**Zadanie 1.**

**Proces modelowania danych** polega na tworzeniu struktury, która odzwierciedla pewien obszar rzeczywistości lub problemu. Polega to na zrozumieniu wymagań biznesowych lub problemu, identyfikowaniu kluczowych informacji i projektowaniu struktury danych, która umożliwi efektywne zarządzanie nimi. Aby przeprowadzić proces modelowania danych, można zastosować następujące kroki:

* Analiza wymagań: Polega na identyfikacji i zrozumieniu potrzeb biznesowych lub problemu, który zostanie uwzględniony w modelu danych. Podczas tego etapu należy ustalić, jakie informacje są niezbędne do reprezentacji i rozwiązania problemu.
* Projektowanie konceptualne: Tworzenie wysokopoziomowego modelu danych, który przedstawia główne obiekty i ich związki. Można użyć narzędzi takich jak diagramy encji i relacji (ERD), aby zobrazować te informacje w sposób czytelny.
* Projektowanie logiczne: Tworzenie bardziej szczegółowego modelu danych, który zawiera informacje o strukturach danych, takich jak tabele, kolumny, klucze główne i obce. W tym etapie określa się także typy danych i ograniczenia.
* Implementacja fizyczna: Przekształcenie modelu logicznego w konkretne struktury danych w bazie danych. To obejmuje tworzenie tabel, indeksów, relacji i innych elementów, które są wymagane do przechowywania danych.
* Testowanie i optymalizacja: Sprawdzanie poprawności modelu danych oraz mierzenie wydajności i efektywności operacji na danych. W tym etapie przeprowadza się testy, identyfikuje i rozwiązuje ewentualne problemy, a także optymalizuje strukturę danych i zapytania.
* Utrzymanie i rozwój: Modelowanie danych to proces, który nie kończy się po wdrożeniu. Wymaga on utrzymania i dostosowania struktury danych do zmieniających się potrzeb biznesowych lub problemów. Może być konieczne rozszerzenie modelu danych wraz z rozwojem systemu lub organizacji.

**Cardinality** (kardynalność) w kontekście modelowania danych odnosi się do relacji między dwiema encjami w bazie danych i określa liczbę wystąpień jednej encji w relacji z inną encją. Wskazuje, jak wiele rekordów w jednej tabeli jest powiązanych z rekordami w innej tabeli. Kardynalność w modelowaniu danych jest ważna, ponieważ pomaga ustalić, jak tabele są ze sobą powiązane i jak informacje są reprezentowane w bazie danych. Określenie właściwej kardynalności jest istotne dla poprawnego zaprojektowania struktury danych i ustalenia ograniczeń integralności referencyjnej, które zapewniają spójność danych. Kardynalność może mieć trzy podstawowe typy:

* One-to-One: Oznacza, że każdy rekord w jednej tabeli jest powiązany tylko z jednym rekordem w drugiej tabeli,
* One-to-Many: Oznacza, że każdy rekord w jednej tabeli może być powiązany z wieloma rekordami w drugiej tabeli, ale rekordy w drugiej tabeli są powiązane tylko z jednym rekordem w pierwszej tabeli.
* Many-to-Many: Oznacza, że wiele rekordów w jednej tabeli może być powiązanych z wieloma rekordami w drugiej tabeli.

**Normalizacja** jest procesem projektowania struktury baz danych w taki sposób, aby uniknąć redundancji danych i zapewnić spójność danych. Polega na podziale danych na mniejsze, bardziej spójne i powiązane ze sobą tabele. Celem normalizacji jest minimalizowanie powtarzających się informacji i unikanie anomalii związanych z wprowadzaniem, aktualizacją i usuwaniem danych. W normalizacji stosuje się normalne formy (1NF, 2NF, 3NF itd.), które narzucają reguły dotyczące organizacji danych w tabelach.

**Denormalizacja** jest procesem wprowadzania redundancji danych do struktury baz danych w celu poprawy wydajności operacji odczytu. W odróżnieniu od normalizacji, denormalizacja jest używana w sytuacjach, gdy wymagana jest szybka i efektywna obsługa zapytań. Poprzez duplikowanie danych i tworzenie połączeń między tabelami, denormalizacja minimalizuje liczbę operacji łączenia i zwiększa szybkość dostępu do danych. Jednak denormalizacja może prowadzić do większej złożoności struktury danych i większego ryzyka utraty spójności danych w przypadku wprowadzania zmian.

