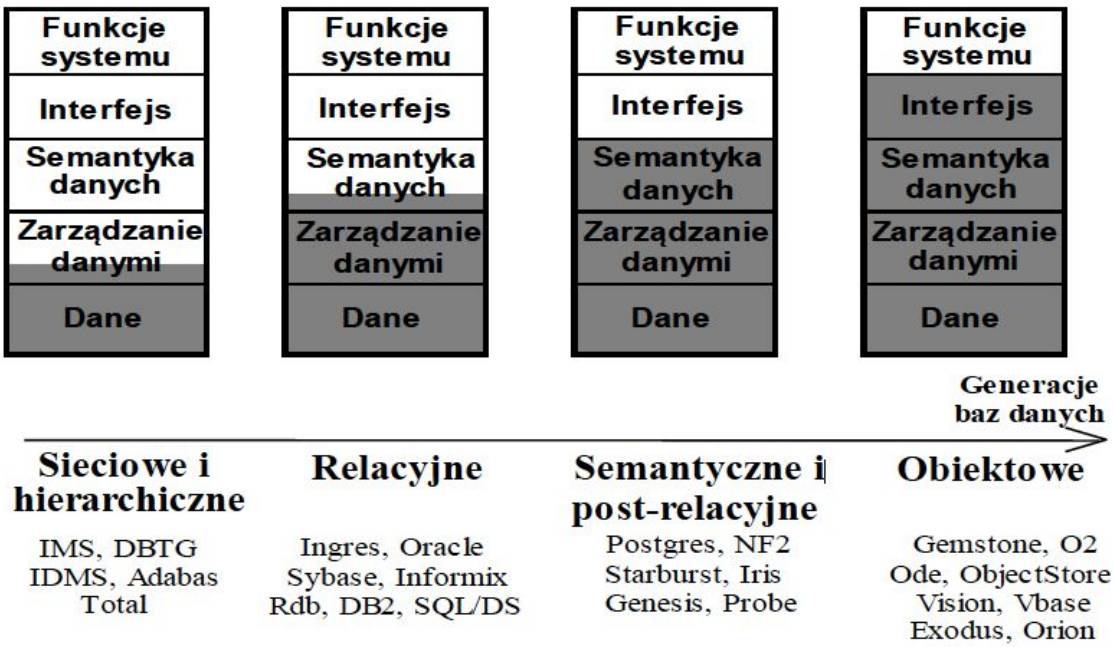
1. Ewolucja systemów opartych na bazach danych

- 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.



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. Omów znaczenia czasu w modelowaniu i strukturalizacji danych

Ilość czasu w modelowaniu i strukturalizacji danych jest wprost proporcjonalna do jakości procesu modelowania i strukturalizacji.

5. Jakie są podstawowe składniki systemu przetwarzania danych i ich wzajemna relacja?

Są to maszyny (elementy wykonujące zadane operacje), programy (dane interpretowane przez maszyny jako algorytmy strukturalne lub obiektowe), dane (jako wejściowe dane "surowe" oraz wyjściowe, przetworzone przez system) oraz ludzie (manipulacja działaniem systemu i danymi, kontrola, administracja).

6. Scharakteryzuj każdy (lub wybrany) ze składników przetwarzania danych w kontekście jego roli w systemie i wykorzystywanych metod.

Maszyny za pomocą struktur mechanicznych i logicznych wykorzystujących prawa fizyki wykonują i interpretują algorytmy. Algorytmy to oprogramowanie, które służy do obróbki i modelowania danych wejściowych w wyjściowe. Ludzie sprawują kontrolę nad systemem i mają możliwość modyfikowania danych oraz administracji procesem. Dodatkowo w literaturze czasami pojawia się dodatkowy składnik "procedury". Procedury precyzują zasady projektowania i użytkowania bazy danych.

