1. Omów ewolucję systemów opartych na bazach danych.

W ciągu ostatnich 60 lat bazy danych wyewoluowały z prostych systemów plików i stały się zaawansowanymi strukturami danych składującymi dane olbrzymiej liczby użytkowników dla wielu różnych aplikacji.
Zanim pojawiły się pierwsze urządzenia pozwalające na składowanie, automatyczne przeszukiwanie oraz przetwarzanie danych dane były zapisywane i przetwarzane ręcznie.
Pierwszy system zarządzania bazami danych został opracowany w latach sześćdziesiątych XXw. przez Charlesa Bachmana. Przetwarzanie danych w tym czasie było oparte na **kartach dziurkowanych i taśmach magnetycznych**. Powstały wtedy dwa kluczowe modele danych: **sieciowy** opracowany przez CODASYL oraz **hierarchiczny** opracowany przez North American Rockwell i adoptowany przez IBM.
W 1970 E. F. Codd zaproponował **relacyjny model danych**. Jednym z pierwszych implementacji modelu relacyjnego były: **Ingres** i **System R**. Pierwsze komercyjne rozwiazania takie jak **Oracle** i **DB2** nie były dostępne aż do 1980. Pierwszym udanym produktem tego typu dla mikrokomputerów był **dBASE** dla systemów operacyjnych CP/M i PC/DOS/MS-DOSC.
W latach osiemdziesiątych XX wieku aktywność badaczy skupiała się na **rozproszonych bazach danych** i **maszynach bazodanowych** oraz na **funkcyjnym modelu danych** lecz nie miały one szerszych zastosowań.
W latach dziewięćdziesiątych uwaga badaczy skupiała się na **obiektowych bazach danych**, np. przestrzenne baz danych, dane inżynieryjne i dane multimedialne. w tych latach rozprzestrzeniały się bazy danych Open Source takie jak: PostgreSQL i MySQL.
Początek XXI wieku to okres zainteresowania bazami danych XML.

2. Czym się różni modelowanie od strukturalizacji danych?

Model danych jest konstrukcją pojęciową pozwalającą w sensowny sposób interpretować dane, tzn. w jaki sposób mamy widzieć informację jaką niosą dane, a nie poszczególne wartości danych. Model danych to zbiór zasad opisujących strukturę danych w bazie danych.
Struktury danych to sposób uporządkowania informacji w komputerze. Na strukturach danych operują algorytmy. Przykłady: rekord, tablica, lista, stos...Odpowiedni wybór struktury może zmniejszyć złożoność obliczeniową, jednocześnie implementacja danej struktury może również stanowić przeszkodę.
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. Z tego powodu w modelu danych trzeba określić ogólny charakter dozwolonych operacji (wymagania). Ponieważ operacje są zazwyczaj związane ze strukturami, stąd ich precyzyjne określenie jest możliwe dopiero w powiązaniu ze strukturą.
Przykładem modelu hierarchiczneg może być struktura katalogów na dysku komputera.

3. Jakie aspekty modelowanych danych są istotne w późniejszym procesie ich strukturalizacji?
4. Omów znaczenia czasu w modelowaniu i strukturalizacji danych.
5. Jakie są podstawowoe składniki systemu przetwarzania danych i ich wzajemna relacja?
6. Scharakteryzuj każdy (lub wybrany) ze składników przetwarzania danych w kontekście jego roli w systemie i wykorzystywanych metod.

7. Co to są procesy ETL?

ETL (Extraction – Transformation – Loading) – ekstrakcja, transformacja i ładowanie. narzędzia wspomagające proces pozyskania danych dla baz danych, szczególnie dla hurtowni danych. Zadaniem narzędzi ETL jest: pozyskanie danych ze źródeł zewnętrznych, przekształcenie danych, załadowanie danych do bazy danych (zazwyczaj będącej hurtownią danych).
Ekstrakcja danych - pobranie danych z pierwotnych źrodeł danych dla potrzeb HD. Dane te pobiera się najczęściej cyklicznie, z powodu ich dużej ilości czyni się to przyrostowo, a operacja ta wymaga dużego obciążenia ŹD. Dane te ulegają szeregu przekształceń, transformacji i integracji przed złożeniem do końcowych struktur danych w HD.

