Proces modelowania danych

Jest to projektowanie schematu bazy danych w hurtowni danych. Definiuje się tam strukturę danych, jakie relacje występują między nimi.

Celem tego procesu jest zapewnienie optymalnej wydajności i spójności w dostępie do danych w hurtowni danych.

W tym procesie robimy przykładowe rzeczy:

- Normalizacja danych
- Wyznaczenie kluczy głównych i obcych
- Określanie indeksów

Do procesu modelowania danych najczęściej stosuje się różne diagramy tj.:

- Diagram ER
- Diagram przepływu danych
- Diagram koncepcyjny i fizyczny,

a także język SQL.

Etapy procesu modelowania danych:

1. Analiza wymagań biznesowych (identyfikacja danych)
2. Projektowanie koncepcyjne (diagramy ER, modele koncepcyjne)
3. Projektowanie logiczne (ustawianie kluczy głównych, obcych, relacje między tabelami)
4. Projektowanie fizyczne (typy danych, partycjonowanie, indeksy, optymalizacja wydajności)
5. Implementacja

Cardinality

Jest to kardynalność. Odnosi się ona do liczby unikalnych wartości w relacji między dwoma zbiorami danych. Określa, jakie relację występują pomiędzy danymi encjami w modelu danych.

Są 3 podstawowe typy cardinality:

1. One-to-One (dla jednej wartości z jednego zbioru, odpowiada tylko jedna wartość z drugiego zbioru, np.: „Polak” i „Numer_Pesel” <- nie ma dwóch osób o takim samym Peselu)
2. One-to-Many (dla każdej wartości w jednym zbiorze może istnieć wiele odpowiadających wartości w drugim zbiorze, np.: „Autor” i „Tytuł_Książek” <- jeden autor może napisać kilka publikacji, książek”)
3. Many-to-Many (wiele wartości w jednym zbiorze, mogą być powiązane z wieloma wartościami w drugim zbiorze, np.: „Student” i „Przedmioty” <- kilku różnych studentów może być zapisanych na ten sam przedmiot, i tak samo student może zapisać się na kilka różnych przedmiotów)

Normalizacja i denormalizacja

Normalizacja – stosujemy ją aby uniknąć redundancji danych i utrzymać integralność danych. Celem jest podział danych na logicznie powiązane tabele, aby nie było powtórzeń informacji oraz by zachować spójność.

Są 3 podstawowe poziomy normalizacji:

1. Pierwsza forma normalna (1NF) – każda komórka tabeli zawiera tylko pojedynczą wartość, a nie zestaw wartości
2. Druga forma normalna (2NF) – spełnia zasady 1NF oraz przenoszone są powiązane dane do osobnych tabel
3. Trzecia forma normalna (3NF) – spełnia zasady 2NF oraz usuwane są zależności funkcjonalne między niestandardowymi, a kluczami głównymi.

Denormalizacja – odwrotność normalizacji. Polega na łączeniu logicznie powiązanych danych z kilku różnych tabel w jedną tabelę, w celu poprawy wydajności odczytu danych. Dzięki niej często unikamy wykonywania skomplikowanych operacji łączenia tabel.

Denormalizacja obejmuje:

- Łączenie danych z różnych tabel w jedną tabelę
- Dodawanie powiązanych danych jako kolumny w jednej tabeli
- Przechowywanie obliczonych wartości jako kolumny w jednej tabeli

Co to jest Datamart?

Jest to podzbiór hurtowni danych, który został zoptymalizowany pod kątem konkretnego obszaru biznesowego. Jest to mała, wyspecjalizowana baza danych, która zawiera dane z określonym obszarem funkcjonalnym. Datamart ma na celu ułatwienie dostępu do informacji i analizy danych.

Cechy Datamartu:

- Skupienie na określonym obszarze biznesowym
- Uporządkowanie i agregacja danych
- Łatwy dostęp
- Wsparcie dla raportowania

Co to jest Lakehouse i jak różni się od Hurtowni

Lakehouse to pojęcie łączące cechy tradycyjnej hurtowni danych i rozwiązań opartych na Data Lake. Przechowuje dane w oryginalnym formacie, zapewniając elastyczność i skalowalność. Jednocześnie wprowadza strukturyzowane widoki danych i mechanizmy zarządzania metadanymi. Różni się od tradycyjnej hurtowni, która przekształca dane i utracają część informacji. Lakehouse umożliwia efektywne zarządzanie i analizę danych przy minimalnej utracie informacji.