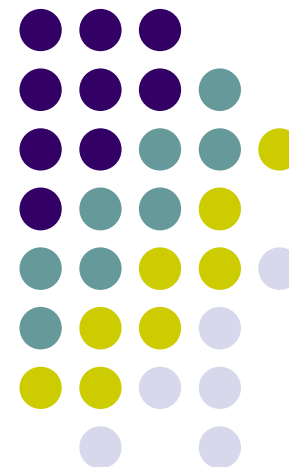


Projektiranje informacijskih sustava

Dizajn pohrane podataka
Ak. god. 2009/2010





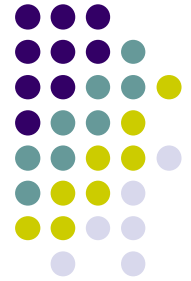
Dizajn pohrane podataka

- Dizajn pohrane podataka definira kako se podaci u sustavu pohranjuju te kako aplikacije rukuju podacima.
- Dizajn treba osigurati da pohrana podataka bude efikasna na način da se osigura zadovoljavajuća brzina odziva pohrane te jednostavan pristup podacima. Česta korisnička primjedba je da je sustav prespor.
- Dizajn pohrane podataka obuhvaća definiranje:
 - formata pohrane podataka,
 - formata podataka (slogova, redaka) u pohrani,
 - optimizaciju pohrane podataka (brzine i veličine).



Formata pohrane podataka

- Dva tipa formata pohrane podataka su:
 - datoteke,
 - baze podataka.
- Datoteke, osobito u starijim sustavima, obično imaju vlastiti definirani format pohrane podataka, ali mogu koristiti i neke uobičajne formate pohrane podataka u datoteku (xml, [csv](#), tab-delimited,...).
- Podaci u datoteci su strukturirani, formatirani u nekom obliku (npr. struktura podataka).



Datoteke

- Datoteke podataka su organizirane sekvencijalno – novi podaci se dodaju na kraj datoteke.
- Podaci se u datotekama povezuju pokazivačima koji sadrže informaciju o lokaciji povezanog podatka.
- Datoteke podataka se često nazivaju vezane liste zbog povezivanja podataka pomoću pokazivača.

Datoteke



Appointment Date	Appointment Time	Duration	Reason	Patient ID	First Name	Last Name	Phone Number	Doctor ID	Doctor Last Name
11/23/2006	2:30	.25 hour	Flu	758843	Patrick	Dennis	548-9456	V524625587	Vroman
11/23/2006	2:30	1 hour	Physical	136136	Adelaide	Kin	548-7887	T445756225	Tantalo
11/23/2006	2:45	.25 hour	Shot	544822	Chris	Pullig	525-5464	V524625587	Vroman
11/23/2006	3:00	1 hour	Physical	345344	Felicia	Marston	548-9333	B544742245	Brousseau
11/23/2006	3:00	.5 hour	Migraine	236454	Thomas	Bateman	667-8955	V524625587	Vroman
11/23/2006	3:30	.5 hour	Muscular	887777	Ryan	Nelson	525-4772	V524625587	Vroman
11/23/2006	3:30	.25 hour	Shot	966233	Peter	Todd	667-2325	T445756225	Tantalo
11/23/2006	3:45	.75 hour	Muscular	951657	Mike	Morris	663-8944	T445756225	Tantalo
11/23/2006	4:00	1 hour	Physical	223238	Ellen	Whitener	525-8874	B544742245	Brousseau
11/23/2006	4:00	.5 hour	Flu	365548	Jerry	Starsia	548-9887	V524625587	Vroman
11/23/2006	4:30	1 hour	Minor surg	398633	Susan	Perry	525-6632	V524625587	Vroman
11/23/2006	4:30	.5 hour	Migraine	222577	Elizabeth	Gray	667-8400	T445756225	Tantalo
11/24/2006	8:30	.25 hour	Shot	858756	Elias	Awad	663-6364	T445756225	Tantalo
11/24/2006	8:30	1 hour	Minor surg	232158	Andy	Ruppel	525-9888	V524625587	Vroman
11/24/2006	8:30	.25 hour	Flu	244875	Rick	Grenci	548-2114	B544742245	Brousseau
11/24/2006	8:45	.5 hour	Muscular	655683	Eric	Meier	667-0254	T445756225	Tantalo
11/24/2006	8:45	1 hour	Physical	447521	Jane	Pace	548-0025	B544742245	Brousseau
11/24/2006	9:30	.5 hour	Flu	554263	Trey	Maxham	663-8547	V524625587	Vroman