7. Co to są procesy ETL?

Procesy ETL (Extract, Transform and Load) to narzędzia wspomagające proces pozyskania danych dla baz danych, szczególnie dla hurtowni danych (baz danych, które są zorganizowane i zoptymalizowane pod kątem pewnego wycinka rzeczywistości którego opisują). *Zadaniem narzędzi ETL jest:*

- pozyskanie danych ze źródeł zewnętrznych,
- przekształcenie danych w formę odpowiadającą strukturze bazy danych,
- załadowanie danych do bazy danych (zazwyczaj będącej hurtownią danych).

8. Scharakteryzuj rodzaje architektur baz danych.
Architektura systemów baz danych rozwijała się od tzw. architektury jednowarstwowej, w kierunku architektury wielowarstwowej.

Architektura jednowarstwowa (tzw. architektura CS, Client-Server) to architektura najprostsza. Termin architektura typu klient-serwer wywodzi się od sposobu interakcji komponentów softwarowych z systemem, mianowicie klient jest procesem który potrzebuje pewnych zasobów a proces serwera tych zasobów dostarcza. W dzisiejszych czasach praktycznie nieużywana w zastosowaniach profesjonalnych.

Architektura dwuwarstwowa wyróżnia dwie warstwy oprogramowania - warstwę serwera (proces serwera) oraz warstwę klienta (proces klienta). Lokalizacje obu procesów mogą znajdować się na jednym komputerze, choć zazwyczaj serwer umieszczany jest na innym komputerze niż procesy klienta komunikując się poprzez sieć. Na serwerze znajdują się dane i oprogramowanie zapewniające dostęp do danych (np. interpretator wywołań SQL) a po stronie klienta realizowana jest logika aplikacji. Klient odpowiada także za przetwarzanie otrzymanych danych i formę przedstawienia wyników.

Zalety: bezpieczeństwo serwera, mały ruch sieciowy, przetwarzanie danych na serwerze.
Wady: Utrudnienia związane z administracją wielu komputerów, duże koszty eksploatacji klientów, brak kontroli nad działaniami użytkowników.

Architektura trójwarstwowa dzieli aplikację bazy danych na trzy części: warstwę dolną (realizuje dostęp do bazy danych), warstwę środkową(reguły dziedziczenia danych) oraz warstwę górną (interfejs użytkownika). Warstwa dolna to prawie zawsze programy wykonywane na serwerze, obsługujące zlecenia warstwy środkowej. Warstwa środkowa może być klientem lub serwerem, natomiast górna - zawsze klientem. W arch. trójwarstwowej mamy doczynienia z tzw. "chudym (cienkim) klientem" (interfejs WWW przez przeglądarkę internetową) lub "grubym klientem" (klient realizuje dodatkowo bardziej złożone interfejsy użytkownika jak i część logiki aplikacji).

Zalety: proste projektowanie, szybka implementacja, podział na komponenty
Wady: Potrzebny silny sprzęt serwerowy, trudniejsze technologie, zwiększony ruch sieciowy.

9. Główne problemy przetwarzania numerycznego

Główne problemy przetwarzania numerycznego to potrzeba tworzenia stosunkowo skomplikowanych i wielkoobjętościowych algorytmów obliczeniowych w stosunku do małej ilości danych obliczanych tą metodą.

- 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

10. Pojęcie OTLP (On-line transactional processing)

Systemy informatyczne możemy podzielić na transakcyjne(OLTP) i analityczne (OLAP). Generalnie można przyjąć, ze systemy OLTP dostarczają danych źródłowych do hurtowni danych, natomiast systemy OLAP pomagają w ich analizie.

OLTP (On-line Transaction Processing) - kategoria aplikacji klient-serwer dotyczących baz danych. Charakteryzuje sie dużą ilością prostych transakcji zapisu i odczytu. Główny nacisk kładziony jest na zachowanie integralności danych w środowisku wielodostępowym oraz na efektywność mierzona liczba transakcji w danej jednostce czasu.
OLAP (On-line Analytical Processing) - charakteryzuje sie natomiast stosunkowo nielicznymi, ale za to złożonymi transakcjami odczytu. Miara efektywności jest czas odpowiedzi. Powszechnie wykorzystuje sie go w technikach związanych z Data Miningiem.