8. Scharakteryzuj rodzaje architektur baz danych.

Trójwarstwowa architektura ANSI-SPARC
1.Warstwa zewnętrzna - Opisuje, jak użytkownicy widzą BD i w jaki sposób uzyskują do niej dostęp. W tej warstwie zawarte są wszystkie istotne dla jej użytkowników informacje o BD. (opisuje strukturę bazy danych z punktu widzenia wybranej grupy użytkowników)
2. Warstwa konceptualna - To zebrany, zbiorowy sposób widzenia BD. Warstwa ta opisuje, jakie dane są przechowywane w bazie i jakie są ich wzajemne związki: -wszystkie encje, ich atrybuty i związki pomiędzy nimi; - reguły spójności danych, czyli więzy; - informacje semantyczne; - informacje dotyczące bezpieczeństwa i

integralności. (opisuje strukturę bazy danych z punktu widzenia globalnego użytkownika, opisuje encje, związki między encjami, atrybuty i ograniczenia)
3.Warstwa wewnętrzna - To fizyczna reprezentacja BD w komputerze. Warstwa ta opisuje sposób przechowywania danych w bazie. Warstwa wewnętrzna jest odpowiedzialna za: - przydział przestrzeni w pamięci dla danych i indeksów; - opis struktury pamięciowej rekordów; - rozmieszczenie rekordów; - techniki kompresji i szyfrowania danych. (opisuje fizyczną organizację bazy danych, schemat wewnętrzny stosuje fizyczny model danych i opisuje ścieżki dostępu do danych oraz fizyczną organizację danych)

9. Jakie są główne problemy przetwarzania numerycznego?

Problemami przetwarzania numerycznego mogą być pojawiające się błędy takie jak: błędy pomiarowe, błędy maszynowe (błędy zaokrąglania - ograniczona precyzja prowadzi do zaokrągleń w pojedynczych kalkulacjach, efekty zaokrągleń akumulują się powoli, błędy zaokrągleń są nieuniknione, sposobem jest tworzenie lepszych algorytmów, odejmowanie niemal równych wartości prowadzi do poważnych strat precyzji, lub odrzucania w arytmetyce zmiennoprzecinkowej), błędy przybliżenia matematycznego, błąd obcięcia, błędy kasowania.

10. Wyjaśnij pojęcie OLTP.

Online Transactional Processing (OLTP) – kategoria aplikacji klient-serwer dotyczących baz danych w ramach bieżącego przetwarzania transakcji obejmujących takie zastosowania jak systemy rezerwacji, obsługa punktów sprzedaży, systemy śledzące itp. W systemach tych klient współpracuje z serwerem transakcji, zamiast z serwerem bazy danych. Kategoria aplikacji klient- serwer dotyczących baz danych w ramach bieżącego przetwarzania transakcji obejmujących takie zastosowania jak systemy rezerwacji, obsługa punktów sprzedaży, systemy śledzące itp. W systemach tych klient współpracuje z serwerem transakcji zamiast z serwerem bazy danych.

11. Scharakteryzuj bazy operacyjne od strony technicznej.

12. Co to jest transakcja w rozumieniu baz danych i jaką pełni rolę?

Transakcja - 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). Przykładem transakcji może być transakcja bankowa jaką jest przelew. W systemach bazodanowych istotne jest, aby transakcja trwała jak najkrócej. Rolą transakcji jest aktualizacja, zmienia ona stan bazy danych z jednego w drugi.

13. Jakie są podstawowe architektury systemów OLTP?

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.Archiwizacja danych obejmuje: dane tworzone i przechowywane bezpośrednio przez użytkownika komputera,pliki danych tworzone przez bazy danych,dane zapisane na wybranej partycji,dane zapisane na całym dysku twardym. Archiwizację można przeprowadzać w regularnych odstępach czasu, tym częściej im cenniejsze, ważniejsze są dane. Aby zmniejszyć objętość takich danych poddawane są onekompresji. Przy częstych archiwizacjach zapisywane mogą być tylko te dane, które zostały zmienione od czasu wykonania ostatniej archiwizacji – jest to tzw. archiwizacja przyrostowa danych. Istnieje oprogramowanie ułatwiające takie archiwizowanie oraz późniejsze odtwarzanie danych.Sposób i rodzaj archiwizacji danych jest ściśle związany z potrzebami użytkownika, systemem operacyjnym, kosztami oraz wymaganym czasem niezbędnym do jej odtworzenia oraz dostępnym oprogramowaniem. Najczęściej im mniej skomplikowana jest archiwizacja danych tym szybciej można je odtworzyć. Dlatego, dane poddane kompresji bądź podzielone na archiwa przyrostowe, mogą wydłużyć taki proces, ale jednocześnie zajmują mniej miejsca na nośniku danych i są przez to mniej kosztowne. W wypadku danych, które nie muszą być szybko odtwarzane, nośniki zawierające archiwum danych można przechowywać w innym miejscu, niż nośniki z danymi oryginalnymi.