Baze podataka



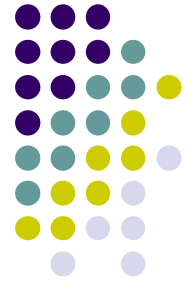
Appointment Date	Appointment Time	Duration	Reason	Patient ID	Doctor ID
11/23/2006	2:30	.5 hour	Flu	758843	V524625587
11/23/2006	2:30	1 hour	Physical	136136	T445756225
11/23/2006	2:45	.25 hour	Shot	544822	V524625587
11/23/2006	3:00	1 hour	Physical	345344	B544742245
11/23/2006	3:00	.5 hour	Migraine	236454	V524625587
11/23/2006	3:30	.5 hour	Muscular	887777	V524625587
11/23/2006	3:30	.25 hour	Shot	966233	T445756225
11/23/2006	3:45	.75 hour	Muscular	951657	T445756225
11/23/2006	4:00	1 hour	Physical	223238	B544742245
11/23/2006	4:00	.5 hour	Flu	365548	V524625587
11/23/2006	4:30	1 hour	Minor surg	398633	V524625587
11/23/2006	4:30	.5 hour	Migraine	222577	T445756225
11/24/2006	8:30	.25 hour	Shot	858756	T445756225
11/24/2006	8:30	1 hour	Minor surg	232158	V524625587
11/24/2006	8:30	.25 hour	Flu	244875	B544742245
11/24/2006	8:45	.5 hour	Muscular	655683	T445756225
11/24/2006	8:45	1 hour	Physical	447521	B544742245
11/24/2006	9:30	.5 hour	Flu	554263	V524625587

Tables related using patient id

Patient ID	First Name	Last Name	Phone Number
136136	Adelaide	Kin	548-7887
222577	Elizabeth	Gray	667-8400
223238	Ellen	Whitener	525-8874
232158	Andy	Ruppel	525-9888
236454	Thomas	Bateman	667-8955
244875	Rick	Grenci	548-2114
345344	Felicia	Marston	548-9333
365548	Jerry	Starsia	548-9887
398633	Susan	Perry	525-6632
447521	Jane	Pace	548-0025
544822	Chris	Pullig	525-5464
554263	Trey	Maxham	663-8547
655683	Eric	Meier	667-0254
758843	Patrick	Dennis	548-9456
858756	Elias	Awad	663-6364
887777	Ryan	Nelson	525-4772
951657	Mike	Morris	663-8944
966233	Peter	Todd	667-2325

Tables related using doctor id

Doctor ID	Last Name
B544742245	Brousseau
T445756225	Tantalo
V524625587	Vroman



Baze podataka

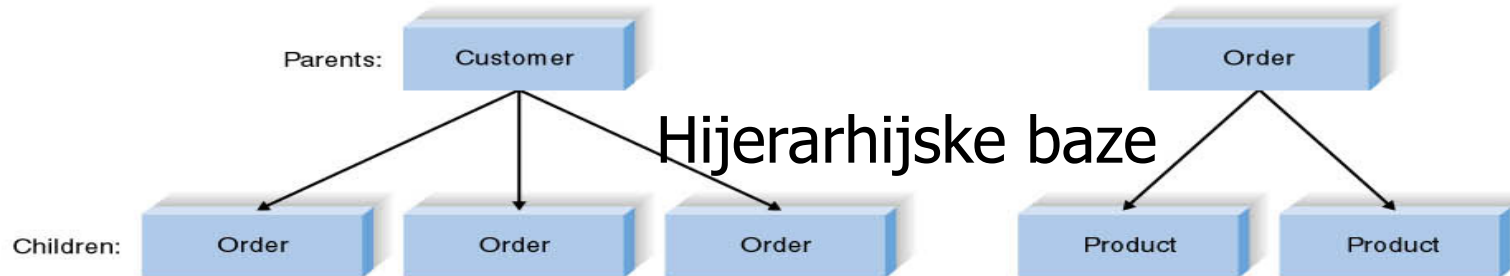
- Database Management System (DBMS) je softver koji kreira i upravlja bazama podataka.
- Postoje različiti tipovi baza podataka:
 1. Naslijeđene (*legacy*) aplikacije obično koriste baze podataka koje se temelje na starijim tehnologijama i danas se više ne koriste u razvoju aplikacija, ali postoji veliki broj sustava koji se temelje na tim “starim” bazama podataka. Primjeri tih “starih” baza podataka su:
 - Hijerarhijske (*hierarchical*) baze
 - Mrežne (*network*) baze



