

Tehnologija upravljanja podatkov

Univerzitetni študij RI in UI, 3. letnik
Magistrski študij RI

Matjaž Kukar, 2023/24





Splošne informacije...

- Predavanja

- Izr. prof. dr. Matjaž Kukar, sreda ob 12:15, P21 (???),
matjaz.kukar@fri.uni-lj.si
- Govorilne ure:
 - ???, kabinet 2.04 (2. nadstropje, desno od dvigala)

- Vaje:

- Doc. dr. Luka Šajn

- Spletna stran:

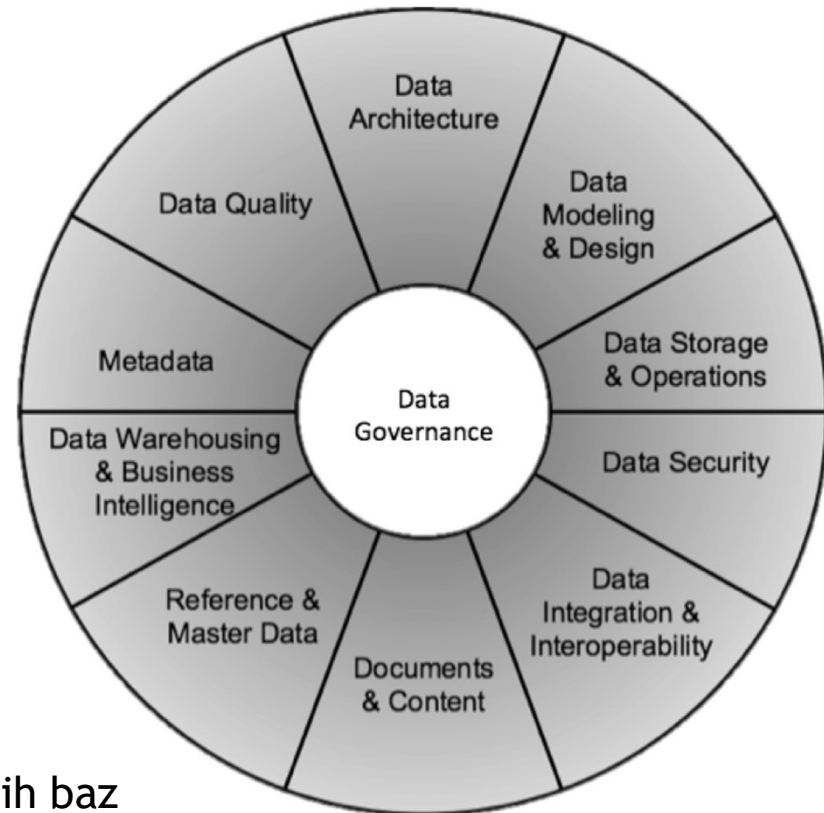
- Tehnologija upravljanja podatkov
<https://ucilnica.fri.uni-lj.si>



Upravljanje s podatki (*data management*)

- Definicija (Data Management Association – DAMA)
Upravljanje s podatki sestavljajo razvoj in izvajanje arhitektur, usmeritev in praktičnih postopkov za podporo celotnemu življenjskemu ciklu podatkovnih potreb sodobnega podjetja.

V kontekstu informacijskih tehnologij:
Tehnologija upravljanja podatkov =
napredne teme s širšega področja podatkovnih baz



Pričakovano predznanje

- **Nadgradnja predmetov Osnove podatkovnih baz/Podatkovne baze**
 - Opisovanje in shranjevanje podatkov v PB
 - Zgradba SUPB, upravljanje z diskom in pomnilnikom
 - Organizacija in indeksiranje datotek
 - Poizvedovanje v PB
 - Relacijski podatkovni model, algebra in račun; SQL
 - Načrtovanje PB
 - Pristopi k načrtovanju PB
 - Konceptualno, logično in fizično modeliranje ???
- **Nadgradnja programerskih predmetov**
 - Osnovno znanje programiranja in uporabe orodij (predvsem Python)
- **Razvoj informacijskih sistemov**
 - Presek nekaterih tem (predvsem načrtovanje), poudarek na PB

