1. Przedstaw sposoby agregacji danych.

Agregacja danych to proces łączenia wielu wartości w jedną, zazwyczaj bardziej ogólną lub bardziej przetworzoną wartość. Sposoby agregacji danych to funkcje takie jak SUM, AVG, MIN, MAX, COUNT, które pozwalają na obliczenie sumy, średniej, minimum, maksimum i liczby rekordów w zestawie danych. Możesz również korzystać z własnych funkcji agregujących, jeśli potrzebujesz bardziej zaawansowanych obliczeń.

1. Omów mechanizmy łączenia danych z wielu tabel.

W SQL można łączyć dane z wielu tabel za pomocą klauzuli JOIN. Najpopularniejsze typy łączeń to INNER JOIN, LEFT JOIN, RIGHT JOIN i FULL JOIN, które pozwalają na pobieranie rekordów z jednej lub więcej tabel na podstawie wspólnych kolumn. To umożliwia łączenie i przetwarzanie danych z różnych źródeł, co jest często niezbędne w analizie danych.

1. Kiedy należy stosować funkcje działające na pojedynczych wierszach, a kiedy funkcje grupowe? Na jakich typach danych działają?

Funkcje działające na pojedynczych wierszach, takie jak funkcje matematyczne czy operatory logiczne, działają na każdym wierszu osobno. Funkcje grupowe, takie jak SUM, AVG, COUNT, działają na grupach wierszy i pozwalają na obliczenia na podstawie określonych kryteriów grupowania. Funkcje grupowe są przydatne do wykonywania obliczeń agregacyjnych, np. obliczania sumy sprzedaży na poszczególnych kategoriach produktów.

1. Omów klasyfikację funkcji działających na pojedynczych wierszach.

Funkcje działające na pojedynczych wierszach można podzielić na kilka kategorii, takie jak matematyczne, tekstowe, daty/czasu, logiczne i wiele innych. Przykłady to funkcje dodawania, mnożenia, konkatencji ciągów tekstowych, porównywania dat czy wykonywania operacji logicznych.

1. Jakie znasz polecenia zmieniające zawartość tabeli? Jakie są ich skutki oraz zakres oddziaływania?

Polecenia zmieniające zawartość tabeli to INSERT, UPDATE i DELETE. INSERT dodaje nowe rekordy, UPDATE aktualizuje istniejące rekordy, a DELETE usuwa rekordy. Skutki tych poleceń mogą być trwałe i wpływają na zawartość bazy danych. Ważne jest, aby używać ich ostrożnie, aby uniknąć utraty danych lub błędów.

1. Jaką rolę pełni Data Dictionary (Słownik Danych) i jak się nim posługiwać?

Data Dictionary to zbiór metadanych opisujących strukturę i zawartość bazy danych. Pomaga on zrozumieć, jakie tabele, kolumny, indeksy i ograniczenia istnieją w bazie danych. Słownik danych ułatwia zarządzanie bazą danych i zrozumienie jej struktury, co jest istotne w analizie danych.

1. W jakim celu buduje się perspektywy? Omów możliwe klauzule polecenia do tworzenia perspektyw.

Perspektywy to wirtualne widoki na dane, które pozwalają użytkownikom na dostęp do określonych części bazy danych bez konieczności dostępu do fizycznych tabel. Klauzule do tworzenia perspektyw obejmują SELECT, FROM, WHERE i inne, które określają, które dane będą widoczne w perspektywie.

1. Operacje na zbiorach – omów składnię poleceń i znaczenie uzyskanych wyników.

Operacje na zbiorach to zestaw operacji, takich jak UNION, INTERSECT, i EXCEPT, które pozwalają na łączenie, przecinanie i różnicowanie wyników zapytań. Składnia tych poleceń pozwala na manipulację zbiorami danych i uzyskiwanie odpowiednich wyników.

1. Przedstaw podzapytania – typy, klauzule, w których mogą wystąpić, operatory.

Podzapytania to zapytania zagnieżdżone wewnątrz innych zapytań. Mogą występować w klauzulach WHERE, SELECT lub FROM. Pozwalają na wykonywanie bardziej zaawansowanych analiz i filtrowanie danych na podstawie wyników innych zapytań.

1. Omów typowe rozwiązania Big Data w obszarze baz/repozytoriów danych.

W obszarze Big Data stosuje się różne technologie, takie jak Hadoop, Apache Spark, NoSQL bazy danych i inne, aby przechowywać i przetwarzać ogromne ilości danych. Te rozwiązania umożliwiają skalowanie i efektywne zarządzanie dużymi zbiorami danych, co jest kluczowe w analizie Big Data.

------------------------------------------------------------------------------------------------------------------------------------------

1. Przedstaw specyfikę środowisk analitycznych stosowanych w Big Data.

Specyfika środowisk analitycznych stosowanych w Big Data jest związana z przetwarzaniem ogromnych ilości danych, które nie mieszczą się w tradycyjnych bazach danych. W tym kontekście, wykorzystuje się rozproszone systemy przetwarzania danych, takie jak Hadoop, Spark czy Flink. Te platformy pozwalają na równoczesne przetwarzanie i analizę danych na wielu węzłach klastra. Środowiska te umożliwiają także integrację różnych źródeł danych, w tym dane strukturalne i nierelacyjne, co jest kluczowe w Big Data.

1. Omów wybrany algorytm stosowany w analityce Big Data.

Jednym z popularnych algorytmów stosowanych w analizie Big Data jest algorytm K-means. Ten algorytm jest używany do grupowania danych w klastry na podstawie podobieństwa między nimi. Działa na zasadzie iteracyjnego przypisywania punktów do klastrów i aktualizacji centrów klastrów. Jest stosowany w dziedzinach takich jak marketing, analiza klientów, przetwarzanie obrazów i wiele innych.

