Zadanie 1

Proces modelowanie danych.
Proces modelowania danych obejmuje projektowanie struktury danych oraz relacji między nimi w celu optymalnego przechowywania i analizy informacji w hurtowni danych.
Może to obejmować takie kroki jak:

1. Analiza wymagań - zrozumienie potrzeb biznesowych i wymagań dotyczących hurtowni danych
2. Projektowanie modelu logicznego - tworzenie się model logiczny danych, który opisuje strukturę danych i ich relacje
3. Projektowanie modelu fizycznego - określa, jak dane będą przechowywane w hurtowni danych
4. Wybór schematu hurtowni danych - Wybiera się odpowiedni schemat hurtowni danych, który może obejmować schemat gwiazdy, schemat płaski lub hybrydowy
5. Implementacja modelu fizycznego - tworzenie fizycznej struktury bazy danych zgodnie z zaprojektowanym modelem fizycznym
6. ETL (Extract, Transform, Load) – przygotowanie procesów ETL, które odpowiadają za pobieranie danych z różnych źródeł, przekształcanie ich do formatu i struktury odpowiednich dla hurtowni danych oraz ładowanie ich do hurtowni danych
7. Testowanie i wdrożenie - po zakończeniu implementacji modelu fizycznego i procesów ETL przeprowadza się testy, aby sprawdzić poprawność

Cardinality
Cardinality (kardynalność) w kontekście modelowania danych odnosi się do relacji między dwoma zbiorami danych, określając liczbę elementów w jednym zbiorze, które mogą być powiązane z elementami drugiego zbioru.
Rodzaje kardynalności:

1. One-to-One (1:1): Oznacza to, że każdy element w jednym zbiorze danych jest powiązany z dokładnie jednym elementem w drugim zbiorze danych, i odwrotnie. Przykładem może być relacja między tabelami "Klient" i "Adres", gdzie każdy klient ma tylko jeden adres, a każdy adres jest przypisany do jednego klienta.

2. One-to-Many (1:N): Oznacza to, że każdy element w jednym zbiorze danych jest powiązany z wieloma elementami w drugim zbiorze danych, ale każdy element w drugim zbiorze jest powiązany tylko z jednym elementem w pierwszym zbiorze. Przykładem może być relacja między tabelami "Autor" i "Książka", gdzie jeden autor może napisać wiele książek, ale każda książka ma tylko jednego autora.

3. Many-to-Many (N:M): Oznacza to, że wiele elementów w jednym zbiorze danych jest powiązanych z wieloma elementami w drugim zbiorze danych, i vice versa. W takim przypadku wymagane jest zastosowanie tabeli pośredniej, zwanej tabelą asocjacyjną, aby śledzić te powiązania. Przykładem może być relacja między tabelami "Student" i "Kurs", gdzie jeden student może być zapisany na wiele kursów, a jeden kurs może mieć wielu studentów.

Normalizacja i denormalizacja

Normalizacja i denormalizacja są dwoma przeciwnymi procesami dotyczącymi organizacji struktury danych w bazach danych.
Normalizacja to proces projektowania struktury bazy danych w taki sposób, aby uniknąć redundancji danych i zapewnić spójność oraz efektywność operacji związanych z bazą danych. Głównym celem normalizacji jest podzielenie danych na mniejsze, logiczne jednostki, nazywane tablicami, które są zależne od kluczy głównych i mają jednoznaczną strukturę. Proces normalizacji opiera się na zasadach normalizacyjnych, takich jak reguła pierwszej postaci normalnej (1NF), reguła drugiej postaci normalnej (2NF), reguła trzeciej postaci normalnej (3NF) itd.
Przykład: W przypadku tabeli "Klient" można normalizować, dzieląc ją na dwie tabele: "Klient" i "Adres", aby uniknąć redundancji informacji o adresie, które powtarzają się dla każdego klienta.

Denormalizacja jest procesem przeciwnym do normalizacji i polega na łączeniu jednostek danych (tablic) w większe struktury, aby zoptymalizować wydajność operacji bazodanowych, takich jak zapytania czy raportowanie. Celem denormalizacji jest zwiększenie wydajności operacji poprzez minimalizację liczby operacji łączenia danych i zmniejszenie liczby tabel w bazie danych. W rezultacie dane mogą być zduplikowane i przechowywane w wielu miejscach, co zwiększa redundancję.
Przykład: W przypadku tabel "Klient" i "Zamówienie" można denormalizować, łącząc część informacji o klientach bezpośrednio z tabelą "Zamówienie". W ten sposób można uniknąć konieczności łączenia tych dwóch tabel podczas wykonywania zapytań dotyczących zamówień klienta.

