Pregled mogućih pitanja za završni pismeni te završni usmeni ispit iz predmeta "Upravljanje podacima"

1. Definicija skladišta podataka. Koji je cilj skladištenja podataka?

SKLADIŠTE PODATAKA je baza podataka koja **sadrži povijesne nepromjenjive podatke** koji se prikupljaju i obrađuju r**adi potpore poslovnom odlučivanju**. Podaci se izvlače iz raznovrsnih izvora te se u skladu s definiranim modelom podataka učitavaju u skladište podataka i integriraju s postojećim podacima.

CILJ skladištenja podataka:

- što bolje i što brže iskoristiti postojeće podatke za dobivanje korisnih informacija
- omogućiti fleksibilan i učinkovit pristup svim relevantnim podacima
- omogućiti dobivanje cjelovite slike o poslovanju poduzeća
- analizirati i razumijeti poslovna kretanja, pratiti i predviđati situaciju na tržištu, što bolje isplanirati sljedeće korake poslovanja, što brže reagirati na promjene, poboljšati odnose s poslovnim partnerima i kupcima

2. Što je to područno skladište podataka (data mart)?

PODRUČNO SKLADIŠTE PODATAKA je skup podataka dizajniran i konstruiran radi potpore odlučivanju pri čijem se dizajniranju slijede principi dizajna skladišta podataka s tim da taj skup podataka poslužuje potrebe **homogene grupe korisnika**.

3. Navedite i objasnite četiri osnovna svojstva skladišta podataka.

Osnovna svojstva:

1. TEMATSKA ORIJENTACIJA - Skladište podataka je orijentirano na **glavna tematska područja** poslovanja neke organizacije (izdavanje računa, naručivanje robe, prijam pacijenata, pristup web-stranici), odnosno **važne entitete** iz stvarnog svijeta (korisnik, proizvod, korisnički račun, ...). Nije orijentirano na procese.

2. INTEGRIRANOST - više izvora podataka u poduzeću i samo jedno skladište

3. **SKUP VREMENSKIH SNIMAKA** - Skladište podataka je **dugi vremenski slijed zabilježenih snimaka** - slojeva podataka na kojima su spremljeni **podaci** o poslovanju poduzeća **za određene trenutke, odnosno vremenska razdoblja**.
4. **NEPROMJENJIVOST PODATAKA** - U skladištu podataka dvije vrste operacija: **učitavanje relevantnih podataka** i **upiti**. Kad su podaci učitani, nad njima više **nema promjena**.Nema ni mijenjanja ni brisanja podataka (osim u slučaju da treba ispraviti pogrešan podatak).

4. Navedite razlike između skladišta podataka i operacijske (tj. transakcijske) baze podataka.

	OLTP	Skladištenje podataka
Sadržaj podataka	Trenutne vrijednosti, detaljni podaci	Povijesni podaci, detaljni i sumarni podaci
Vrijednost podataka	Vrlo promjenjiva	Postojana
Namjena	Vođenje poslovnog sustava, dnevne operacije	Izvješćivanje o stanju poslovnog procesa, analiza
Jedinica obrade	Transakcija	Upit
Korisnici	Službenici	Analitičari i menadžeri
Raspoloživost	Izuzetno važna	Manje važna
Izmjena podataka	Polje po polje	Nema direktne izmjene
Radne karakteristike	Čitanje/pisanje	Čitanje
Interakcija korisnika	Predodređena	Ad-hoc
Pristup zapisima	Desecima	Milijunima
Fokus	Spremanje podataka	Dobivanje informacija

5. Što obuhvaća izvlačenje podataka iz izvora i koji problemi se tu pojavljuju?

Izvlačenje podataka (eng. *extracting*) podrazumijeva čitanje i razumijevanje izvornih podataka, odabir podataka koji su najkvalitetniji i najrelevantniji za poslovnu analizu, te preslikavanje tih podataka radi daljnje obrade.