Baze podataka

- Hijerarhijske baze koriste hijerarhiju za definiranje relacija među podacima (npr. IBM-ov Information Management System (IMS) (XML ??)).
- Podaci su organizirani hijerarhijski u obliku stabla.
- Podržavaju relacije 1:1 i 1:M dok se M:N relacije teško ili nikako ne mogu prikazati hijerarhijskim relacijama podataka.
- Prednost im je brzo izvođenje pretraživanja.
- Za razliku od hijerarhijskog model podataka u kojem svaki slog ima jedan roditeljski slog i više slogova djece (stablata struktura), mrežne baze omogućavaju da svaki slog ima više roditeljskih slogova i slogova djece (npr. Turbolmage, IDMS,..). Slogovi tvore mrežnu struktura odakle i ime.

Baze podataka



Notice how Order serves as a child to Customer and a parent to Product.

Sample Records:

Customer as parent

1035 Black	...
	235 11/23/05 ...
1556 Fracken	...
	236 11/23/05 ...
	243 11/26/05 ...
2274 Goodin	...
	237 11/23/05 ...
	245 11/26/05 ...
	260 11/30/05 ...
	275 12/7/05 ...
4254 Bailey	...
	234 11/23/05 ...
	242 11/26/05 ...
9500 Chin	...
	233 11/23/05 ...
	244 11/26/05 ...
	262 11/30/05 ...