1. Na czym polega MapReduce?

MapReduce to model programowania i framework stworzony przez Google do równoległego przetwarzania danych. Polega na podziale operacji na dwie fazy: mapowania i redukcji. MapReduce jest wykorzystywany w Hadoop do przetwarzania danych na klastrze. Mapowanie polega na przekształceniu danych wejściowych w parę klucz-wartość, a redukcja na przetwarzaniu tych par w wynikowe dane.

1. Co to jest Deep Learning, podaj przykład.

Deep Learning to rodzaj uczenia maszynowego oparty na sztucznych sieciach neuronowych. Przykładem może być sieć neuronowa konwolucyjna (CNN), wykorzystywana do rozpoznawania obrazów. CNN uczy się hierarchii cech obrazów, co pozwala na skuteczne rozpoznawanie obiektów na zdjęciach.

1. Jakimi cechami charakteryzują się typowe problemy Big Data?

Typowe problemy Big Data charakteryzują się ogromnymi wolumenami danych, dużą różnorodnością źródeł i formatów danych, wysoką prędkością generowania danych oraz potrzebą przetwarzania w czasie rzeczywistym. Ponadto, często występuje potrzeba analizy danych w wielu wymiarach (wielowymiarowa analiza) oraz wykrywania wzorców i zależności w danych.

1. Omów przykładowe techniki stosowane w rozpoznawaniu wzorców.

Przykładowe techniki stosowane w rozpoznawaniu wzorców to drzewa decyzyjne, algorytmy k-means, regresja logistyczna, maszyny wektorów nośnych (SVM) i wiele innych. Te techniki pozwalają na identyfikowanie struktury i wzorców w danych, co jest istotne w analizie Big Data.

1. Na czym polega przetwarzanie rozproszone?

Przetwarzanie rozproszone to technika, w której obliczenia są wykonywane na wielu węzłach obliczeniowych, co pozwala na przyspieszenie przetwarzania danych. Jest to istotne w analizie Big Data, gdzie przetwarzanie na jednym komputerze jest niewystarczające.

1. Omów wybraną metodykę opisującą sposób realizacji procesu wytwórczego modelu analitycznego.

Metodyka realizacji procesu wytwórczego modelu analitycznego obejmuje etapy takie jak zbieranie danych, ich czyszczenie i przygotowanie, wybór i szkolenie modelu, ocenę modelu oraz jego wdrożenie. Przykładem metodyki może być CRISP-DM (Cross-Industry Standard Process for Data Mining).

1. Wymień kluczowe założenia będące warunkami zastosowania modeli predykcyjnych do wspomagania procesów decyzyjnych.

Kluczowe założenia dla zastosowania modeli predykcyjnych to jakość danych wejściowych, reprezentatywność próby, odpowiedni wybór modelu i algorytmu, oraz regularna aktualizacja modelu. Model powinien być dostosowany do konkretnego problemu biznesowego i kontekstu.

1. Jak mierzymy jakość modelu prognostycznego?

Jakość modelu prognostycznego mierzy się za pomocą miar takich jak RMSE (Root Mean Square Error), MAE (Mean Absolute Error), ROC-AUC, F1-score, czy log-likelihood. Te miary pozwalają ocenić, jak dobrze model prognozuje rzeczywistość i jakie błędy popełnia.

------------------------------------------------------------------------------------------------------------------------------------------

1. Omów w jaki sposób wykorzystanie systemu kontroli wersji wpływa na efektywność procesu wytwórczego rozwiązań analitycznych.

Wykorzystanie systemu kontroli wersji (VCS) jest kluczowe dla efektywności procesu wytwórczego rozwiązań analitycznych. VCS umożliwia śledzenie zmian w kodzie źródłowym i kontrolę nad wersjami projektu. Dzięki temu zespół analityków może współpracować nad projektem, unikając konfliktów w kodzie i utraty danych. VCS ułatwia także monitorowanie zmian, odnajdywanie błędów oraz wdrożenia. Dzięki historii zmian można również analizować, jakie czynniki wpłynęły na wyniki analizy danych. Ostatecznie, VCS przyczynia się do zwiększenia przejrzystości procesu analitycznego i pozwala na zwiększenie jakości rozwiązań analitycznych.

1. Wyjaśnij co to jest reprodukowalność procesu analitycznego i dlaczego jest ona ważna w praktyce gospodarczej.

Reprodukowalność procesu analitycznego oznacza zdolność do powtórzenia analizy na tych samych danych i uzyskania tych samych wyników. Jest to istotne w praktyce gospodarczej, ponieważ pozwala na zachowanie spójności wyników, kontrolę nad procesem i możliwość weryfikacji wyników. Reprodukowalność jest ważna zwłaszcza w naukach przyrodniczych, gdzie wyniki muszą być zweryfikowane przez innych naukowców. W biznesie pozwala na zachowanie spójności procesu analitycznego, co jest kluczowe dla podejmowania decyzji opartych na danych.

1. Omów podstawowe sposoby zapewnienia reprodukowalności procesu analitycznego.

Sposoby zapewnienia reprodukowalności procesu analitycznego obejmują dokumentację procesu, korzystanie z systemów kontroli wersji, stosowanie jasnych i przekazywalnych procedur oraz rejestrowanie wszystkich działań podczas analizy danych. Ważne jest również korzystanie z narzędzi do wirtualizacji środowiska oraz przechowywanie danych źródłowych w sposób, który umożliwia ich odtworzenie.

1. Wyjaśnij co to jest próg odcięcia w modelach klasyfikacyjnych oraz omów od czego zależy jego optymalna wartość w przypadku wykorzystania takiego modelu do wspomagania podejmowania decyzji.

