Temat1- Geneza

1. Ewolucja systemów opartych na bazach danych

Początki to np. ręczne spisy danych na papirusie w Egipcie. Ręczne zapisywanie danych trwało aż do końca XIX wieku. W XIX wieku stworzone zostały urządzenia, które dane odczytywały z kart perforowanych. Następnie w latach czterdziestych XX wieku zostały opracowane pierwsze komputery, które przechowywały dane jak i program w pamięci operacyjnej. W tym samym czasie zostały stworzone pierwsze taśmy magnetyczne, na których można było zapisać znacznie więcej danych niż na kartach perforowanych. W obecnych czasach dane przechowywane są na serwerach. Istnieją specjalne języki do posługiwania się SZBD, np. SQL.

Temat2 – Modele danych

2. Czym się rożni modelowanie od strukturalizacji danych?

Modelowanie danych określa jak dane mają być umieszczone w strukturach, i jest podstawą do strukturalizacji. Pozwala sensownie interpretować dane. Jest logicznym projektem bazy danych. Z kolei struktury zależą od używanych narzędzi i nie interpretują danych. Są częścią modelu i są sposobem na zapis danych.

- Model danych jest konstrukcją pojęciową pozwalającą w sensowny sposób interpretować dane, tzn.w taki sposób aby widzieć *informację* jaką niosą dane, a nie poszczególne wartości danych.

- Struktury (danych) - Model danych określa reguły zgodnie z którymi dane są umieszczane w strukturach. Struktury zależą od wykorzystywanych narzędzi i nie zawierają pełnej interpretacji danych oraz informacji o sposobie ich wykorzystania.

3. Jakie aspekty modelowanych danych są istotne w późniejszym procesie ich strukturalizacji?

- powinny być ***elastyczne*** czyli służyć różnym celom i realizować wymagania różnych użytkowników
- powinny być ***zrozumiałe*** czyli posługiwać się jednoznacznymi i ściśle zdefiniowanymi pojęciami
- pomagają zrozumieć, jakie dane i informacje są potrzebne organizacji

4. Znaczenie czasu w modelowaniu i strukturalizacji danych

Temat3 – Systemy przetwarzania danych

5. Składniki systemu przetwarzania danych i ich relacje

6. Charakterystyka składników przetwarzania danych

7. Procesy ETL (Extraction – Transformation – Loading)

Wypełnienie struktur Hurtowni Danych (HD) następuje na drodze złożonego procesu, który określa sięmianem ETL(Extraction – Transformation – Loading) – ekstrakcja, transformacja i ładowanie.

Ekstrakcjądanychnazywa siępobranie danych z pierwotnych ŹródełDanych(ŹD, czyli z operacyjnych systemów informacyjnych) dla potrzeb HD. Dane z ŹD od HD pobiera się
najczęściej cyklicznie. Z uwagi na zazwyczaj dużąliczbępobieranych danych czyni sięto przyrostowo, a operacja pobrania wymaga najczęściej dużego obciążenia ŹD.
Dane pobrane z ŹD ulegająnajczęściej szeregu przekształceń, transformacji i integracji przed złożeniem do końcowych struktur danych w HD. Jest to spowodowane koniecznością
ich oczyszczenia i unifikacji, a także ma na celu wykonanie niezbędnych obliczeń i porównań. Czynności przekształcania i wpasowywania danych w końcowe struktury HD
nazywa się transformacjąi ładowaniem.

8. Rodzaje architektur baz danych

[Istnieją dwa rodzaje, jednowarstwowe i dwu-lub-wielowarstwowe.

- Jednowarstwowe bazy danych to takie, z którymi użytkownik ma bezpośredni kontakt.
- Dwu lub wielowarstwowe bazy danych to takie, z którymi użytkownik porozumiewa się za pomocą pośredniczących aplikacji. Jest to dobra architektura w przypadku potrzeby typu klient-serwer.]

Jednowarstwowe bazy danych to takie, które wykonują natychmiast wszelkiego rodzaje zmiany, zaś program, który udostępnia użytkownikowi zawartość bazy ma z nim bezpośredni kontakt.