Problemi:

- Čitanje podataka iz raspoloživih izvora koji posjeduju **nejasne i nedokumentirane vrijednosti** podataka, a odnosi među podacima imaju dosta **nekonzistentnosti i zalihosti**.
- Za zastarjele sustave koji se koriste tekstualnim datotekama, mrežnim i hijerarhijskim bazama podataka smještenim na *mainframe* računalima koristi se naziv **naslijeđeni sustavi**. Ti su sustavi izvor povijesnih podataka, a **često ne podržavaju on-line režim rada**.
- **Metapodaci**, tj. podaci o imenu datoteka, nazivu polja, ograničenjima i tipovima podataka, mijenjaju se tokom vremena. Te **promjene nisu uvijek dokumentirane**, te se ne može sa sigurnošću utvrditi kada je došlo do promjena u naslijeđenoj aplikaciji. Ljudi koji su razvijali te aplikacije rijetko su još dostupni.
- Pri ispitivanju izvora podataka **problem nekonzistentnosti između više odvojenih sustava** dolazi u središte pozornosti.
- Podaci su raspoređeni u više operacijskih baza podataka i među njima postoji znatna **zalihost**.
- **Podaci** koji se odnose na iste entitete **razlikuju se po formi i sadržaju**. Svaki izvor podataka izrađen je na temelju vlastitog skupa zahtjeva - tijekom razvojnog procesa ne vodi se briga o drugim bazama pa nema usklađenog pogleda na podatke.

6. Što sve uključuje transformacija podataka?

Transformacija podataka u *pripremnom spremištu podataka* uključuje:

- **čišćenje** podataka
- **kombiniranje više izvora** podataka i **integraciju** podataka
- **prilagodbu izvornih podataka modelu podataka skladišta**
- **nametanje novih ključeva**
- **izgradnju agregacija** (tj. organiziranje pohrane određenih podataka dobivenih korištenjem agregatnih funkcija SUM, AVG, COUNT) radi poboljšanja izvedbe čestih upita

7. Opišite pregled kvalitete i čišćenje podataka.

Pregled kvalitete podataka u operacijskim sustavima:

- zalihost podataka,
- nedostatak zajedničkog standarda za pohranjivanje podataka,
- jednostavne greške pri unosu podataka,
- nedostatak određenih podataka,
- raznovrsni podaci upisani u tekst slobodne forme

Čišćenje podataka uključuje:

- ispravljanje pogrešaka nastalih pri unosu podataka
- rad s podatkovnim elementima koji nedostaju
- raščlanjivanje podataka u standardne formate
- osiguravanje domenskog i entitetskog integriteta

8. Koja su tri osnovna načina korištenja skladišta podataka?

Tri osnovna načina korištenja:

- **izrada izvještaja** (*reports*)
- **analitička obrada OLAP** (*On-line analitical processing*)
- **dubinska analiza podataka** (*data mining*)

9. Objasnite zvjezdastu shemu (dimenzijski model). Kakvi se atributi biraju za mjere (ili činjenice)?

ZVJEZDASTA SHEMA je posebni model kojim se oponaša podatkovna struktura višedimenzionalnog polja i koji je optimiziran za brzu i jednostavnu izvedbu složenih višedimenzijskih upita. Zvjezdasta shema se sastoji od jedne velike središnje tablice (tzv. **činjenične tablice**) i **više dimenzijskih tablica.** Svi **strani ključevi na dimenzijske tablice** zajedno čine **složeni primarni ključ činjenične tablice**. To znači da je redak činjenične tablice određen kombinacijom primarnih ključeva svih dimenzijskih tablica. U **činjeničnoj tablici** (engl. *fact table*) nalaze se također **mjere**.

Mjere (engl. *measures*) su brojčani atributi s kontinuiranim (neprekidnim) skupom vrijednosti.

10. Što znači da su dimenzijske tablice hijerarhijski organizirane? Zašto su tako organizirane? Jesu li dimenzijske tablice u 3. normalnoj formi? Objasnite na primjeru.

Hijerarhija je vrlo važna u dimenzijskom modeliranju, budući da ona **omogućuje dobivanje detaljnijeg ili sumarnijeg višedimenzijskog pogleda na podatke**. Krajni korisnik će obično prvo promatrati sumarne podatke, a zatim će dio podataka gledati detaljnije.

Dimenzijske tablice su **denormalizirane** radi jednostavnosti dizajna i učinkovitijeg izvođenja upita - u dimenzijskim tablicama mogu postojati tranzitivne funkcijske zavisnosti (tj. dimenzijske tablice **nisu u 3. normalnoj formi**).

Primjer: atribut **godina** (funkcijski ovisi o atributu ***mjesec*** i tranzitivno o atributu ***kljuc_datum***)

11. Objasnite razliku između zbrojivih, poluzbrojivih i nezbrojivih mjera (činjenica). Gdje se u zvjezdastom spoju pohranjuju vrijednosti mjera?