Próg odcięcia w modelach klasyfikacyjnych to wartość graniczna, powyżej której obserwacje są przyporządkowane do jednej klasy, a poniżej do drugiej. Optymalna wartość progu odcięcia zależy od kontekstu biznesowego i priorytetów. Wyższy próg oznacza bardziej rygorystyczny model, co może zmniejszyć fałszywe pozytywy, ale zwiększyć fałszywe negatywy. Niższy próg może zwiększyć odsetek fałszywych pozytywów, ale zmniejszyć fałszywe negatywy. Optymalny próg odcięcia powinien być dostosowany do wymagań biznesowych i kosztów błędów.

1. Wyjaśnij do czego wykorzystywana jest regularyzacja w procesie budowy modeli predykcyjnych.

Regularyzacja w procesie budowy modeli predykcyjnych to technika, która pomaga zapobiec przeuczeniu się modelu. Polega na dodaniu do funkcji celu (funkcji straty) kary za zbyt skomplikowany model. Regularyzacja ma na celu zminimalizowanie współczynnika przystosowania modelu do danych treningowych kosztem jego zdolności do generalizacji na danych testowych. Dzięki temu unika się nadmiernego dopasowania modelu do szumu w danych i poprawia się jego zdolność do przewidywania na nowych danych.

1. Wyjaśnij różnicę pomiędzy wnioskowaniem obserwacyjnym, interwencyjnym i kontrfaktycznym.

Wnioskowanie obserwacyjne polega na analizie danych obserwowanych w naturalnych warunkach. Wnioskowanie interwencyjne to wnioskowanie oparte na działaniach i eksperymentach, a kontrfaktyczne opiera się na badaniu tego, co by się stało, gdyby pewne działania nie zostały podjęte. Różnice między nimi wynikają z różnych poziomów ingerencji w badane zjawisko i rodzaju pytań, jakie można na ich podstawie postawić.

1. Wyjaśnij na czym polega paradoks Simpsona.

Paradoks Simpsona to zjawisko, w którym relacje statystyczne między zmiennymi zmieniają się w zależności od kontekstu lub grupy danych. Przykładem może być sytuacja, w której w jednej grupie dane sugerują związek między zmiennymi A i B, a w innej grupie brak tego związku, chociaż ogólny trend jest przeciwny. Paradoks ten podkreśla znaczenie uwzględniania kontekstu i ostrożności w analizach statystycznych.

1. Przedstaw korzyści ekonomiczne z przetwarzania danych w chmurze.

Przetwarzanie danych w chmurze oferuje wiele korzyści ekonomicznych, takich jak elastyczność, skalowalność, niskie koszty początkowe i brak konieczności inwestowania w sprzęt. Firmy mogą dostosować zasoby do bieżących potrzeb, co prowadzi do oszczędności kosztów i lepszej wydajności. Ponadto, korzystanie z usług chmurowych pozwala na dostęp do zaawansowanych narzędzi i technologii bez konieczności ich zakupu i utrzymania.

1. Omów technologie serveless w gromadzeniu i przetwarzaniu danych na potrzeby procesów analitycznych.

Technologie serverless w gromadzeniu i przetwarzaniu danych pozwalają na zarządzanie zasobami na poziomie funkcji, co eliminuje konieczność zarządzania infrastrukturą. Serverless oferuje skalowalność, oszczędność kosztów i efektywność, gdy potrzebne są dynamiczne obliczenia i przetwarzanie danych w czasie rzeczywistym.

1. Przedstaw metody przechowywania danych dużych rozmiarów w chmurze.

Istnieje wiele metod przechowywania danych dużych rozmiarów w chmurze, w tym bazy danych NoSQL, przechowywanie obiektowe, systemy plików i usługi magazynowania danych. Wybór odpowiedniej metody zależy od konkretnych potrzeb, rodzaju danych i wymagań wydajnościowych. Przechowywanie w chmurze oferuje skalowalność, odporność na awarie

------------------------------------------------------------------------------------------------------------------------------------------

1. Omów skalowanie dokumentowych baz danych typu noSQL w chmurze na przykładzie DynamoDB.

Skalowanie dokumentowych baz danych typu NoSQL w chmurze, na przykładzie DynamoDB, polega na elastycznym dostosowaniu zasobów bazy danych do bieżących potrzeb. DynamoDB to usługa bazodanowa w chmurze dostępna w Amazon Web Services (AWS), która oferuje automatyczne skalowanie poziome oraz zarządzanie niezawodnością. Dzięki funkcjonalności auto-skali, DynamoDB może dostosować się do wzrostu obciążenia i zapewnić ciągłą dostępność danych. Skalowanie polega na zmianie pojemności odczytu/zapisu oraz replikacji w bazie danych w zależności od wymagań. Jest to przydatne w przypadku aplikacji o zmiennym obciążeniu lub z okresowymi wzrostami ruchu.

1. Omów skalowanie procesów analitycznych w chmurze.

Skalowanie procesów analitycznych w chmurze jest procesem dostosowywania mocy obliczeniowej do potrzeb analizy danych. Chmura obliczeniowa, takie jak AWS, umożliwia dynamiczne przydzielanie zasobów obliczeniowych, co pozwala na równoczesne przetwarzanie dużych zbiorów danych. Przykłady to usługi EC2, EMR i Redshift w AWS. Dzięki temu możesz wykonywać zaawansowane analizy, takie jak analiza Big Data, bez konieczności inwestowania w stały sprzęt. Możesz również dostosować zasoby do bieżących zadań, co jest wydajne i kosztoszczędne.

1. Omów Function as a service – model przetwarzania oparty o architekturę Lambda.

Function as a Service (FaaS) to model przetwarzania w chmurze, oparty na architekturze Lambda. Lambda pozwala na wykonywanie kodu w odpowiedzi na zdarzenia, bez konieczności zarządzania infrastrukturą. To elastyczny sposób tworzenia mikrousług i aplikacji serwerless, co znacznie upraszcza rozwijanie i wdrażanie aplikacji. Lambda umożliwia także integrację z innymi usługami chmurowymi, co jest przydatne w przypadku analizy danych w czasie rzeczywistym.