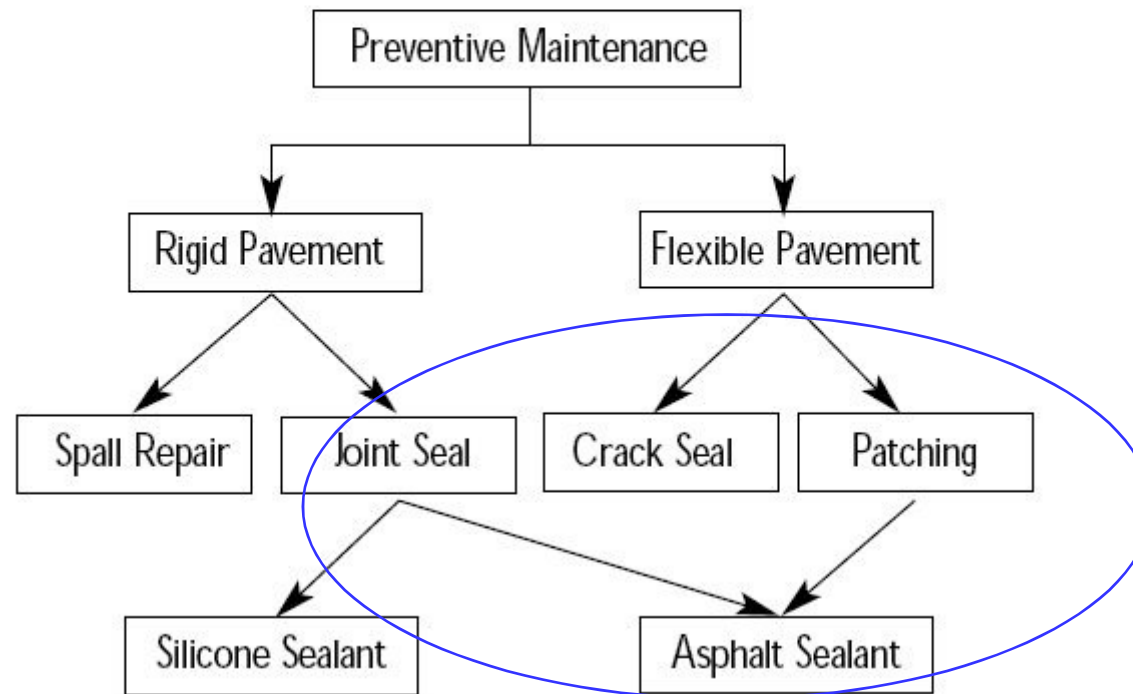
Order as parent

233 11/23/05 ...	
	444 Wine Gift Pack
	222 Bottle Opener
	555 Cheese Tray
234 11/23/05 ...	
	222 Bottle Opener
235 11/23/05 ...	
	555 Cheese Tray
	222 Bottle Opener
236 11/23/05 ...	
	333 Jams & Jellies
	222 Bottle Opener
237 11/23/05 ...	
	111 Wine Guide
242 11/26/05 ...	
	444 Wine Gift Pack
243 11/26/05 ...	
	333 Jams & Jellies
	222 Bottle Opener
	555 Cheese Tray

Baze podataka



Mrežna baza





Baze podataka

- Oba tipa baza podataka omogućuju efikasno pohranjivanje najrazličitijih podataka, ali im je veliki nedostatak složenost aplikacija za rad sa tim podacima.
- Kôd aplikacije mora održavati pokazivače među slogovima što je podložno greškama.
- Nekada je hardver bio skup, a vrijeme programera zanemarivo u odnosu na hardver pa su ovakvi tipovi baza podataka bili često korišteni. No pad cijena hardvera doveo je do sve većeg širenja relacijskih baza.

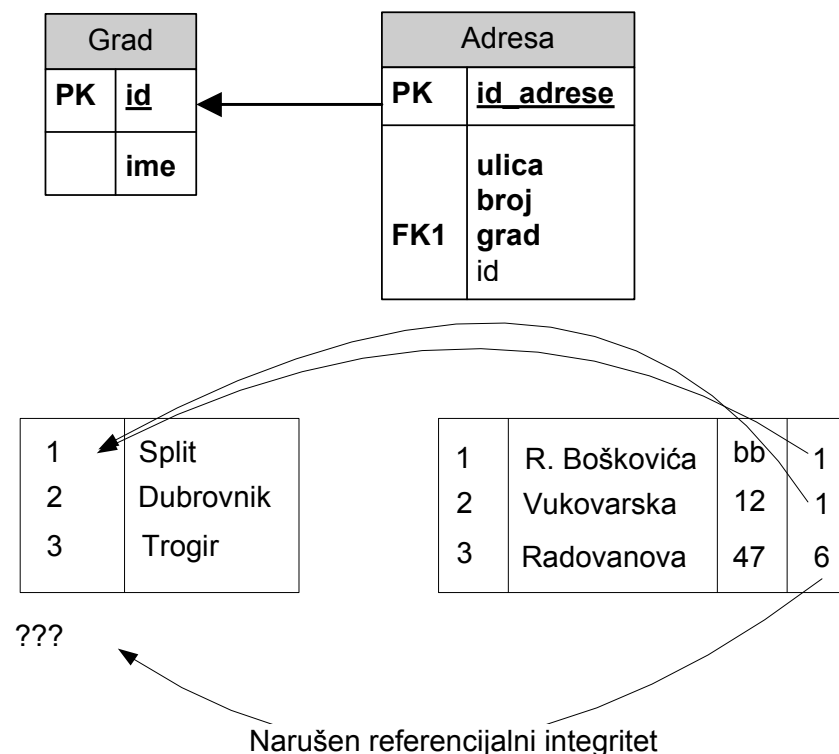
Baze podataka



2. Relacijske baze podataka su najpopularniji tip baza podataka danas.
 - Podaci su u relacijskoj bazi grupirani po tablicama.
 - Relacijska baza je kolekcija tablica koje sadrže slogove.
 - Tablice sadrže primarni ključ (*primary key*) kao polja u slogu koja imaju jedinstvenu vrijednost za sve slogove.
 - Relacije među podacima (tj. tablicama) se postavljaju preko sekundarnih ključeva (*secondary key*) koji su kopije primarnih ključeva iz tablica među kojima se definira relacija.