**Datamart** to specjalnie zaprojektowana i zoptymalizowana część lub podzbiór hurtowni danych (data warehouse), który zawiera szczegółowe informacje o określonym obszarze biznesowym lub dziedzinie. Datamart jest tworzony w celu ułatwienia analizy i raportowania w określonym obszarze, dostarczając użytkownikom gotowe dane w łatwo przyswajalnej formie. Datamart jest zoptymalizowany pod kątem konkretnych potrzeb analitycznych, co oznacza, że dane w nim są uporządkowane i dostosowane do konkretnej struktury zapytań i analiz. Zawiera często wstępnie przeliczone miary i agregaty, które ułatwiają szybkie generowanie raportów i wykresów. Datamart może być tworzony jako samodzielna struktura lub jako część większego rozwiązania hurtowni danych.

**Lakehouse** to hybrydowa architektura, która integruje możliwości przechowywania i przetwarzania dużych ilości danych w formie surowej (raw data) charakterystycznej dla data lake oraz optymalizację i strukturalną organizację danych charakterystyczną dla data warehouse. Dane są gromadzone bez uprzedniej strukturyzacji i transformacji. Dzięki temu można zgromadzić duże ilości danych różnego typu, w różnych formatach i z różnych źródeł.

Róznice hurtowni danych i lakehouse:

* Przechowywanie danych: W hurtowni danych dane są zazwyczaj przechowywane w strukturalizowanej i przetworzonej postaci, często w formie znormalizowanej. Natomiast w lakehouse dane są przechowywane w surowej (raw) formie, bez wcześniejszej transformacji, co umożliwia gromadzenie różnorodnych danych w różnych formatach.
* Przetwarzanie danych: W hurtowni danych operacje przetwarzania i transformacji danych są zazwyczaj wykonywane przed ich załadowaniem do hurtowni. W lakehouse przetwarzanie danych może odbywać się na surowych danych bez wcześniejszej transformacji, dzięki wykorzystaniu silników przetwarzania danych, takich jak Apache Spark.
* Elastyczność: Lakehouse oferuje większą elastyczność w zakresie przechowywania i analizy różnych typów danych. Można łatwo przechowywać dane strukturalne i niestrukturalne w różnych formatach, takich jak CSV, JSON, Parquet itp. Hurtownie danych są zwykle bardziej ograniczone w zakresie obsługiwanych formatów danych.
* Aktualizacje danych: W lakehouse można łatwo dodawać, aktualizować i usuwać dane w surowej formie, co pozwala na bardziej dynamiczne zarządzanie danymi. W hurtowni danych aktualizacje danych są zazwyczaj bardziej skomplikowane i czasochłonne.
* Koszt: Implementacja lakehouse może być tańsza w porównaniu do tradycyjnej hurtowni danych, ponieważ nie wymaga tak dużego nakładu na transformację danych przed ich załadowaniem.
* Przetwarzanie w czasie rzeczywistym: Lakehouse może oferować możliwość przetwarzania danych w czasie rzeczywistym, co umożliwia analizę strumieniową i monitorowanie w czasie rzeczywistym. Hurtownie danych zazwyczaj koncentrują się na przetwarzaniu wsadowym i analizie danych historycznych.

**Zadanie 2.**

Kostka OLAP (OLAP Cube) to struktura danych wielowymiarowych, która jest wykorzystywana w technologii OLAP (Online Analytical Processing). Kostka OLAP umożliwia efektywne przechowywanie, analizę i eksplorację dużych zbiorów danych w sposób wielowymiarowy. Składa się z hierarchii wymiarów, które reprezentują różne atrybuty danych oraz miar, które są wartościami liczbowymi, podlegającymi analizie.

Kostka OLAP jest zbudowana wokół jednego lub więcej wymiarów, takich jak czas, geografia, produkt czy klient, które dostarczają kontekstu dla danych. Dzięki strukturze wielowymiarowej, użytkownicy mogą w prosty sposób analizować dane z różnych perspektyw, przeglądać sumy, zagregowane wartości, wykresy i raporty.

Aby przeprowadzić analizę w kostce OLAP, wykorzystuje się język zapytań, taki jak DAX (Data Analysis Expressions), który umożliwia formułowanie skomplikowanych zapytań, obliczenia i agregacje danych. DAX jest używany do definiowania miar, tworzenia hierarchii, wykonywania obliczeń kalkulowanych i filtrowania danych.

Kostka OLAP jest popularnym narzędziem w analizie biznesowej i wspiera takie operacje jak wiercenie (drill-down), wznoszenie (roll-up), przycinanie (slice), selekcja (dice) czy przemieszczanie (pivot). Dzięki temu użytkownicy mogą odkrywać zależności, analizować trendy, generować raporty i podejmować decyzje oparte na głębokiej analizie danych.

Kostka OLAP jest wykorzystywana w różnych dziedzinach, takich jak planowanie strategiczne, analiza sprzedaży, prognozowanie, zarządzanie zapasami, analiza finansowa i wiele innych, gdzie istnieje potrzeba kompleksowej analizy dużych zbiorów danych w kontekście różnych wymiarów i miar.