1. Omów tworzenie i zarządzanie bezpieczeństwem środowisk analitycznych dla języków Python i R w chmurze.

Tworzenie i zarządzanie bezpieczeństwem środowisk analitycznych dla języków Python i R w chmurze obejmuje różne aspekty, takie jak kontrola dostępu, szyfrowanie danych i monitorowanie. Chmura obliczeniowa, takie jak AWS, oferuje narzędzia do zarządzania bezpieczeństwem, takie jak IAM (Identity and Access Management), które pozwalają na definiowanie uprawnień dostępu. Ponadto, usługi, takie jak Amazon S3, pozwalają na szyfrowanie danych w spoczynku i w transmisji. Monitoring i audytowanie to kolejne kluczowe elementy, które pomagają w śledzeniu dostępu i wykrywaniu ewentualnych naruszeń bezpieczeństwa.

1. Omów zarządzanie bezpieczeństwem, użytkownikami i prawami dostępu w chmurze – użytkownicy, role, polityki i grupy.

Zarządzanie bezpieczeństwem, użytkownikami i prawami dostępu w chmurze obejmuje konfigurację identyfikacji i uwierzytelniania użytkowników, przypisywanie ról, tworzenie polityk dostępu oraz grup użytkowników. Przykładem jest usługa IAM w AWS, która umożliwia tworzenie i zarządzanie użytkownikami, grupami i rolami, a także definiowanie precyzyjnych polityk dostępu do zasobów. To kluczowy aspekt zarządzania infrastrukturą w chmurze.

1. Przedstaw systemy zarządzania relacyjną bazą danych w chmurze i ich zastosowania w analityce danych.

Systemy zarządzania relacyjną bazą danych w chmurze, takie jak Amazon RDS, pozwalają na przechowywanie, zarządzanie i skalowanie danych relacyjnych w środowisku chmurowym. Są one przydatne w analizie danych, gdy konieczne jest zachowanie integralności danych i stosowanie zapytań SQL. Przykłady zastosowań to analiza transakcyjna, raportowanie i business intelligence.

1. Przedstaw modele przetwarzania danych w chmurze: IaaS (Infrastructure-as-aService), PaaS (Platform-as-a-Service) oraz SaaS (Software-as-a-Service).

Modele przetwarzania danych w chmurze to IaaS (Infrastructure-as-a-Service), PaaS (Platform-as-a-Service) i SaaS (Software-as-a-Service). IaaS oferuje dostęp do infrastruktury obliczeniowej, takiej jak wirtualne maszyny. PaaS to środowisko do budowania aplikacji, a SaaS to gotowe aplikacje dostępne w chmurze. Modele te różnią się poziomem kontroli nad infrastrukturą i aplikacjami oraz elastycznością.

1. Omów kwestie etyczne związane z Big Data.

Kwestie etyczne związane z Big Data dotyczą prywatności, ochrony danych, uczciwego przetwarzania i konsekwencji analizy danych. W analizie danych należy zachować zgodność z przepisami ochrony danych osobowych i dbać o transparentność oraz zgodność z zasadami etycznymi, aby uniknąć nadużyć.

1. Omów cechy danych istotne w procesie analizy danych.

Cechy danych istotne w procesie analizy danych obejmują jakość, kompletność, spójność, dostępność, aktualność i odporność na błędy. Dobre jakościowo dane są kluczowe dla uzyskania rzetelnych wyników analizy.

1. Przedstaw, na czym polega zmienność danych i jak ją uwzględnić w wizualizacji danych.

Zmienność danych to ich różnorodność i dynamiczność w czasie. Wizualizacja danych powinna uwzględniać zmienność, stosując odpowiednie techniki, takie jak animacje, interaktywność i wykresy dynamiczne, aby lepiej zrozumieć zmiany w danych.

------------------------------------------------------------------------------------------------------------------------------------------

1. Przedstaw, na czym polega niepewność w analizie danych i jak można wpływać na jej wielkość.

Niepewność w analizie danych polega na stopniu niepewności lub błędów związanymi z danymi, które mogą wpłynąć na wyniki analizy. Wpływa na nią wiele czynników, takich jak jakość danych, metody zbierania danych, brak danych, błędy pomiarowe itp. Aby zmniejszyć niepewność, można stosować różne techniki, takie jak normalizacja danych, usuwanie wartości odstających, uzupełnianie brakujących danych, czy też stosowanie statystycznych metod estymacji. Ważne jest także prowadzenie badań w sposób powtarzalny, dzięki czemu inni badacze mogą zweryfikować wyniki.

1. Jakie znaczenie mają metadane w analizie danych.

Metadane w analizie danych to informacje o danych, które opisują ich strukturę, znaczenie i sposób gromadzenia. Są kluczowe, ponieważ pozwalają zrozumieć i interpretować dane. Przykłady metadanych to nazwy kolumn, jednostki miary, opisy, daty zbierania danych itp. Metadane ułatwiają również zarządzanie danymi, przeszukiwanie i identyfikowanie odpowiednich zbiorów danych.

1. Wymień i omów układy współrzędnych stosowane przy wizualizacji danych.

Wizualizacja danych wykorzystuje różne układy współrzędnych, takie jak:

a. Układ kartezjański: Wykorzystuje oś X i oś Y do przedstawienia danych numerycznych na płaszczyźnie.

b. Układ biegunowy: Przedstawia dane wokół punktu centralnego, używając promienia i kąta.

c. Układ czasowy: Ukazuje dane w zależności od czasu na osi X.

d. Układ geograficzny: Wykorzystuje długość i szerokość geograficzną do przedstawienia danych na mapie.

44. Wymień i omów metody wizualizacji proporcji.

Metody wizualizacji proporcji służą do przedstawiania stosunków procentowych i udziałów. Przykłady to wykres kołowy, wykres słupkowy, wykres pudełkowy itp.