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 w rozumieniu baz danych i jaką pełni rolę?

Transakcja to zbiór operacji na bazie danych, które stanowią w istocie pewną całość i jako takie powinny być wykonane wszystkie lub żadna z nich. Warunki jakie powinny spełniać transakcje bardziej szczegółowo opisują zasady ACID (Atomicity, Consistency, Isolation, Durability - Atomowość, Spójność, Izolacja, Trwałość).

13. Jakie są podstawowe architektury systemów OLTP?
Są to architektury SOA (zorientowane na usługi systemy interfejsowe) oraz architektury Web.

14. Na czym polega mechanizm archiwizacji danych?
Archiwizacja danych w rozumieniu informatyki, jest to czynność wykonywania kopii danych w pamięci masowej, w celu ich długotrwałego przechowywania. Istnieje ponad 40 mechanizmów archiwizacji danych z czego najprostszym jest kopiowanie plików na nośnik inny, niż oryginalnie dane się znajdują.

15. Przedstaw koncepcję Hurtowni Danych i scharakteryzuj ich rodzaje.
Hurtownie danych są bazami danych integrującymi dane z wszystkich pozostałych systemów bazodanowych w przedsiębiorstwie. Ta integracja polega na cyklicznym zasilaniu hurtowni danymi systemów produkcyjnych (może być tych baz lub systemów dużo i mogą być rozproszone).
Architektura bazy hurtowni jest zorientowana na optymalizację szybkości wyszukiwania i jak najefektywniejszą analizę zawartości. Stąd bywa, że hurtownie danych nie są realizowane za pomocą relacyjnych baz danych, gdyż takie bazy ustępują szybkością innym rozwiązaniom.
W ramach architektury hurtowni wyróżniany jest poziom danych detalicznych oraz warstwa agregatów/kostek tematycznych.

Cele HURTONI DANYCH:
- przetwarzanie analityczne (OLAP), wspomaganie decyzji (DSS), archiwizacja danych, analiza efektywności i wsparcie dla systemów CRM (np. dobieranie strategii marketingowych na podstawie danych o klientach i sprzedaży).

Rodzaje systemów OLAP:
- **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.

16. Jaką rolę pełnią metadane w HD (HURTOWNI DANYCH)?
Opisują definicje, znaczenie, pochodzenie i identyfikują zależności danych w obrębie hurtowni danych i w powiązaniu z systemami źródłowymi.
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)).

[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/godziny
Wykonane operacje	DML	Select
Czasowy zakres danych	30-60 dni	2-10 lat
Organizacja danych	wg aplikacji	tematyczna
Rozmiar	Małe-Duże	Duże-wielkie
Intensywność operacji dyskowych	Mała-średnia	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).

Fakty i wymiary powiązane są z Kostką OLAP, która jest strukturą danych pozwalającą na szybką analizę danych. Przechowuje ona dane w sposób bardziej przypominający wielowymiarowe arkusze kalkulacyjne niż tradycyjną, relacyjną bazę danych. Można ją również zdefiniować jako zdolność manipulowania i analizowania danych z różnych punktów widzenia. Rozmieszczenie danych w kostkach pokonuje ograniczenia relacyjnych baz danych.

19. Wymień i omów schematy danych wykorzystywane w HD

- **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.

22. 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
- *Strojenie hurtowni danych* - rekonfiguracja serwera bazy danych, tworze-niedodatkowych indeksów i materializowanych perspektyw

23. Rodzaje Hurtowni Danych
- **scentralizowana** - 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.
- **warstwowa** - Architektura warstwowa, jak sama nazwa wskazuje, uzupełnia HD kolejnymi warstwami,
podsumowującymi dane. Warto ją stosować gdy jest wiele źródeł danych które trzeba podsumować.
- **federacyjna** - Architektura federacyjna oznacza aktywną współpracę kilku powiązanych HD (w jednym lub wielu
systemach). Globalna hurtownia jest czymś wirtualnym, a poszczególne HD odpowiadają działom.

24. Operatory

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