Potpuno **zbrojive** mjere mogu se zbrajati po svim dimenzijama koje opisuju činjenicu.

Poluzbrojive mjere mogu se zbrajati samo po nekim dimenzijama. Mjera je **zbrojiva nad određenom dimenzijom** ako se njene vrijednosti mogu agregirati **po pripadajućoj hijerarhiji** koristeći operator zbrajanja.

Za **nezbrojive** mjere ne može se izvršiti agregacija niti po jednoj vremenskoj dimenziji. Najčešće se radi o atributima koji bi se trebali naći u dimenzijama, ali zbog načela logičkog oblikovanja ostaju kao opis činjenice.

Mjere se pohranjuju u **činjeničnoj tablici.**

12. Kako se izvode upiti nad zvjezdastim spojem?

Izvedba upita:

- **U dimenzijskim tablicama** se nađu sve **vrijednosti ključa** koje zadovoljavaju postavljena ograničenja.
- Od njih se spoje sve moguće **kombinacije složenih ključeva** koje će se tražiti u činjeničnoj tablici.
- Svi nađeni podaci **u tablici činjenica** se zatim **grupiraju** i **sumiraju** prema specifikacijama korisnika.

13. Za zadani primjer odrediti dimenzije, hijerarhijske razine i mjere u zvjezdastoj shemi.

14. Objasnite sporopromjenjive dimenzije.

Vrijednosti u sporo promjenjivim dimenzijskim tablicama su prilično statične, no i one će se u određenim slučajevima mijenjati tijekom vremena.

Primjeri:

- Kupac X je promijenio adresu.
- Promijenio se opis proizvoda A.
- Promijenila se zemljopisna podjela na regije.

15. Objasnite korištenje iste dimenzijske tablice u više različitih uloga.

Za dimenziju koja se pojavljuje u više uloga napravi se jedna dimenzijska tablica te se za svaku ulogu napravi virtualna relacija (*view*) preko koje se pristupa podacima iz dimenzijske tablice. Ista dimenzijska tablica je tako više puta spojena na činjeničnu tablicu – za svaku ulogu stavi se u činjeničnu tablicu jedan atribut koji je strani ključ na dimenzijsku tablicu i dio je složenog ključa činjenične tablice. Preporučuje se u svakoj virtualnoj relaciji **preimenovati nazive atributa**. Dimenzija koja se najčešće koristi u više uloga je vremenska dimenzija.

16. Što su agregacijske tablice? Navedite nekoliko primjera. U kojim slučajevima treba kreirati agregacijske tablice?

Agregirane činjenice (mjere), tj. činjenice koje su unaprijed sumirane i pohranjene, najčešće se dobivaju tako da se činjenice s osnovne razine zrnatosti zbrajaju, no činjenice se mogu i prebrojavati ili se može računati njihova minimalna, maksimalna ili srednja vrijednost i sl. Agregirane činjenice se najčešće spremaju u **posebne činjenične tablice**, odvojeno od podataka osnovne razine. Svaka agregacijska razina ima svoju činjeničnu tablicu. Pojedine **agregacijske tablice** se mogu staviti *off-line* i zatim opet *on-line*, a da to nema utjecaja na ostale podatke.

Primjer:

- Prodaja po:
  kategoriji proizvoda, danu i trgovini
  proizvodu, mjesecu i gradu
  tromjesečju i regiji

Agregacijske tablice se kreiraju u slučaju kada treba brzo izvesti upite nad podacima nad kojima se često vrši grupiranje, a imaju velik broj redaka. To se određuje prema početnim korisničkim zahtjevima, a zatim se vrše korekcije tijekom korištenja.

17. Navedite i objasnite barem tri temeljne razlike između "tradicionalnog" i "stvarnovremenskog" skladištenja podataka.

▪ **tradicionalni** skladišni sustav učitava podatke u skladište tijekom "mirovanja" poslovanja (*downtime*) kada neće opterećivati postojeće informatičke resurse – tzv. *batch loading*	▪ **stvarnovremenski** skladišni sustav konstantno učitava podatke u skladište – tzv. *stream loading*
▪ skladište predstavlja sliku povijesnih podataka	▪ skladište predstavlja sliku povijesnih podataka ali i prikaz trenutnih poslovnih događaja
▪ skladište podataka služi kao potpora za strateško, dugoročno odlučivanje	▪ skladište podataka je potpora za strateško (dugoročno) ali i taktičko (operativno) odlučivanje
▪ postoji stroga i vidljiva razlika između "operativne" i "analitičke" informatičke podrške poduzeća	▪ operativni i analitički sustavi predstavljaju zatvorenu petlju

18. Što znači da skladišni sustav radi u skoro stvarnom vremenu (*near real time*)?