15. Przedstaw koncepcję Hurtowni Danych i scharakteryzuj ich rodzaje.

Hurtownie danych (HD) są złożonymi systemami informatycznymi, które przetwarzają i łączą dane pochodzące z różnych źródeł w zunifikowane struktury, aby nadać im jakość i formę niezbędną dla celów analitycznych. Taka definicja wymiar jakości danych. Hurtownia danych jest wyższym szczeblem abstrakcji niż zwykła relacyjna baza danych (choć do jej tworzenia używane są także podobne technologie). W skład hurtowni wchodzązbiory danych zorientowanych tematycznie (np. hurtownia danych klientów). Dane te często pochodzą z wielu źródeł, są one zintegrowane i przeznaczone wyłącznie do odczytu. W praktyce hurtownie 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: Źródło danych – bazy danych przedsiębiorstwa, najczęściej relacyjne, Obszar przejściowy – dane pobrane z systemów źródłowych są oczyszczane i dostosowane do wymagań hurtowni danych, Warstwa prezentacji – warstwa dostępna dla użytkowników końcowych w postaci raportów i analiz, Warstwa metadanych: metadane biznesowe: tabele wymiarów, data marty, agregaty, tabele faktów oraz metadane techniczne: mapowania i transformacje danych od systemu źródłowego do systemu docelowego.

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.
metadane biznesowe - przechowują definicje biznesowe na temat danych, zawierają ogólne opisy wszystkich wartości występujących w hurtowni danych, z których korzystają użytkownicy (tabele wymiarów, data marty, agregaty, tabele faktów)
metadane techniczne - reprezentują obraz procesu ETL. Metadane te ukazują mapowania i transformacje danych od systemu źródłowego do systemu docelowego procesu ładowania. Głównie używane przez developerów hurtowni danych, specjalistów procesu ETL, analityków technicznych. (mapowania i transformacje danych od systemu źródłowego do systemu docelowego).
Każdy z tych typów niesie ze sobą nieco odmienne spojrzenie na dane, ale oba są niezbędne w prawidłowym wykorzystaniu systemów hurtowni danych.

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

18. Omów pojęcia faktów i wymiarów w kontekście systemów OLAP.

Najczęściej używa się wielowymiarowych bazy danych z uwagi na analityczny model danych o postaci kostki wielowymiarowej obejmujący: *fakty zwane też miarami, np. liczba sprzedanych samochodów; *wymiary, np. miesiące, regiony sprzedaży.
Wymiary tworzą zazwyczaj hierarchie, np. dla czasu będzie to rok-kwartał-miesiąc-dzień. Dzięki temu możliwa jest interakcyjna zmiana poziomu szczegółowości (ziarnistości) oglądanej informacji. W bardziej złożonych przypadkach hierarchie mogą rozgałęziać się, np. podział na tygodnie jest niezgodny z podziałem na miesiące.
Wymiar czasu wymaga specjalnego traktowania: *Zwykle jest ukryty — nie tworzy sie dla niego osobnej tabeli wymiaru; *Ma nature sekwencyjna: jest uporzadkowany; *Sposób agregacji po czasie zalezy od sensu miary

19. Wymień i omów schematy danych wykorzystywane w HD.

Schematy logiczne:
schemat gwiazdy-zawiera centralną tabelę faktów, ma wymiary zdenormalizowane, tabela faktów(olbrzymi zbiór faktów takich jak informacje o sprzedaży) jest połączona z tabelami wymiarów(mniejsze, statyczne informacje o obiektach, których dotyczą fakty) poprzez klucze główne i klucze obce
schemat płatków śniegu-zawiera centralną tabelę faktów, ma wymiary znormalizowane
schemat konstelacji faktów-schemat stanowiący kombinację schematów gwiazd współdzielących niektóre wymiary, różne tabele faktów mogą odwoływać się do różnych poziomów danego wymiaru

20. Omów rodzaje implementacji modelu danych w systemach OLAP.

Model wielowymiarowy moze być implementowany albo w serwerach relacyjnych (tzw. ROLAP) albo w serwerach wielowymiarowych (tzw. MOLAP). W pierwszym przypadku, dane są składowane w tabelach relacyjnych. W drugim przypadku, są one składowane w wielwymiarowych tablicach.Często w tej samej bazie danych reprezentuje się informacje częściowo w implementacji ROLAP, a częściowo w MOLAP. Taki sposób reprezentacji nazywa się hybrydowym – HOLAP (ang. Hybrid OLAP).
ROLAP: *fakty przechowywane w tabelach faktów;*wymiary przechowywane w tabelach wymiarów. MOLAP: dane przechowywane w wielowymiarowych tabelach, zwanych potocznie kostkami. HOLAP(relacyjno-wymiarowa): *dane elementarne przechowywane w tabelach, *dane zagregowane przechowywane w kostkach.

21. Co to jest i jak przebiega eksploracja danych?

Eksploracja danych - proces automatycznego odkrywania nietrywialnych, 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 danych to etap odkrywania wiedzy w bazach danych KDD (Knowledge Discovery in Databases).
Metody eksploracji danych: klasyfikacja/regresja, grupowanie/analiza skupień, odkrywanie sekwencji, odkrywanie charakterystyk, analiza przebiegów czasowych, odkrywanie asocjacji, wykrywanie zmian i odchylen, eksploracja WWW, eksploracja tekstów.