1. Wymień i omów metody wizualizacji relacji.

Metody wizualizacji relacji pozwalają na przedstawienie związków między różnymi zmiennymi. Przykłady to wykres punktowy, wykres siatki, mapa cieplna itp.

1. Wymień i omów metody wizualizacji danych geolokalizacyjnych.

Metody wizualizacji danych geolokalizacyjnych wykorzystują dane związane z lokalizacją. Przykłady to mapy, mapy cieplne, wykresy rozkładu punktów na mapie itp.

1. Wymień obiekty bazy danych i omów ich przeznaczenie.

Obiekty bazy danych to tabele, indeksy, widoki, procedury przechowywane itp. Służą do przechowywania, zarządzania i dostępu do danych. Tablice przechowują faktyczne dane, indeksy przyspieszają wyszukiwanie, a widoki tworzą wirtualne tabele na podstawie istniejących danych.

1. Wymień i omów metody wizualizacji szeregów czasowych.

Metody wizualizacji szeregów czasowych obejmują wykresy liniowe, wykresy słupkowe, wykresy punktowe. Pozwalają na analizę zmian w czasie i identyfikację trendów.

1. Przedstaw, na czym polega uwzględnienie kontekstu a analizie danych.

Uwzględnienie kontekstu w analizie danych oznacza branie pod uwagę informacji otaczających dane, takich jak okoliczności, warunki, czynniki wpływające na wyniki. Kontekst pomaga zrozumieć, dlaczego pewne zdarzenia miały miejsce i pozwala na bardziej kompleksową analizę

1. Wyjaśnij co to jest system kontroli wersji na przykładzie systemu Git i zaproponuj typowy workflow.

System kontroli wersji, tak jak Git, służy do zarządzania historią zmian w kodzie źródłowym i dokumentach. Typowy workflow w Git obejmuje tworzenie repozytorium, dodawanie zmian do repozytorium, commitowanie zmian, tworzenie gałęzi, łączenie zmian (merge), rozwiązywanie konfliktów i publikację zmian na serwerze zdalnym. Systemy kontroli wersji ułatwiają współpracę, śledzenie zmian i przywracanie wcześniejszych wersji kodu.

------------------------------------------------------------------------------------------------------------------------------------------

1. Omów wybraną technikę redukcji wymiaru danych, jej zalety i wady.

Technika redukcji wymiaru danych, tak jak np. analiza głównych składowych (PCA), jest wykorzystywana do zmniejszenia ilości atrybutów w zbiorze danych, jednocześnie zachowując jak najwięcej informacji. Główne zalety to zmniejszenie złożoności danych, co może poprawić wydajność algorytmów analizy danych i zmniejszyć ryzyko overfittingu. Jednakże, redukcja wymiaru może prowadzić do utraty informacji istotnych dla analizy. Istnieje także ryzyko wprowadzenia błędów, jeśli nie dokładnie przemyślimy procesu redukcji

1. Omów pojęcie obliczeń równoległych i podstawowe problemy, które pojawiają się przy obliczeniach równoległych.

Obliczenia równoległe to podejście, w którym wiele obliczeń wykonywanych jest jednocześnie, co pozwala na przyspieszenie przetwarzania danych. Problemy związane z obliczeniami równoległymi obejmują konieczność synchronizacji procesów, zarządzania dostępem do zasobów wspólnych i zapewnienie spójności wyników.

1. Omów pojęcie estymatora odpornego na wybranych przykładzie.

Estymator odporny, np. mediana, jest odporny na obserwacje odstające w danych. Estymatory te nie są wrażliwe na skrajne wartości, co może być korzystne w analizie danych, gdzie obecność outlierów jest częsta.

1. Omów technikę regularyzacji na wybranym przykładzie, np. regresji LASSO.

Regularyzacja, np. regresja LASSO, to technika służąca do kontrolowania overfittingu w modelach predykcyjnych. Poprzez wprowadzenie kary za duże współczynniki, regularyzacja pomaga w selekcji istotnych zmiennych i poprawie ogólnej wydajności modelu.

1. Co oznacza określenie 3V oraz 5V w kontekście problematyki Big Data?

W kontekście Big Data, 3V oznacza duży wolumen, dużą różnorodność i dużą prędkość przetwarzania danych. 5V dodaje do tego wartość i ważność danych. Oznacza to, że Big Data charakteryzują się dużą ilością, różnorodnością, szybkim tempem i znaczeniem danych.

1. Wyjaśnij pojęcia danych ustrukturyzowanych i nieustrukturyzowanych.

Dane ustrukturyzowane mają klarowną strukturę, taką jak bazy danych. Dane nieustrukturyzowane nie posiadają ustalonej struktury, takie jak tekst, obrazy czy multimedia.

1. Przedstaw architektury: Lambda i Kappa.

Architektura Lambda łączy przetwarzanie wsadowe (batch) i strumieniowe (streaming) w celu analizy danych w czasie rzeczywistym. Kappa to architektura oparta na przetwarzaniu strumieniowym, która pomaga w rozwiązywaniu problemów przetwarzania danych w czasie rzeczywistym.

1. Przedstaw kluczowe cechy uczenia i predykcji w trybie wsadowym (offline learning) i przyrostowym (online learning).

Uczenie w trybie wsadowym to proces trenowania modelu na całym dostępnym zbiorze danych. Uczenie przyrostowe odświeża model na bieżąco w miarę napływania nowych danych. Kluczowe cechy obejmują wydajność, możliwość dostosowywania się do zmieniających się danych i koszty obliczeniowe.

1. Podaj przykład i omów w jakich sytuacjach wskazane jest zastosowanie modelu przetwarzania OLTP.

Przykładem zastosowania modelu przetwarzania OLTP (Online Transaction Processing) jest system obsługi zamówień w sklepie internetowym. Ten model zapewnia obsługę transakcji na żywo, takie jak przetwarzanie zamówień i płatności.

