Dateneo®

Claves subrogadas

Las claves subrogadas son columnas que representan de manera única cada dimensión dentro de un modelo de datos. Esto garantiza que el modelo sea completo, no contenga duplicados y permita realizar uniones entre diferentes modelos en un almacén de datos.

Enfoques tradicionales: claves subrogadas con enteros

En sistemas de almacenamiento tradicionales, era común usar claves subrogadas con enteros monotónicamente incrementales (MIISKs). Este enfoque tenía varias ventajas:

- Relacionar entidades de manera clara e intuitiva.
- Reducir costos de almacenamiento y facilitar la indexación, haciendo más rápidas las uniones entre tablas.

Sin embargo, también presentaba desafíos:

- Mantenimiento complejo para preservar la integridad referencial.
- Dificultad para reconstruir claves en caso de pérdida de datos.
- Dependencia de ejecutar procesos específicos para generar y propagar las claves.

Claves subrogadas basadas en hash

Un enfoque alternativo es utilizar funciones de hash criptográficas para generar claves subrogadas a partir de los datos mismos. Estas funciones son determinísticas, lo que significa que producen el mismo resultado para los mismos datos de entrada. Las claves basadas en hash ofrecen beneficios como:

- Idempotencia: cada ejecución genera los mismos resultados, asegurando consistencia en los datos.
- Facilidad para manejar transformaciones paralelas sin necesidad de coordinar procesos específicos.
- Menor acoplamiento entre transformaciones, ya que las claves pueden ser regeneradas a partir de los datos originales.

Sin embargo, este enfoque también tiene desventajas:

- Potenciales colisiones (aunque son extremadamente improbables con algoritmos como MD5).
- Problemas de rendimiento en datasets masivos debido al uso de cadenas largas como claves.
- Mayores costos de almacenamiento comparados con las claves enteras.

Ejemplo: clave Subrogada para una Dimensión de Productos

id_dim_producto	id_producto_origen	nombre_producto
a1b2c3d4	1	Producto A
e5f6g7h8	2	Producto B
i9j0k1l2	3	Producto C

En este ejemplo, la 'Clave subrogada (id_dim_producto)' se genera a partir de los datos del producto (Nombre del producto, etc.), utilizando una función de hash.

Con SQL podría generarse de esta manera: `md5(concat(id_producto_origen,nombre_producto))`

El 'id_producto_origen' representa la clave original del producto.

Decisión final

La elección entre enteros monotónicamente incrementales y hashes depende de factores como:

- Tamaño de los datos y requisitos de rendimiento.
- Costos de mantenimiento y complejidad de la implementación.
- Restricciones impuestas por consumidores de datos o herramientas existentes.

Ambas estrategias tienen sus ventajas y limitaciones, y la mejor elección dependerá de las necesidades específicas del negocio y la infraestructura disponible.