Co to jest Datamart

Datamart to skoncentrowany, tematyczny zbiór danych, który jest często wykorzystywany w kontekście hurtowni danych. Jest to specjalnie zaprojektowana struktura danych, która gromadzi informacje dotyczące określonej dziedziny biznesowej lub obszaru tematycznego. Datamarty są tworzone w celu ułatwienia analizy danych i raportowania w konkretnych dziedzinach, takich jak sprzedaż, marketing, zasoby ludzkie, finanse itp.
Korzyści z użycia datamartów:

1. Uproszczenie analizy
2. Szybkość dostępu
3. Łatwość użytkowania

Co to jest Lakehouse i jak różni się od Hurtowni

Lakehouse to pojęcie, które odnosi się do nowego podejścia do przechowywania i przetwarzania danych, które łączy cechy tradycyjnej hurtowni danych i magazynu danych z elastycznością i skalowalnością technologii big data. Lakehouse łączy w sobie cechy hurtowni danych (data warehousing) i Data Lake, tworząc jednocześnie jednolitą i spójną platformę danych.
Oto kilka kluczowych różnic między Lakehouse a tradycyjną hurtownią danych:

1. Struktura danych: W tradycyjnej hurtowni danych dane są zwykle przechowywane w znormalizowanych tabelach, które są ściśle zdefiniowane i ograniczone przez schemat. W przypadku Lakehouse dane są przechowywane w formie nieprzetworzonej (raw) w Data Lake, które można opisać jako rozległe, nieustrukturyzowane i elastyczne repozytorium danych.

2. Przetwarzanie danych: W hurtowni danych przetwarzanie danych jest zwykle oparte na schemacie i strukturalnych operacjach, takich jak złączenia, agregacje i filtrowanie. Natomiast w Lakehouse przetwarzanie danych wykorzystuje technologie big data, takie jak Apache Spark, które oferują skalowalność i możliwość wykonywania operacji na danych niezależnie od ich struktury.

3. Elastyczność i skalowalność: Lakehouse zapewnia elastyczność i skalowalność, umożliwiając przechowywanie danych w różnych formatach (np. Parquet, Avro, JSON) i obsługując różne typy danych, w tym dane strukturalne i nieustrukturyzowane. Ponadto, Lakehouse wykorzystuje technologie big data, które mogą łatwo skalować się w zależności od potrzeb, umożliwiając przetwarzanie dużych ilości danych.

4. Real-time analytics: Lakehouse wspiera analizę w czasie rzeczywistym, co oznacza, że dane mogą być analizowane i raportowane na bieżąco, nie tylko w trybie wsadowym. To pozwala na szybkie reagowanie na zmiany w danych i podejmowanie decyzji w oparciu o aktualne informacje.

5. Koszty: Tradycyjne hurtownie danych często wymagają kosztownych struktur i narzędzi, aby utrzymać i zarządzać danymi. Lakehouse wykorzystuje otwarte źródła i technologie big data, które są bardziej kosztowo efektywne, co może przyczynić się do zmniejszenia kosztów wdrożenia i utrzymania infrastruktury danych.

Zadanie 2

Znajdź informację i napisz krótką notatkę co to jest kostka OLAP (OLAP CUBE, Dax).

Kostka OLAP (OLAP Cube) to wielowymiarowa struktura danych używana do analizy danych z różnych perspektyw i agregacji. Jest częścią technologii OLAP (Online Analytical Processing) i pozwala na eksplorację dużych zbiorów danych. Kostka składa się z wymiarów, które reprezentują różne aspekty danych, takie jak czas czy lokalizacja, oraz hierarchii, które organizują dane w struktury drzewiaste. Komórki w kostce przechowują wartości liczbowe lub miary. Wykorzystuje się zapytania DAX (Data Analysis Expressions) do analizy danych w kostce OLAP. Kostki OLAP są powszechnie stosowane w analizie biznesowej, raportowaniu i podejmowaniu decyzji. Umożliwiają użytkownikom eksplorację danych, odkrywanie zależności i tworzenie interaktywnych raportów.