1. Podaj przykład i omów w jakich sytuacjach wskazane jest zastosowanie modelu przetwarzania OLAP.

Przykładem zastosowania modelu przetwarzania OLAP (Online Analytical Processing) jest system do analizy i raportowania danych biznesowych. OLAP umożliwia interaktywną analizę danych, generowanie raportów i odkrywanie wzorców w danych.

------------------------------------------------------------------------------------------------------------------------------------------

1. Wyjaśnij pojęcie i zastosowania biznesowe hurtowni danych.

Hurtownia danych to centralne repozytorium, w którym gromadzone są dane z różnych źródeł w celu ułatwienia ich analizy i wykorzystania w procesach biznesowych. Biznesowe zastosowania hurtowni danych obejmują raportowanie, analizę, generowanie wizualizacji danych, wspieranie procesów podejmowania decyzji oraz prognozowanie. Hurtownie danych umożliwiają firmom efektywne zarządzanie danymi, poprawę jakości informacji, a także tworzenie bardziej spójnych i dostępnych źródeł danych dla pracowników.

1. Omów problem czasu w strumieniowym przetwarzaniu danych, czym jest Watermark.

Problem czasu w strumieniowym przetwarzaniu danych polega na konieczności analizy danych w czasie rzeczywistym, co jest istotne w przypadku takich dziedzin jak telekomunikacja, medycyna czy finanse. Watermark jest znacznikiem czasowym, który oznacza moment pojawienia się danej informacji w strumieniu. Pomaga on w kontrolowaniu opóźnień w przetwarzaniu i rozumieniu, jak świeże są dane.

1. Przedstaw różnicę pomiędzy wsadowym i strumieniowym sposobem przetwarzania danych.

Różnica pomiędzy wsadowym a strumieniowym przetwarzaniem danych polega na podejściu czasowym. Wsadowe przetwarzanie opiera się na analizie danych w partiach, podczas gdy strumieniowe przetwarzanie to ciągłe analizowanie napływających danych w czasie rzeczywistym.

1. Opisz dwa biznesowe zastosowania analizy danych w czasie rzeczywistym.

Analiza danych w czasie rzeczywistym ma wiele zastosowań biznesowych, takich jak monitorowanie zachowań klientów w sklepach online, wykrywanie oszustw transakcyjnych w bankowości, zarządzanie łańcuchem dostaw, monitorowanie zdrowia pacjentów w medycynie i wiele innych.

1. Wymień i omów metodyki procesu eksploracji danych.

Metodyki procesu eksploracji danych to kroki i techniki wykorzystywane do odkrywania wzorców i wiedzy z danych. Przykłady to CRISP-DM, KDD, SEMMA, i inne.

1. Omów dwie główne grupy metod eksploracji danych.

Dwie główne grupy metod eksploracji danych to eksploracja statystyczna (oparta na analizie statystycznej) i eksploracja oparta na uczeniu maszynowym (wykorzystująca algorytmy uczenia maszynowego do wydobywania wiedzy).

1. Omów metody selekcji zmiennych i obserwacji do modelowania data mining.

Metody selekcji zmiennych i obserwacji pomagają wybrać najważniejsze dane do modelowania w procesie data mining. Przykłady to analiza wariancji, metody filtrowe, a także techniki oparte na algorytmach uczenia maszynowego

1. Metody klasyfikacji danych – przedstaw różnice i podobieństwa pomiędzy nimi.

Metody klasyfikacji danych, takie jak drzewa decyzyjne, regresja logistyczna, i maszyny wektorów nośnych, różnią się pod względem zastosowań i podejścia, ale wszystkie służą do przewidywania przynależności obiektów do określonych klas na podstawie dostępnych danych.

1. Przedstaw model drzewa decyzyjnego.

Model drzewa decyzyjnego to graficzna reprezentacja podejmowania decyzji. Drzewo składa się z węzłów reprezentujących testy na zmiennych i liści reprezentujących przewidywane klasy obiektów.

1. Omów modele lasów losowych.

Modele lasów losowych to zbiory drzew decyzyjnych, które działają razem w celu poprawienia dokładności klasyfikacji i redukcji ryzyka overfittingu. Są one wykorzystywane w zadaniach klasyfikacji i regresji, a ich główną zaletą jest zdolność do radzenia sobie z dużą ilością danych i zmiennych.

------------------------------------------------------------------------------------------------------------------------------------------

1. Przedstaw modele sztucznych sieci neuronowych.

Modele sztucznych sieci neuronowych to rodzaj algorytmów wykorzystywanych w uczeniu maszynowym. Próbują one naśladować działanie ludzkiego mózgu, składając się z neuronów i połączeń między nimi. Popularne modele to np. jednokierunkowe sieci neuronowe (feedforward neural networks), które składają się z warstw neuronów, sieci rekurencyjne (recurrent neural networks), które uwzględniają poprzednie stany, oraz głębokie sieci neuronowe (deep neural networks), które mają wiele warstw ukrytych.

1. Omów metody grupowania danych.

Metody grupowania danych to techniki wyodrębniania podobnych obserwacji z zestawu danych. Przykłady to k-means, hierarchiczne grupowanie, DBSCAN i wiele innych. Grupowanie pomaga w odkrywaniu struktur w danych i jest często stosowane w analizie Big Data, na przykład do segmentacji klientów.

1. Omów metody analizy danych transakcyjnych.

Analiza danych transakcyjnych to proces wydobywania wiedzy z danych zawierających transakcje, np. dane o zakupach klientów. Metody te mogą zawierać statystyczne analizy, jak analiza koszyka zakupowego, czy techniki oparte na uczeniu maszynowym do przewidywania zachowań klientów.

1. Przedstaw metody estymacji parametrów modelu regresji logistycznej.

Metody estymacji parametrów modelu regresji logistycznej to techniki używane w statystyce i analizie danych do dopasowania modelu regresji logistycznej do danych obserwacyjnych. Przykładowe metody to metoda największej wiarygodności i metoda najmniejszych kwadratów.