Predpriprava na predavanja in vaje

- Primeri na predavanjih-vajah v Pythonu. Ga obvladate?
 - Uradna Python dokumentacija, <http://docs.python.org>
 - J. Demšar: *Python za programerje*. FRI, Ljubljana, 2009.
 - Learn Python the hard way, <http://learnpythonthehardway.org/book/>
 - Learn Python (interaktivno), <http://www.learnpython.org/>
- Aktualna verzija (oktober 2022):
 - Python 3.10; 3.7 ali novejši je OK (**priporočena distribucija Anaconda**)
- Zakaj Python?
 - manjši obseg pisanja (večina primerov na enem zaslonu)
 - hitrejša programiranje
 - uporaba v podatkovni znanosti



Vsebina predavanj

- Eksterni vidiki obvladovanja podatkov:
 - Načrtovanje podatkovnih baz
 - Konceptualno, logično in fizično načrtovanje
 - Normalizacija relacij
 - Analiza uporabniških zahtev
 - Podatkovne baze in podatkovna skladišča
 - Namen in načrtovanje podatkovnih skladišč
 - Zagotavljanje kvalitete shranjenih podatkov
 - Analiza shranjenih podatkov (OLAP, Data Mining)
 - Sodobne nerelacijske podatkovna baze (NoSQL)
 - Dokumentne: MongoDB, grafne: Neo4j, stolpične: MonadDB
 - ... vaše izkušnje?



Vsebina predavanj

- Interni vidiki obvladovanja podatkov:
 - Dostop do podatkov
 - Zagotavljanje dostopnosti in konsistentnosti podatkov (upravljanje s transakcijami)
 - Upravljanje sočasnosti dostopa do podatkovne baze
 - Varovanje in obnavljanje podatkovne baze
 - Relacijska algebra in SQL (ponovitev)
 - Optimizacija in evalvacija poizvedb
 - Načrtovanje izvajanja poizvedb
 - Vrednotenje zahtevnosti osnovnih operacij
 - Alternativne strategije izvajanja poizvedb
 - Upravljanje delno strukturiranih in nestrukturiranih podatkov
 - JSON, XML, prostorski podatki, tekst, zvok, slika ...

Izhodiščna literatura

1. Relacijski SUPB

- a) Thomas M. Connolly, Carolyn E. Begg (2015). Database Systems, A Practical Approach to Design, Implementation and Management, 6th Edition, Pearson
- b) Ramez Elmasri, Shamkant B. Navathe (2016). Fundamentals of Database Systems, 7th Edition, Pearson
- c) S. Sumathi, S. Esakkirajan: Fundamentals of Relational Database Management Systems, Springer, 2007
- d) Carlos Coronel, Steven Morris (2018). Database Systems: Design, Implementation and Management, 13th Edition, Cengage Learning
- e) Raghu Ramakrishnan, Johannes Gehrke (2003). Database Management Systems, Third Edition, McGraw-Hill
- f) Paul Wilton and John W. Colby (2005): Beginning SQL, Wrox

Izhodiščna literatura

2. Nerelacijski (NoSQL) SUPB

- a) Ian Robinson, Jim Webber and Emil Eifrem: Graph Databases, O'Riley (free download), <https://neo4j.com/graph-databases-book>
- b) Dan McCreary and Ann Kelly: Making Sense of NoSQL: A guide for managers and the rest of us, Manning Publications, 2013
- c) Pramod J. Sadalage and Martin Fowler: NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence , Addison-Wesley, 2012
- d) Alex Giamas: Mastering MongoDB 3.x: An expert's guide to building fault-tolerant MongoDB applications, Apress, 2017

3. Spletni viri (sproti)



Vsebina vaj

I. Praktična obravnava nekaterih tem s predavanj

- Spoznavanje s programsko opremo
- Načrtovanje podatkovne baze in podatkovnega skladišča
- Programski dostop do PB
- Normalizacija, priprave na normalizacijo, denormalizacija
- Transakcije
- Optimizacija poizvedb
- Nerelacijski SUPB

II. Domače naloge (vsaj 50%)

III. Seminarska naloga v obliki projekta (vsaj 50%)

- Na koncu semestra (december, januar)
- Nekoliko obsežnejša kot domače naloge
- Obvezen predstavitveni seminar (vaje ali predavanja)
- Pravočasnost izdelave, oddaje in zagovora!



