad pueden ser mucho mayores en el caso de las técnicas de minería de datos que en los de las otras herramientas de análisis del data warehouse.
- Las técnicas de minería de datos se enriquecen con datos depurados de transacciones altamente dimensionales.
- La minería de datos requiere una serie de herramientas que se extienden más allá de las herramientas de análisis estadístico tradicional.
- Las herramientas de análisis estadístico tradicional no son adecuadas para datos muy dimensionales con una mezcla de datos numéricos y categóricos.

MINERÍA DE DATOS

- La minería de datos incluye generalmente los siguientes tipos de herramientas:
 - ✓ Herramientas de acceso a datos para extraer y muestrear datos de transacciones de acuerdo con criterios complejos para grandes bases de datos.
 - ✓ Herramientas de visualización de datos que permiten que una persona que toma decisiones adquiera una comprensión más profunda e intuitiva de los datos.
 - ✓ Una rica colección de modelos para agrupar, pronosticar y determinar reglas de asociación de grandes cantidades de datos. Los modelos implican redes neuronales, algoritmos genéticos, inducción en árboles de decisiones, algoritmos para descubrir reglas, redes de probabilidad y otras tecnologías de sistema experto.
 - ✓ Una arquitectura que proporciona optimización, procesamiento cliente-servidor y consultas paralelas para escalar a grandes cantidades de datos.