1. Interpretacja wyników oszacowań parametrów modelu regresji logistycznej.

Interpretacja wyników modelu regresji logistycznej polega na analizie oszacowanych parametrów, takich jak współczynniki regresji. Może to obejmować analizę wpływu zmiennych niezależnych na prawdopodobieństwo sukcesu (lub inny wynik) oraz ocenę istotności tych zmiennych.

1. Weryfikacja istotności oszacowań parametrów regresji logistycznej.

Weryfikacja istotności oszacowań parametrów regresji logistycznej zwykle polega na obliczeniu statystyk testowych, takich jak statystyka Wald'a lub testu zgodności. Wartości te pomagają określić, czy dany parametr jest istotny statystycznie.

1. Metody oceny dopasowania modelu regresji logistycznej do danych empirycznych.

Ocena dopasowania modelu regresji logistycznej do danych obejmuje zastosowanie miar takich jak ROC-AUC, krzywe ROC, krzywe Precision-Recall oraz statystyk takich jak deviance. Pomagają one określić jakość dopasowania modelu.

1. Metody identyfikacji obserwacji odstających i wpływowych w regresji logistycznej.

Do identyfikacji obserwacji odstających i wpływowych w regresji logistycznej można wykorzystać statystyki Cooka i DFBETA. Obserwacje, które znacząco wpływają na wyniki modelu, mogą być uważane za potencjalne obserwacje odstające.

1. Omów model wielomianowej regresji logistycznej.

Model wielomianowej regresji logistycznej to rozszerzenie standardowego modelu regresji logistycznej, w którym używane są wyższe potęgi zmiennych niezależnych. To pozwala uwzględnić nieliniowe zależności między zmiennymi a wynikiem.

1. Omów model proporcjonalnych szans.

Model proporcjonalnych szans (ang. proportional odds model) to rodzaj modelu regresji logistycznej, który jest używany do analizy wielu poziomów uporządkowanej zmiennej zależnej. Model ten zakłada stały stosunek szans pomiędzy różnymi poziomami zmiennej zależnej. Jest szczególnie przydatny w kontekście analizy danych uporządkowanych, takich jak skale ocen.

------------------------------------------------------------------------------------------------------------------------------------------

1. Metody doboru zmiennych objaśniających w modelach regresji.

W modelach regresji istotne jest wybranie odpowiednich zmiennych objaśniających, aby uzyskać dokładne prognozy. Istnieje wiele metod doboru zmiennych, takie jak selekcja krokowa, analiza składowych głównych (PCA), las losowy czy LASSO. Selekcja krokowa polega na iteracyjnym dodawaniu lub usuwaniu zmiennych w modelu na podstawie ich istotności statystycznej. PCA redukuje wielowymiarowe dane do mniejszej liczby składowych, co może pomóc w identyfikacji istotnych zmiennych. Las losowy ocenia wpływ każdej zmiennej na wynik i pozwala na ich ranking. LASSO dodaje karę za dużą liczbę zmiennych, co promuje prostsze modele.

1. Jakość danych w analizach biznesowych. Znaczenie i metody oceny.

Jakość danych odgrywa kluczową rolę w analizach biznesowych. Nieprawidłowe, brakujące lub niekompletne dane mogą prowadzić do błędnych wniosków. Istnieją różne metody oceny jakości danych, takie jak ocena spójności, kompletności, dokładności i aktualności danych. Techniki takie jak profilowanie danych i analiza jakości danych pomagają zidentyfikować problemy. Ważne jest również utrzymanie danych w dobrej jakości przez stałą kontrolę i zarządzanie nimi.

1. Imputacja danych. Istota i znaczenie.

Imputacja danych to proces zastępowania brakujących wartości w danych. Jest istotna, ponieważ brakujące dane mogą zakłócić analizy i prognozy. Istnieją różne metody imputacji, takie jak imputacja medianą, średnią, regresją czy metodami bazującymi na algorytmach. Imputacja powinna być wykonywana ostrożnie, aby nie wprowadzać błędów ani zniekształcać wyników.

1. Imputacja wielokrotna: opis metody, wybór modelu do imputacji oraz estymacja parametrów.

Imputacja wielokrotna to zaawansowana technika imputacji danych. Polega na wielokrotnym losowym uzupełnianiu brakujących danych, co pozwala na uwzględnienie niepewności. Wybór modelu do imputacji może zależeć od charakterystyki danych, a estymacja parametrów może być dokonywana na podstawie różnych prób.

1. Porównaj modele z efektami stałymi oraz modele z efektami losowymi. Przedstaw podstawowe różnice i zastosowania obu typów modeli.

Modele z efektami stałymi zakładają, że wpływ zmiennych niezależnych jest taki sam dla wszystkich obserwacji, podczas gdy modele z efektami losowymi uwzględniają zmienność między różnymi grupami obserwacji. Modele z efektami stałymi są prostsze, ale mogą pomijać ważne źródła zmienności. Modele z efektami losowymi są bardziej elastyczne i mogą uwzględniać efekty grupowe.

1. Regresja kwantylowa: opis i zastosowania w analityce biznesowej.

Regresja kwantylowa to technika modelowania, która pozwala analizować różnice w kwantylach rozkładu zmiennych niezależnych. Jest przydatna w analizie danych biznesowych, gdzie istotne są różnice w dolnych lub górnych częściach rozkładu.

1. Regresja adaptacyjna: model, opis estymacji i zastosowania w analityce biznesowej.

Regresja adaptacyjna to technika modelowania, która dostosowuje się do zmieniających się warunków danych. Jest użyteczna w analizie biznesowej, gdzie zmienne i relacje między nimi mogą się zmieniać w czasie.