Orodja (minimalne verzije)

- PowerDesigner 12.5 (licenca FRI)
- MariaDB \geq 10.9 (priporočeno 10.11) ali MySQL \geq 8 (priporočeno 8.0.34)
- PostgreSQL \geq 14 (priporočeno 15)
- MySQL Workbench, HeidiSQL
- Python \geq 3.9 + dodatki
- Gonilniki za dostop do baze (npr. ODBC)
- (konec semestra) MongoDB, Neo4j, MonadDB

- Lastni računalniki?
- Docker?



Izpitni red

1. Obveznosti vaj: iz domačih nalog in seminarske naloge morate doseči posamično najmanj 50% možnih točk (pogoj za pristop k pisnemu izpitu).
2. Pisni izpit morate za pozitivno oceno pisati najmanj 50%
3. Vaje sestavljajo pol ocene in veljajo eno šolsko leto!
4. Sodelovanje na predavanjih/vajah se nagrajuje (subjektivno, do 10%)

**Torej: domače naloge \geq 50%, seminarska naloga \geq 50%,
izpit \geq 50%**

Kako izgleda vaše predznanje v praksi?

- Programiranje: **Python**, Java, C/C++/C#, ...
- Programski dostop do podatkovne baze
- Načrtovanje podatkovnih baz (ER): OPB/PB, RIS
- Normalizacija PB: OPB/PB, RIS
- SUPB: **MySQL/MariaDB, PostgreSQL**, SQLite/DuckDB, Oracle, Microsoft SQL Server, NoSQL ?????
- Orodja ORM (SQLAlchemy, Hibernate)

Teme seminarских nalog

- Specifična tematika, večja količina podatkov
- Praktična izvedba
- **Dva** študenta v skupini
- Teme (objavljene bodo na učilnici):
 - Praktična izvedba obdelave (analiza, vizualizacija, ...) nad izbranimi podatki (več nalog), na platformi SUPB + orodje
 - Specifične pregledne teme z vzorčno implementacijo (več nalog)
 - Stare teme seminarских nalog za vzorec

Domača naloga

- Python, Python, Python !!!
 - Vzpostavite si delovno okolje
 - Naredite čim več primerov
 - Ogromno literature na spletu
- Razmislite o svojih kompetencah s področja predmeta in o tem kakšen tip seminarske naloge ali seminarja bi vam najbolj ustrezal
- Oblikujte skupine po **natanko dva** študenta

Terminologija

- Sistem za upravljanje s podatkovnimi bazami (SUPB)
- Podatkovna baza (data base)
- Podatkovno skladišče (data warehouse)
- Sodobni (ne)relacijski SUPB:
 - NoSQL = Not only SQL
 - NewSQL

Sistemi za upravljanje s PB (SUPB)

- Angleško: Database Management System (DBMS)
- Sistem za upravljanje s podatkovno bazo – SUPB je programska oprema za obvladovanje velikih količin podatkov, shranjenih v vnapij točno določeni obliki (logičnem podatkovnem modelu)
- Alternativa – shranjevanje v aplikaciji lastni obliki; problemi: neprenosljivost, nefleksibilnost ...
- Obstaja veliko vrst SUPB. Omejili se bomo predvsem na relacijske in nekatere sodobne nerelacijske:
 - Oracle, Microsoft SQL Server, Postgres, MySQL/MariaDB,
 - MongoDB, Neo4j, MonadDB, ...

Podatkovna baza in podatkovno skladišče

- Podobno, vendar ne enako!
- Podatkovna baza (PB oz. DB):
 - OLTP sistem (on-line transactional processing)
 - opisuje trenutno stanje
- Podatkovno skladišče (PS oz. DW):
 - OLAP sistem (on-line analytical processing)
 - opisuje zgodovino vsebin OLTP sistema (pogosto več OLTP sistemov)
- Oba pristopa tečeta na SUPB, vendar z različnimi prioritetai izvajanja (OLTP - hitro izvajanje transakcij, OLAP - hitra analiza)
- Razlikujejo se tudi postopki načrtovanja

Načrtovanje PB, PS, NoSQL

- Podatkovne baze
 - konceptualno načrtovanje:
 - obsežne PB, veliko število tabel
 - korak pri načrtovanju inf. sistemov
 - logično načrtovanje (normalizacija relacij oz. tabel):
 - direktno za manjše PB (nekaj deset tabel)
 - preverjanje rezultatov konceptualnega načrtovanja
 - fizično načrtovanje
- Podatkovna skladišča
 - načrtovanje zvezdnih shem
- Načrtovanje nerelacijskih PB
 - Relativno novo, nezrelo in neuskklajeno področje
 - Poudarek na horizontalni skalabilnosti!