Baze podataka

- Većina RDBMS (*relational database management system*) podržava referencijalni integritet podataka.
- Referencijalni integritet se odnosi na podatke koji su u relaciji osiguravajući ispravnost primarnih i sekundarnih ključeva.



Baze podataka



- Tablice relacijske baze podataka sadrže definirani skup kolona i promjenljivi broj redaka sa instancama vrijednosti kolona.
- Structured Query Language (SQL) je standardni jezik za pristup podacima u tablicama.
- SQL upit se primjenjuje na čitavu tablicu, za razliku od drugih programskih jezika koji obično manipuliraju pojedinačnim zapisom, a ne cijelom tablicom tj. kolekcijom zapisa. Dohvaćanje podataka iz više tablica radi se povezivanjem tablica preko ključeva u jednu tablicu.
- RDBMS sustavi su Microsoft Access, Oracle, DB2 (IBM), Informix (IBM), Microsoft SQL server, MySQL, PostgreSQL,...

Baze podataka

Relacijska baza

Cust ID is the primary key of Customer

Customer

Cust ID	Last Name	First Name
4254	Baily	Ryan
9500	Chin	April
1556	Fracken	Chris
2487	Hancock	Bill
2243	Harris	Linda
1035	Black	John
1123	Williams	Mary
9501	Kaplan	Bruce
4453	Min	Julie
9505	Marvin	Sandra
2282	Lau	Mark
5927	Lee	Diane
2241	Jones	Chris
2242	DeBerry	Ann
2274	Goodin	Dan
9507	Nelson	Dave
2264	White	Anthony

Order

Order Number	Date	Cust ID	Amount	Payment Type
234	11/23/05	4254	\$30.00	MC
235	11/23/05	9500	\$20.00	VISA
236	11/23/05	1556	\$20.00	VISA
237	11/23/05	2487	\$60.00	AMEX
238	11/23/05	2243	\$50.00	MC
239	11/23/05	1035	\$50.00	AMEX
240	11/23/05	1556	\$20.00	VISA
241	11/23/05	1123	\$40.00	MC
242	11/24/05	9501	\$30.00	VISA
243	11/24/05	4453	\$30.00	VISA
244	11/24/05	9505	\$20.00	VISA
245	11/24/05	2282	\$20.00	AMEX
246	11/24/05	5927	\$60.00	MC
247	11/24/05	2241	\$50.00	VISA
248	11/24/05	4254	\$50.00	AMEX
249	11/24/05	2242	\$50.00	AMEX
250	11/24/05	2274	\$20.00	MC
251	11/24/05	9507	\$10.00	MC
252	11/24/05	2487	\$60.00	MC
253	11/24/05	2264	\$40.00	AMEX

Cust ID is a foreign key in Order

Payment Type is a foreign key in Order

Referential integrity:

- All Payment Type values in Order must exist first in the Payment Type table
- All Cust ID values in Order must exist first in the Customer table

Payment Type

Payment Type	Description
MC	Mastercard
VISA	VISA
AMEX	American Express

Payment Type is the primary key of the Payment Type table



Baze podataka

3. Objektne (ili objektno-orijentirane) baze podataka informacije pohranjuju kao objekte u objektno-orijentiranom programiranju. Objekt ima podatke i procese koji se primjenjuju na tim podacima.
- Objekt je definiran klasom. Postoji mogućnost nasljeđivanja.