Dwuwarstwowe bazy danych - klient porozumiewa się z serwerem za pomocą specjalnych sterowników. Jeśli chodzi o samo połączenie to jest ono zależne od samego serwera, natomiast kontrolowanie poprawności danych zależy od klienta. Rozwiązanie takie wiąże się ze sporym obciążeniem programu klienckiego.

Architektura trójwarstwowa: dzieli aplikacje bazy danych na trzy, współpracujące ze sobą części:

- Warstwę dolna, realizującą dostęp do bazy danych,
- Warstwę środkową, zawierającą reguły dziedziczenia danych,
- Warstwę górna, stanowiącą interfejs użytkownika

9. Główne problemy przetwarzania numerycznego

- liczby typu INTEGER są reprezentowane dokładnie, ich arytmetyka jest dokładna, zagrożenia przy przetwarzaniu to dzielenie oraz nadmiar i niedomiar
- liczby typu REAL są reprezentowane na ogół niedokładnie, ich arytmetyka jest przybliżona

Temat 4- OLTP (operacyjnie)

10. Pojęcie OTLP (On-line transactional processing)

Online Transaction Processing (OLTP) – kategoria aplikacji klient-serwer dotyczących baz danych . Są zoptymalizowane do wykonywania bieżących czynności ewidencyjnych przy użyciu tzw. transakcji, tak aby zawierać zawsze aktualne informacje o stanie ewidencjonowanych obiektów. Transakcyjne bazy danych służą do działań operacyjnych, np. bieżącego ewidencjonowania działań, zasobów, produktów, usług, czy klientów przedsiębiorstwa.

11. Scharakteryzuj bazy operacyjne od strony technicznej

- przechowuje dane dynamiczne
- szybka modyfikacja danych
- wymaga ciągłego monitorowania, gdyż dokładne informacje mogą być cenne dla biznesu (gdzie najczęściej jest wykorzystywana)
- może przechowywać różne typy danych
- umożliwia wymianę informacji w całej firmie
- przyspiesza proces pobierania dużych ilości informacji z maksymalną wydajnością

12. Co to jest transakcja i jaka pełni role?

Transakcja – operacja lub ciąg operacji mający się wykonywać na danych z bazy (np.zapis, odczyt, wykonywanie obliczeń). Często zdarza się, że dwie lub kilka operacji musi się wykonać w określonej kolejności i niedopuszczalne jest, żeby choć jedna z nich się nie wykonała dlatego poprzez grupowanie zamyka się je w całość i dzięki temu istnieje gwarancja, że albo wszystkie operacje się wykonają albo żadna (dobrym przykładem jest przelew bankowy – nie może zdarzyć się tak, że pieniądze z jednego konta zostaną odjęte a nie zostaną dodane do drugiego)

13. Podstawowe architektury OLTP
14. Mechanizm archiwizacji danych

Temat5- OLTP (analitycznie)

15. Przedstaw koncepcję Hurtowni Danych i scharakteryzuj ich rodzaje

+++

Analityczne bazy danych OLAP (on-line analytical processing)

Modele danych:

- Relacja
- Gwiazda I płatek śniegu [relacyjmu OLAP - ROLAP]

  To typowy sposób organizacji danych dla bazy danych relacyjnej

  Obejmuje tabele faktów – olbrzymi zbiór informacji i tabele wymiarów – mniejsze informacje o obiektach, których dotyczą frakty

  Techniki ROLAP:

        - Indeksy bitmapowe – dla każdej wartości klucza indeksowego w tabeli wymiaru tworzymy wektor bitowy podający, które krotki w tabeli faktów zawierają tę wartość
        - Perspektywy zmaterializowane – w HD przechowujemy gotowe odpowiedzi na kilka użytecznych zapytań
- Kostka (tablice wielowymiarowe) [wielowymiarowy OLAP - MOLAP]

Operatory

- Slice & dice – przecinanie I rzutowanie
- Roll-up – zwijanie
- Drill-down – zmiana poziomu szczegółowości –rozwijanie
- Pivoting – obracanie –zmienia położenie wymiaru na wykresie

+++

Hurtownia danych jest to analityczna baza danych. Jej istotą jest nakierunkowanie na efektywność przetwarzania i prezentacji danych, a przedmiotem przetwarzania są duże zbiory danych (bazy operacyjne – małe zbiory). HD wykorzystują dane z systemow OLTP, a organizacja jest podporzadkowana efektywności analizy. Hurtownia danych ma wspomagac przetwarzanie informacji dla celow strategicznych i analitycznych (w przeciwienstwie do systemow transakcyjnych realizujacych przetwarzanie dla celow operacyjnych)

Rodzaje systemów OLAP:

Rodzaje:
ROLAP (relacyjny) – przechowują dane (często w postaci źródłowej) oraz tabele wymiarów w relacyjnych bazach danych. Wykonuje obliczenia na bieżąco dla przedstawienia podsumowań i wyników w wielowymiarowym formacie.