Zakaj razlike?

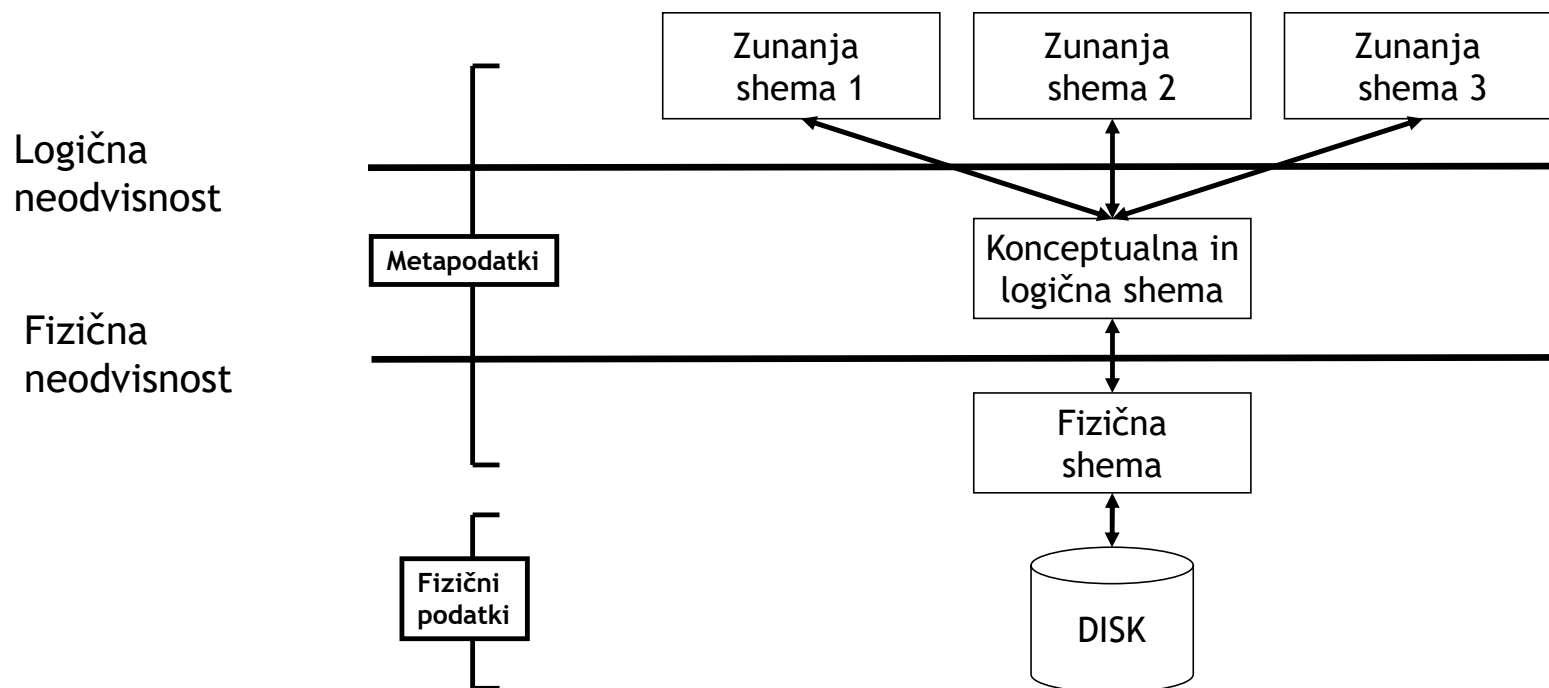
- PB: hitro (sprotno) izvajanje vnaprej definiranih transakcij
- PS: hitro izvajanje spontanih analiz podatkov
- NoSQL: nerelacijski podatkovni modeli, skalabilnost!