Object-Oriented Model

Object 1: Maintenance Report

Date	
Activity Code	
Route No.	
Daily Production	
Equipment Hours	
Labor Hours	

Object 1 Instance

01-12-01
24
I-95
2.5
6.0
6.0

Object 2: Maintenance Activity

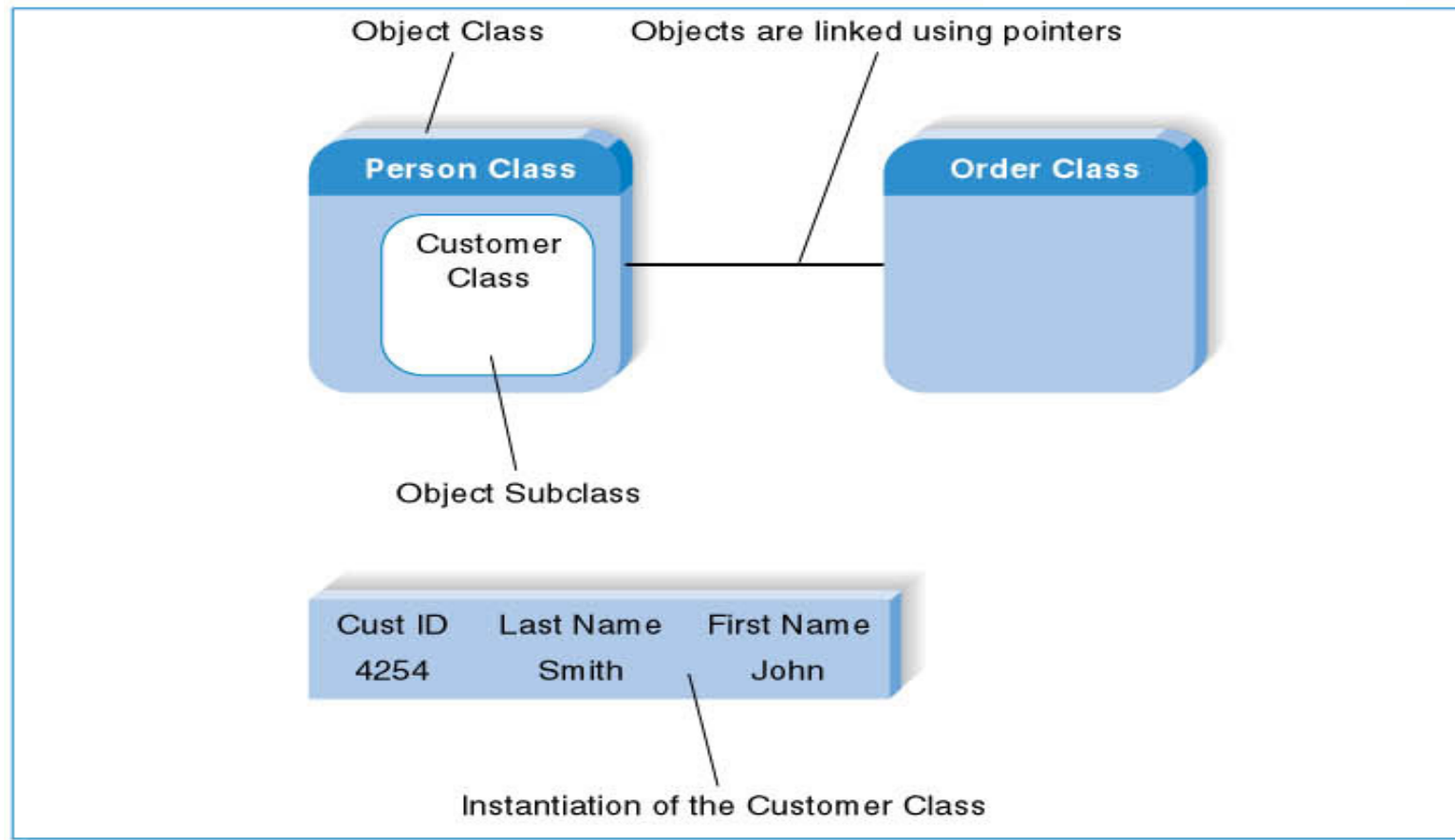
Activity Code	
Activity Name	
Production Unit	
Average Daily Production Rate	