Skladišni sustav radi u skoro stvarnom vremenu ako je latencija između pojave poslovnog događaja i njegove evidencije u skladištu podataka značajno smanjena, odnosno ako se osvježavanje vrši u puno manjem intervalu nego što je to uobičajeno recimo jedan dan. U ovom slučaju nema čekanja na razdoblja smanjenog intenziteta pojave novih događaja kako bi se podaci evidentirali u skladištu.

19. Opišite različite načine pohrane stvarnovremenskih podataka u sustavu stvarnovremenskog skladištenja.

Četiri načina pohrane RT podataka:

- **povećavanje frekvencije učitavanja ETL sustava** - npr. učitavanje dva puta dnevno
- **direktno učitavanje (Direct Trickle Feed)** - prikupljati nove podatke u stvarnom vremenu , transformirati ih i unositi u skladište podataka, problem skalabilnosti –skladište je optimizirano za upite, ne za česti unos i ažuriranje podataka - preopterećenje sustava
- **učitavanje i preklapanje (Trickle & Flip)** - podaci se stalno učitavaju u tablice koje se nalaze izvan skladišta, u određenim vremenskim intervalima podaci se ažuriraju u skladištu
- **stvarnovremenska particija** - zasebna stvarnovrem. baza koja je optimizirana za česta učitavanja i stvarnovrem. režim rada - može biti i izvan SUBP-a gdje je skladište, nema degradacije performansi i problema skalabilnosti, ali ima više posla oko održavanja i prilagođavanja upita

20. Navedite obilježja polustrukturiranih podataka i objasnite ih na odgovarajućim primjerima.

Polustrukturirane podatke karakterizira barem jedno od sljedećih dvaju osnovnih **obilježja**:

- **Nepravilna struktura**

  Shema nije strogo ograničavajuća i bez iznimaka kao u slučaju potpunu strukturiranih podataka u relacijskoj bazi podataka.

  Veliki skupovi podataka često se sastoje od manjih podatkovnih jedinica čija struktura je međusobno slična, ali ne i posve jednaka.

  {autor: {prezime: "Poe"}, {ime: "Edgar"}, {srednje: "Allan"}}
  {autor: {prezime: "Gide"}, {ime: "Andre"}}
  {autor: {prezime: "Jesenjin"}, {ime: "Sergej"}, {očevo: "Aleksandrovič"}}

- **Implicitna struktura**

  Shema podatkovne strukture ne mora biti pohranjena odvojeno (u posebnoj shemi ili dokumentu), nego može biti integrirana zajedno s njihovim sadržajem.

  Informacije koje su obično vezane uz shemu ovdje se nalaze zajedno s podacima.

  {kompozitor:
  {prezime:"Verdi"}{ime="Giuseppe"}
  {djelo: {opera:"Traviata"}}
  {djelo: {opera:"Don Carlos"} } }

21. Kad je XML dokument dobro oblikovan, a kad je valjan?

Dobro oblikovani dokument zadovoljava pravila oblikovanja XML dokumenta:

- Svaki element mora imati zatvarajući tag.
- Tagovi su "case sensitive".
- Ako neki element nema nikakav sadržaj, može se koristiti sljedeća sintaksa: *<kolicina mjerJed="kg"/>* što je isto kao: *<kolicina mjerJed="kg"> </kolicina>*
- Tagovi se smiju ugniježđivati ali se ne smiju isprepletati.
- Vrijednosti atributa moraju biti u navodnicima.
- Preporučuje se započeti XML dokument deklaracijom XML-a.
- Znakovi: < > & ' " moraju se uvijek pisati na sljedeći način: *< > & ' "*

XML dokument je **valjan** ako je dobro oblikovan i slijedi pravila koja propisuje njemu pridružen DTD ili XML Schema.

22. Definirati u DTD-u i XML Schemi određeni tip elementa sa zadanom kombinacijom pojavljivanja podelemenata (što uključuje slijed podelemenata, izbor jednog od podelemenata, kardinalnost podelemenata, atribute, kardinalnost atributa,...).