Poglavje 1

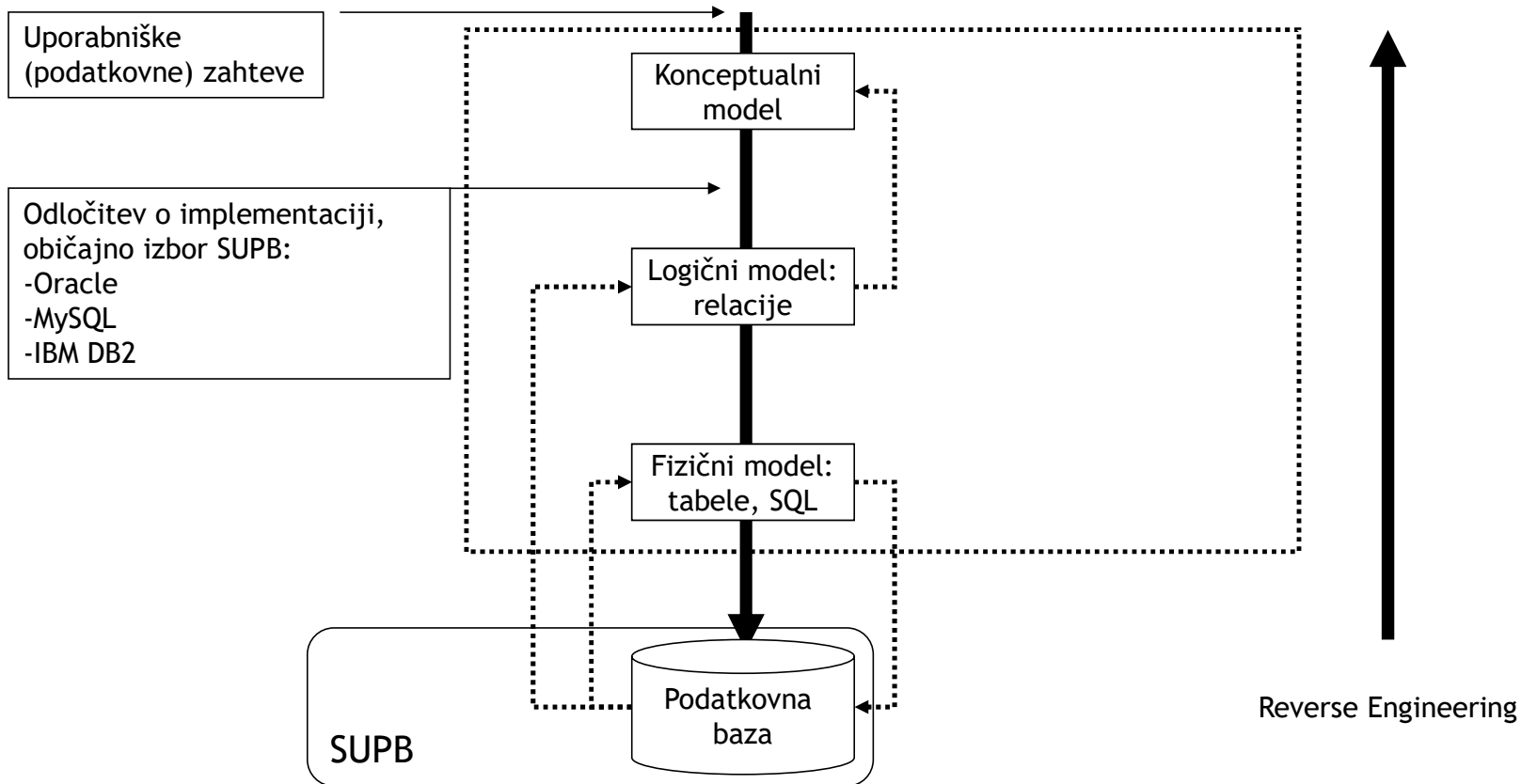
**Konceptualno načrtovanje
transakcijskih podatkovnih baz**

Trinivojska predstavitev podatkov...