MOLAP(multiwymiarowy)- przekładają transakcje na wielowymiarowe widoki. Dane są organizowane w postaci wielowymiarowych kostek. W porównaniu do relacyjnych systemów, systemy MOLAP cechuje duża wydajność. Najbardziej istotną wadą jest możliwość przetrzymywania znacznie mniejszej ilości danych od systemów ROLAP.

[Hurtownia Danych to system informatyczny (będący tematyczną bazą danych) przetwarzających dane z różnych źródeł w spójną całość którą da się przeanalizować. Są trzy rodzaje HD, **(nawet nie wiem czy dobrą rzecz opisuję)**

- scentralizowana
- warstwowa
- federacyjna

Rodzaj (architektura) scentralizowana jest najprostszą architekturą. Upraszcza dostęp do danych. Najlepiej stosować w organizacjach o scentralizowanej strukturze, i utworzenie kilku scentralizowanych hurtowni, nie tylko jednej.

Architektura warstwowa, jak sama nazwa wskazuje, uzupełnia HD kolejnymi warstwami, podsumowującymi dane. Warto ją stosować gdy jest wiele źródeł danych które trzeba podsumować.

Architektura federacyjna oznacza aktywną współpracę kilku powiązanych HD (w jednym lub wielu systemach). Globalna hurtownia jest czymś wirtualnym, a poszczególne HD odpowiadają działom.]

16. Jaką rolę pełnią metadane w HD?

Metadane w systemie hurtowni danych mają szczególnie ważne znaczenie, gdyż opisują definicje, znaczenie, pochodzenie i identyfikują zależności danych w obrębie hurtowni danych i w powiązaniu z systemami źródłowymi.

W hurtowni danych występują dwa główne typy metadanych:

- Metadane **Biznesowe (przechowują definicje biznesowe na temat danych** np**. Nazwa Tabeli Hurtowni Danych, Nazwa Kolumny HD, Nazwa biznesowa)**

- Metadane **Techniczne (reprezentują obraz procesu ETL - mapowania i transformacje danych od systemu źródłowego do systemu docelowego,** np**. Źródłowa baza danych, Docelowa baza danych (hurtownia danych))**.

Każdy z tych typów niesie ze sobą nieco odmienne spojrzenie na dane, ale oba są niezbędne w prawidłowym wykorzystaniu systemów hurtowni danych.

[Metadane to zapisane w specjalnym repozytorium informacje o zadawanych zapytaniach w celu optymalizacji zapytań. Zawierają też wiele innych danych, np. opis logiczny danych, informacje o źródłach, informacje o aktualizacjach.]

17. Porównaj systemy OLTP i Hurtownie Danych (jako przykład OLAP)

Cecha	**OLTP**	**HD**
Czas odpowiedzi	ułamki sekundy/sekundy	Sekundy - godziny
Wykonywane operacje	DML	select

Czasowy zakres danych	30-60dni	2-10lat
Organizacja danych	Wg aplikacji	tematyczna
rozmiar	Male-duze	Duże-wielkie
Intensywność operacji dyskowych	Male-srednia	wielka

[OLTP jest systemem transakcyjnym, modyfikujący zawartość bazy danych, a OLAP to system analizujący dostarczone dane. OLTP zajmuje się świeżą, małą lub średnią ilością danych, bardzo szybko. Z kolei OLAP obsługuje gigantyczne bloki danych ze sporego zakresu czasowego, co trwa dość długo.]

18. Omów pojęcia faktów i wymiarów w kontekście systemów OLAP

Fakt - pojedyncze zdarzenie będące podstawą analiz (np. sprzedaż)
Wymiar - cecha opisująca dany fakt, pozwalająca powiązać go z innymi pojęciami modelu przedsiębiorstwa (np. klient, data, miejsce,produkt).