23. Provjeriti je li prema zadanoj XML Schemi neki zadani XML dokument valjan.

24. Različiti primjeri lokacijskih staza u XPathu.

25. Ispisati rezultat određenog upita zadanog u jeziku XQuery.

26. Postaviti odgovarajući XQuery upit da bi se dobio traženi rezultat u XML formatu.

27. Navedite prednosti i nedostatke preslikavanja iz XML strukture u relacijsku strukturu.

Prednost: - relacijski sustavi – dobro poznata, zrela i raširena tehnologija

Nedostaci: - XML – polustrukturirani podaci, mogu imati nepravilnu strukturu (relacije uvijek imaju pravilnu strukturu)
- nema podudarnosti između elemenata i atributa u XML-u i relacija i atributa u relacijskoj bazi

28. Objasnite pohranu XML-a u velike objekte (LOB).

Da bi se omogućila pohrana XML-a u objektno-relacijske sustave, uvedeni su novi tipovi podataka i proširen je jezik SQL. Tzv. Veliki objekti (LOB – Large OBject) su prilagođeni za pohranu XML dokumenata ili njihovih dijelova.

Sadržaj dokumenta se može spremiti u jedan stupac tablice koristeći tipove podataka CLOB (Character Large OBject) ili BLOB (Binary Large OBject), odnosno tipove podataka koji se temelje na CLOB-u ili BLOB-u. Ako se radi pohrani na temelju CLOB-a:

- dokument se čuva kao tekstualna datoteka
- ovisno o implementaciji, za dohvat podataka mogu se koristiti:
  upitni jezici za XML (XQuery, XPath)
  SQL/XML - proširenje SQL-a za XML
  metode za pretraživanje teksta

29. Navedite svojstva izvornih sustava za upravljanje XML bazom podataka. Navedite barem tri takva sustava.

Izvorni (eng. native) sustavi za upravljanje XML bazom podataka su prvenstveno namijenjeni pohrani XML dokumenata. Pohranjuju i obrađuju podatke u XML formatu, bez mapiranja I prebacivanja podataka u relacijsku ili neku drugu strukturu.

Svojstva izvornih XML baza podataka:
- Osnovna logička jedinica pohrane je XML dokument.
- Imaju definiran logički model za XML dokumente. Taj model mora uključivati barem elemente, atribute, tekst i poredak elemenata.

Primjeri sustava:
1. eXist
2. Tamino
3. Oracle Berkeley DB XML

30. Navedite definiciju dubinske analize podataka.

Analiza **velikih, opservacijskih** skupova podataka s ciljem pronalaženja **neočekivanih veza i uzoraka** u skupovima podataka ili sumarnog prikaza skupa podataka na način da vlasniku ili korisniku podataka pruža **nove, razumljive i korisne informacije**.

31. Navedite i kratko objasnite osnovnu podjelu metoda dubinske analize podataka.

Vizualizacijske i interaktivne metode – upoznavanje s podacima

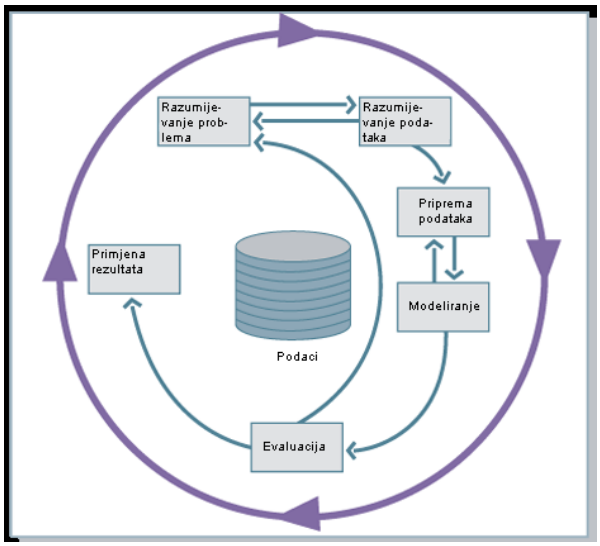
Deskriptivno modeliranje – podaci govore sami za sebe; pronalaze se uzorci i izvode zaključci

- Grupiranje (eng. *clustering*)
- Asocijacijska pravila

Prediktivno modeliranje – analitičar zna što traži ili što želi predvidjeti

- Klasifikacija
- Regresija

32. Objasnite od kojih se faza sastoji model procesa dubinske analize podataka CRISP-DM.



33. Za određeni primjer te uz zadane minimalne parametre (podrške, pouzdanosti i poboljšanja) potrebno je odrediti koja će asocijacijska pravila u konačnici biti prezentirana analitičaru.