- Podatki so v PB opisani na treh nivojih:
 - Zunanje sheme (zunanji, uporabniški nivo)
 - Konceptualna (logična) shema
 - Fizična (notranja) shema



1.1 Trije nivoji načrtovanja – trije modeli..



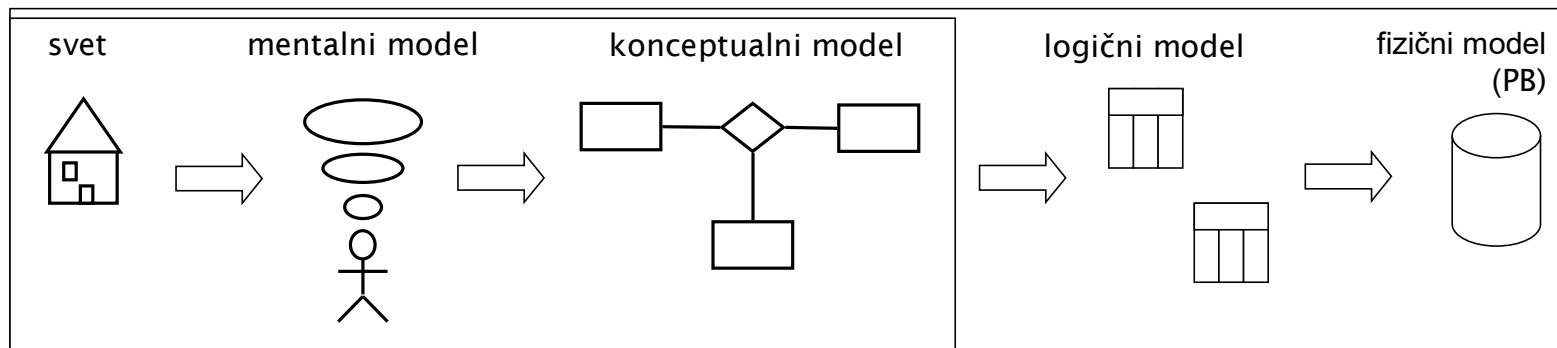


Konceptualni podatkovni model

- Formalizem s katerim opišemo, kaj bi želeli hraniti v PB ter kakšne povezave obstajajo med elementi, ki jih želimo hraniti, se imenuje konceptualni model.
- Konceptualni model je torej način, kako na visoki ravni abstrakcije razumljivo in ne preveč tehnično opišemo podatke, ki jih želimo hraniti ter skrijemo nepomembne podrobnosti.
- Konceptualni model odraža uporabnikovo zaznavanje realnega sveta oziroma poslovnega problema, ter načrtovalčevo videnje problema (stična točka).

1.1 Trije nivoji načrtovanja – trije modeli

- Načrtovanje PB: od realnega sveta do fizične podatkovne baze za potrebe poslovne domene



1.2 Kaj je konceptualno načrtovanje?..

- Konceptualno načrtovanje je postopek opredelitve podatkovnih potreb oz. zahtev (poslovne ali druge) domene s pomočjo konceptualnega modela
- Konceptualno načrtovanje preko konceptualnega modela poskrbi za opis pomena podatkov, potrebnih za poslovno domeno
- Konceptualno načrtovanje je neodvisno od dejanskega podatkovnega modela (relacijski, ...)
- Konceptualnega načrtovanja ne moremo avtomatizirati, za njegovo izvedbo je odgovoren analitik. Gre za prenos semantike v model.