1. Metoda k-średnich i jej zastosowanie w ocenie wartości klienta w czasie CLV.

Metoda k-średnich to technika grupowania danych, która może być wykorzystana do segmentacji klientów. W analizie CLV może pomóc w identyfikacji grup klientów o podobnych zachowaniach i potrzebach.

1. Wymień i omów zastosowania biznesowe modeli oceny wartości klienta w czasie CLV.

Modele CLV pozwalają firmom ocenić, ile dany klient przyniesie wartości w przyszłości. To jest istotne w zarządzaniu relacjami z klientami, marketingu i strategii cenowej.

1. Przedstaw plusy, minusy sekwencyjnego przetwarzania danych oraz jego inne alternatywy.

Sekwencyjne przetwarzanie danych polega na analizie danych w kolejności ich napływania. Jest przydatne w analizie strumieniowej, gdzie dane są dostarczane w czasie rzeczywistym. Jedną z zalet jest możliwość szybkiego reagowania na zmiany w danych. Jednak może być trudne do zarządzania i wymagać dużej mocy obliczeniowej. Istnieją również alternatywy, takie jak przetwarzanie wsadowe i przetwarzanie w pamięci podręcznej, które mają swoje plusy i minusy.

------------------------------------------------------------------------------------------------------------------------------------------

1. Jakie statystyki opisowe są odporne na wartości nietypowe?

Istnieją statystyki opisowe, które są odporne na wartości nietypowe, czyli obserwacje odstające. Jednym z popularnych rozwiązań jest zastosowanie mediany zamiast średniej arytmetycznej. Mediana to wartość, która dzieli dane na dwie równe części, co sprawia, że jest mniej podatna na ekstremalne wartości. Inną przydatną miarą jest kwartyl pierwszy (Q1) i trzeci (Q3), które definiują granice, między którymi mieszczą się 50% danych. Dzięki nim można wykryć obserwacje odstające.

1. Jakie statystyki opisowe należy stosować w przypadku prób pobranych z populacji o rozkładzie innym niż rozkład normalny?

W przypadku prób pobranych z populacji o rozkładzie innym niż normalny, warto stosować statystyki opisowe odpornie na założenia o rozkładzie normalnym. Można używać miar pozycyjnych, takich jak mediana, kwartyle czy odchylenie od mediany. Innym podejściem jest transformacja danych w celu uzyskania rozkładu zbliżonego do normalnego, co umożliwia zastosowanie klasycznych statystyk opisowych.

1. Przedstaw plusy i minusy struktur danych: analitycznej i transakcyjnej.

Struktury danych analityczne są optymalne do analizy i raportowania danych. Są zoptymalizowane pod kątem szybkiego odczytu i przetwarzania, ale mogą być trudne w utrzymaniu. Struktury transakcyjne są zaprojektowane do obsługi operacji na danych, takich jak dodawanie, aktualizacja i usuwanie. Są bardziej odporne na błędy i gwarantują integralność danych, ale są mniej efektywne w analizie.

1. Omów trzy elementy bezpieczeństwa informacji: Poufność, Integralność, Dostępność.

Poufność oznacza zapewnienie, że tylko uprawnione osoby mają dostęp do danych. Integralność dotyczy zapewnienia, że dane nie zostaną nieuprawnienie zmienione. Dostępność polega na zapewnieniu dostępu do danych w dowolnym momencie. Plusy to ochrona informacji, ale minusy to koszty i ograniczenia.

1. Co to jest Spear Phishing (ukierunkowany atak na osobę lub organizację)?

Spear Phishing to ukierunkowany atak, w którym atakujący podszywa się pod zaufaną osobę lub organizację, próbując oszukać odbiorcę i uzyskać poufne informacje. Atakujący dostosowuje wiadomości, by wydawały się autentyczne, co zwiększa skuteczność ataku.

1. Przedstaw podstawowe zasady cyberbezpieczeństwa dla MŚP (Małe i Średnie Przedsiębiorstwa).

Podstawowe zasady cyberbezpieczeństwa dla MŚP obejmują regularne aktualizacje oprogramowania, szkolenie pracowników w zakresie bezpieczeństwa, stosowanie silnych haseł, backup danych i monitorowanie aktywności sieciowej.

1. Na czym polega interpretacja języka programowania, podaj przykłady języków interpretowalnych oraz kilka różnych interpreterów.

Interpretacja języka programowania polega na tłumaczeniu kodu źródłowego na instrukcje zrozumiałe dla komputera. Przykłady języków interpretowalnych to Python, JavaScript, i Ruby. Różne interpretery to CPython dla Pythona, Node.js dla JavaScriptu itp.

1. Omów sposób instalacji i zarządzania bibliotekami (pakietami) w środowisku języka Python, podaj przykłady popularnych bibliotek.

W środowisku języka Python biblioteki można instalować za pomocą narzędzi takich jak pip lub conda. Przykładowe popularne biblioteki to numpy do obliczeń numerycznych, pandas do analizy danych, i matplotlib do tworzenia wykresów.

1. Omów techniki iterowania na przykładzie wybranego języka programowania, np. R, Python.

Techniki iterowania pozwalają na przechodzenie przez elementy kolekcji danych. W Pythonie można używać pętli for do iteracji przez listy, krotki, słowniki, i wiele innych struktur danych. W R również można stosować pętle, ale często używa się funkcji takich jak apply() do bardziej efektywnej iteracji.

1. Omów koncepcję funkcji oraz zakresu ważności zmiennych na przykładzie wybranego języka programowania, np. R, Python.

Koncepcja funkcji polega na grupowaniu zestawu instrukcji w jedną jednostkę, która może być wywoływana wielokrotnie z różnymi argumentami. Zakres ważności zmiennych określa, gdzie dana zmienna jest dostępna i ma znaczenie. W Pythonie zmienne lokalne istnieją tylko w obrębie funkcji, a zmienne globalne są dostępne w całym programie. W R zakresy zmiennych są bardziej złożone, ale ogólnie podobne zasady obowiązują.