34. Navedite probleme koji se javljaju pri upotrebi metode stvaranja asocijacijskih pravila.

Problemi:

- vrlo mnogo potencijalnih uzoraka – računalna zahtjevnost
- određivanje razine parametara
- velik broj pravila – preklapanje, nepreglednost

35. Objasnite moguće uporabe asocijacijskih pravila u maloprodaji.

Primjene asocijacijskih pravila u maloprodaji:

- smještaj artikala
- izrada kataloga i ponuda
- preporučiteljski sustavi

36. Uloga UML-a u oblikovanju baze podataka, u čemu je osnovna razlika u odnosu na oblikovanje pomoću ER dijagrama?

Koristi se za:

- Vizualizaciju
- Specifikaciju
- Izradu modela
- Dokumentiranje

Oblikovanje podataka (engl. data modeling) – sve više se koristi UML za prikaz E-R dijagrama (dijagrami koji opisuju relacijsku bazu podataka).

UML – omogućava opis cjelokupnog procesa razvoja relacijske ili objektno relacijske baze od zahtjeva (engl. business requirements) do fizičkog modela.

37. Nabrojite osnovne UML dijagrame (barem 5).

UML dijagrami:

- Use case
- Sekvencijski
- Dijagram klasa
- Objektni
- Komponentni
- Dijagram razmještaja

* * *

38. Što je to informatika o okolišu?

ENVIROMATICS je područje znanstvenih i tehničkih aktivnosti koje objedinjuje sva „state-of-the-art“ znanja potrebna za dohvaćanje i upravljanje prostorno-vremenskim podacima o prirodi i okolišu podložnom ljudskom djelovanju, kako bi se omogućilo stvaranje, upravljanje i širenje relevantnih informacija o okolišu.

39. Nabrojite svojstva OBJEKTA kao osnovnog koncepta objektnog modela podataka.

OBJEKT kao osnovni **koncept modela podataka** ima sljedeća **svojstva**:

a) Objekt je **jedinstvena posebnost** u prostoru i vremenu i može se jedinstveno prepoznati
b) Objekti imaju **svojstva**, tj. **atribute**
c) Objektima se **rukuje** pomoću **metoda posluživanja**
d) Obilježje objekta je njegovo **ponašanje** i **stanje** koje se **mijenja**
e) Objekti se mogu **klasificirati**
f) Objekti mogu biti **sastavljeni**
g) Objekti **razmjenjuju poruke**

40. Koje se metode modeliranja podataka koriste u postupcima modeliranja objekata ?

U postupku modeliranja objekta koriste se sljedeće **metode**:

- Apstrakcija
- Agregacija
- Podatkovno apstrahiranje
- Proceduralno apstrahiranje
- Metoda učahurivanja
- Metoda nasljeđivanja

41. Koji su osnovni postupci KONCEPTUALNOG i LOGIČKOG MODELIRANJA OBJEKATA ?

Osnovni **postupci *KONCEPTUALNOG*** i ***LOGIČKOG MODELIRANJA OBJEKATA*** su:

- ODREĐIVANJE OBJEKTNOG SUSTAVA
- ODREĐIVANJE TIPOVA OBJEKATA
- OBLIKOVANJE KLASIFIKACIJSKIH I SASTAVNIH STRUKTURA OBJEKATA
- ODREĐIVANJE ATRIBUTA
- UTVRĐIVANJE VEZA I KARDINALNOSTI
- ODREĐIVANJE DOGAĐAJA
- ODREĐIVANJE METODA POSLUŽIVANJA
- ODREĐIVANJE SPOJNICA PORUKA
- PROVJERA MODELA

42. Usporedite model entiteti-veze i objektni model podataka.

Bitna je razlika između objektnih modela i modela entiteti-veze u definiciji metoda posluživanja (servisa).

Model entiteti-veze ne predstavlja konceptualni opis kompletnog informacijskog sustava jer opisuje samo podatkovnu komponentu informacijskog sustava.

Objektni model je semantički bogatiji jer opisuje i procesnu komponentu informacijskog sustava pa predstavlja konceptualni sustava opis kompletnog sustava.