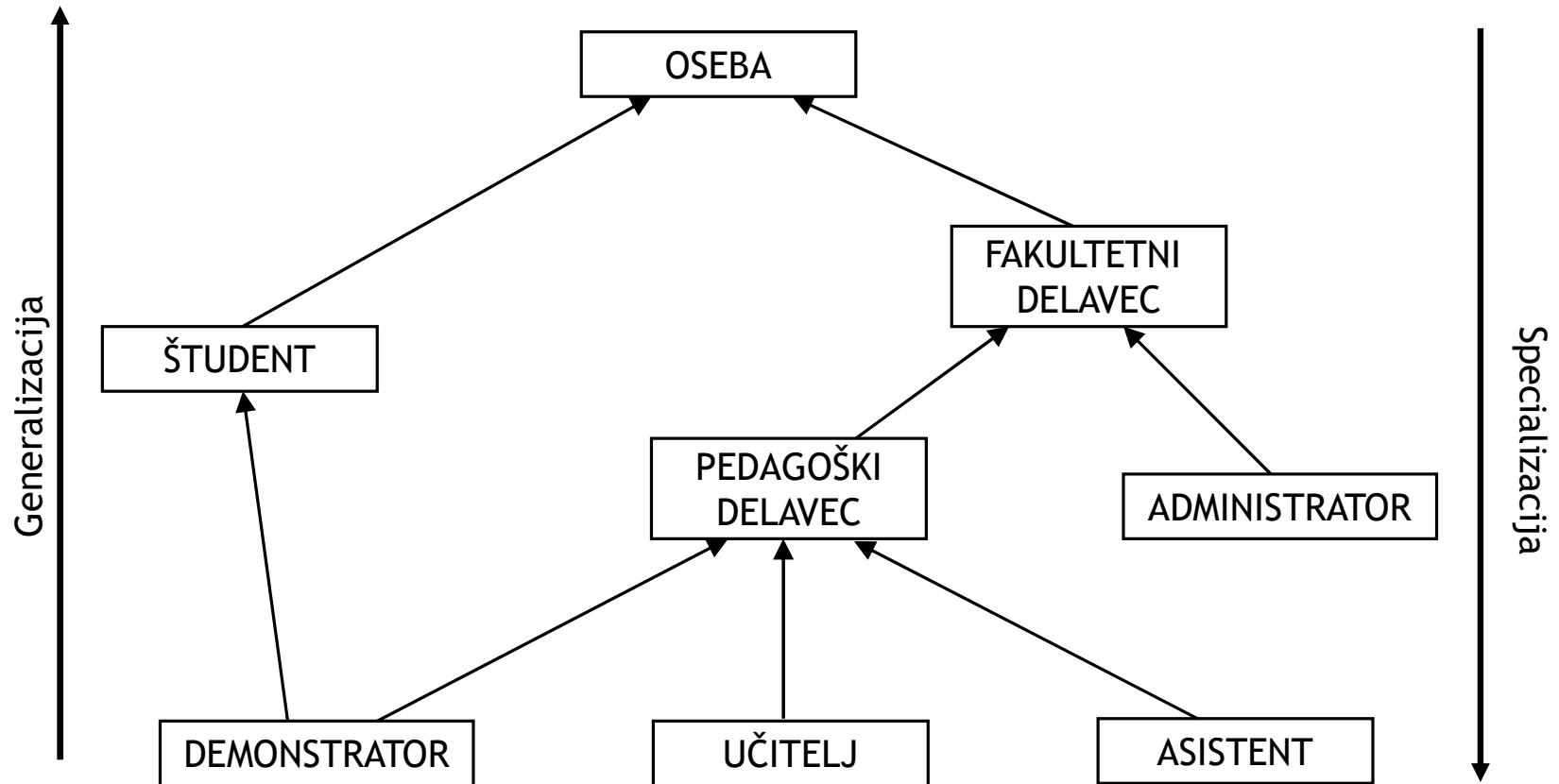
1.2 Kaj je konceptualno načrtovanje?

- Je najbolj kritično, saj se napake narejene pri konceptualnem načrtovanju prenašajo naprej na naslednje modele
- Pri konceptualnem načrtovanju je zelo pomembno sodelovanje uporabnikov in interakcija z uporabniki. Uporabniki so nosilci znanja o poslovni domeni, so poznavalci semantike in običajno vedo, kaj hočejo, a to težko natančno izrazijo
- Konceptualno načrtovanje mora upoštevati tudi poslovna pravila (kot omejitve modela)