Stąd wynika, że każdy fakt istnieje w wielowymiarowej przestrzeni, np. fakt pojedynczej sprzedaży istnieje w wielowymiarowej przestrzeni, w której poszczególne wymiary to: czas, struktura sprzedaży, struktura klientów, struktura produktów itp.

[Fakty, inaczej miary, oraz wymiary to składniki wielowymiarowej kostki OLAP. Miały to wartości numeryczne, np. ilość sprzedanych sztuk lub koszt zakupu, mogą być sumami wszystkich wartości, pojedynczymi wartościami, lub sumami niektórych wartości w wymiarze. Wymiary to dane opisowe. Pogrupowane w poziomy, pozwalają na kontrolę szczegółowości analizy wymiaru.]

19. Wymień i omów schematy danych wykorzystywane w HD

1. Schemat gwiazda
    a. centralna tabela faktow,
    b. wymiary zdenormalizowane,
    c. tabele faktow polaczone z tabelami wymiarow przez klucze glowne/obce
2. Schemat płatka sniegu
    a. Centralna tabela faktow
    b. Wymiary znormalizowane
    c. odtwarza hierarchię wymiarów.
3. Schemat konstelacji faktow

a. Rozne tabele faktow mogą odwolywac się do roznych poziomow danego wymiaru

[Są trzy schematy danych.
- Gwiazdy – istnieje pojedyncza centralna tabela faktów, a wymiary są zdenormalizowane. Pojedyncza tabela miar/faktów jest połączona z tabelami wymiarów przez klucze główne i obce.
- Płatka śniegu – istnieje pojedyncza, centralna tabela faktów. Wymiary są znormalizowane.
- Konstelacja faktów – kombinacja wielu schematów gwiazd ze wspólnymi wymiarami. Różne tabele faktów mogą odwoływać się do różnych poziomów danego wymiaru.]

20. Omów rodzaje implementacji modelu danych w systemach OLAP

ROLAP Relacyjna implementacja modelu

1. Powiazane ze soba tabele relacyjne: tabele faktów i wymiarów
2. Schematy logiczne:
    a. Schemat gwiazdy
    b. Schemat płatka sniegu
    c. Konstelacja faktów
3. Materializowane perspektywy dla agregatów
4. Logiczny model wielowymiarowy definiowany poprzez OLAP Cata-log lub na poziomie aplikacji analitycznej

MOLAP Wielowymiarowa reprezentacja modelu

1. Dane fizycznie składowane w postaci wielowymiarowej
2. W Oracle jako analityczne przestrzenie robocze (ang. Analytic Workspaces - AW)

21. Co to jest i jak przebiega eksploracja danych?

Eksploracja danych to jedna z metod analizy danych. Umożliwia odkrywanie zależności „ukrytych w danych”, jest to proces automatycznego odkrywania dotychczas nieznanych, potencjalnie użytecznych reguł, zależności, wzorców, schematów, podobieństw lub trendów w dużych repozytoriach danych.
Celem eksploracji danych jest analiza danych i procesów dla lepszego ich zrozumienia

Odkrywane w procesie eksploracji danych wzorce mają najczęściej postać reguł logicznych, klasyfikatorów (np. drzew decyzyjnych), zbiorów skupień, wykresów, równań liniowych, itp.

Eksploracja to nie OLAP! - nie dysponujemy pełną wiedzą o przedmiocie analizy.

KROKI IMPLEMENTACJI SYSTEMU

Analiza wymagań – zgromadzenie wiedzy o wymaganiach biznesowych
W zakresie przetwarzania analitycznego
Projekt logiczny hurtowni danych - pojęciowa definicja wymaganych
Struktur danych
Implementacja struktur fizycznych hurtowni danych – tworzenie bazy danych, tabel, indeksów, materializowanych perspektyw
Implementacja oprogramowania ETL- konstrukcja modułów programo-wych służących do zasilania hurtowni danych nowymi danymi
Realizacja aplikacji analitycznych- implementacja programów dla użytkowników końcowych
Strojeniehurtownidanych-rekonfiguracjaserwerabazydanych,tworze-niedodatkowychindeksów i materializowanych perspektyw