1.3 Tehnike konceptualnega načrtovanja..

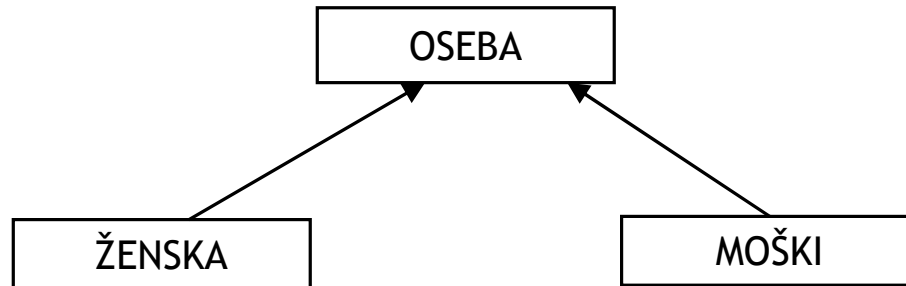
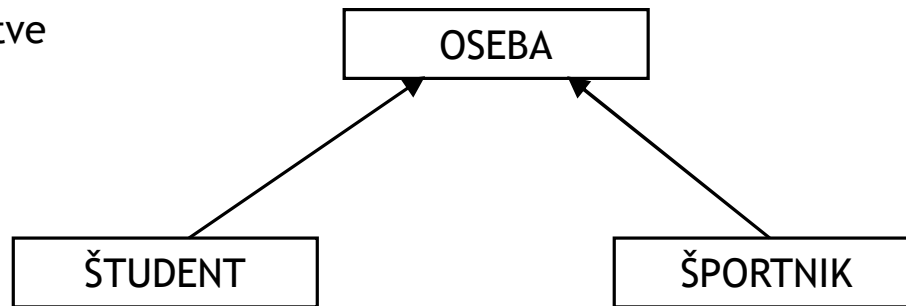
- Percepcija
- Abstrakcija
- Objekti, koncepti (tipi) in término
- Klasifikacija (objekti→koncepti)
- Ureditev hierarhij konceptov (tipov)
- Agregacija (sestavljanje konceptov)

1.3 Tehnike konceptualnega načrtovanja..

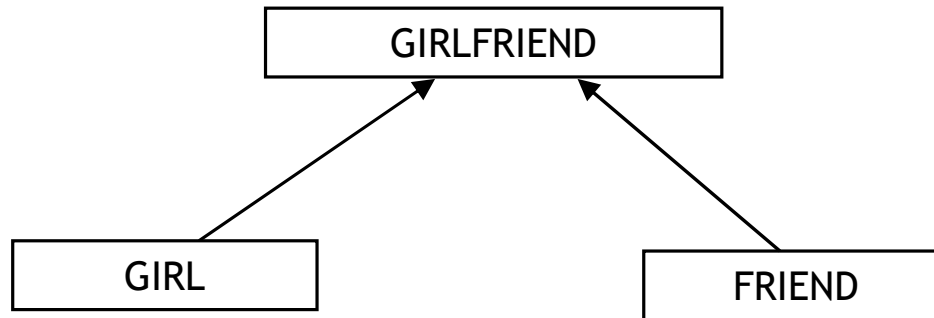


1.3 Tehnike konceptualnega načrtovanja..

Primeri ureditve tipov:
- pokrivanje?



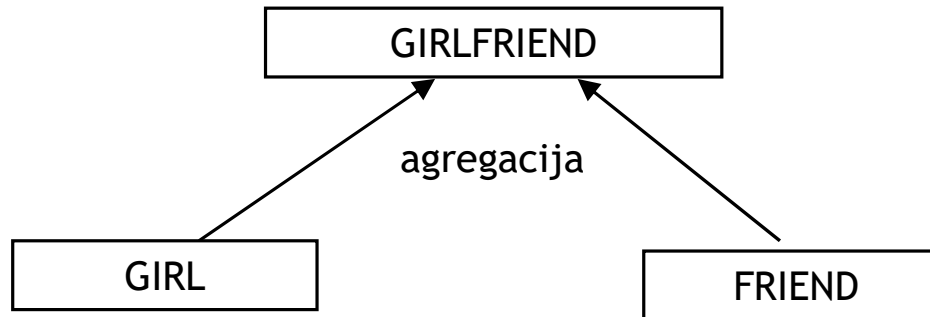
1.3 Tehnike konceptualnega načrtovanja..



??????

1.3 Tehnike konceptualnega načrtovanja

- Preko agregacije definiramo nov – agregiran (sestavljen) tip na osnovi obstoječih tipov



- Kakšne je razlika z generalizacijo? Agregirani tip vsebuje VSE lastnosti agregiranih tipov in ne samo skupnih!
- Pomensko agregirani tip ponuja nekaj novega.