Baze podataka



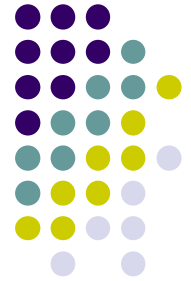
Objektna baza



Baze podataka

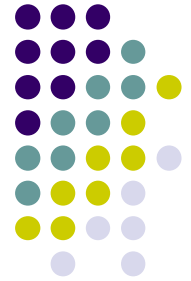


- OODBMS (*object-oriented database management system*) se najčešće koriste za multimedijalne aplikacije ili sustave sa kompleksnim podacima (video, zvuk, grafika, ...) (npr. Jasmine, Matisse <http://www.matisse.com/>,...) . (Npr. pohrana pdf, doc dokumenata u relacijske baze).
- Većina objektnih baza podataka nudi nekakav jezik za upite (query language). ODMG (Object Data Management Group) (<http://www.odbms.org/ODMG/>) je pokušao napraviti standardizaciju jezika za OQL-om (*object query language*). Osim nedostatka standardnog jezika za upite, nedostaju i standardizirani pristupi objektnim bazama (ODBC, JDBC,...), backup i sl. funkcionalnosti uobičajne za relacijske baze podataka.
- Stoga se često koriste hibridni OODBMS koji uključuju i objektne i relacijske osobine u jednoj bazi.



Baze podataka

- Relacijske baze koriste dvodimenzionalnu strukturu podataka u obliku kolona i redaka u tablici. Ovakav model pohrane podataka dizajnirao je Codd 1970 u IBM-u i značajno unaprijedio postojeće modele pohrane podataka. No ovaj model zahtjeva veliku količinu dodatnih podataka (tzv. *overhead*) pa se proširio tek sa razvojem hardvera. Ove baze su jednostavne za dizajn i njihovo upravljanje (DBMS), ali su relativno složene za dohvaćanje podataka, osobito kada se radi o složenijim upitima.
- 4. Višedimenzionalne baze – MDB (*multidimensional database*) podataka su najnoviji tip bazi podataka koji je usmjeren olakšavanju dohvata podataka iz baze.

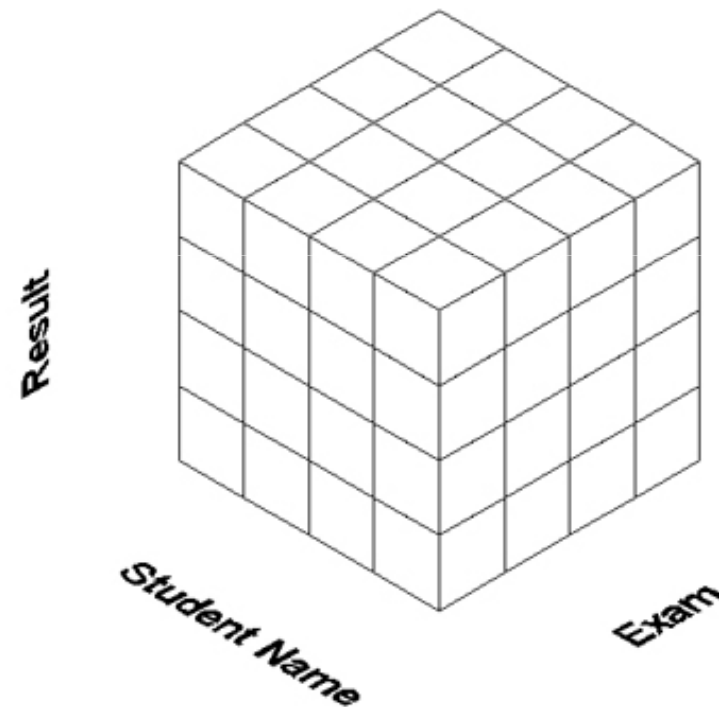


Baze podataka

- MDB je tip baze koji je optimiziran za tzv. podatkovna spremišta (*data warehouse*) i tzv. aplikacije za online analitičko procesiranje (online analytical processing - OLAP).
- Podatkovnim spremištima se nazivaju centralna spremišta svih važnih podataka poslovnog sustava. Mogu ga činiti različite baze podataka i datoteke koje zajednički čuvaju sve te podatke.
- OLAP aplikacije omogućavaju korisniku jednostavno i selektivno dohvaćanje podataka. OLAP aplikacije koje dohvaćaju podatke iz višedimenzionalnih baza se naziva MOLAP (*multidimensional OLAP*).
- MDB baze se često kreiraju sa ulaznim podacima iz relacijskih baza.

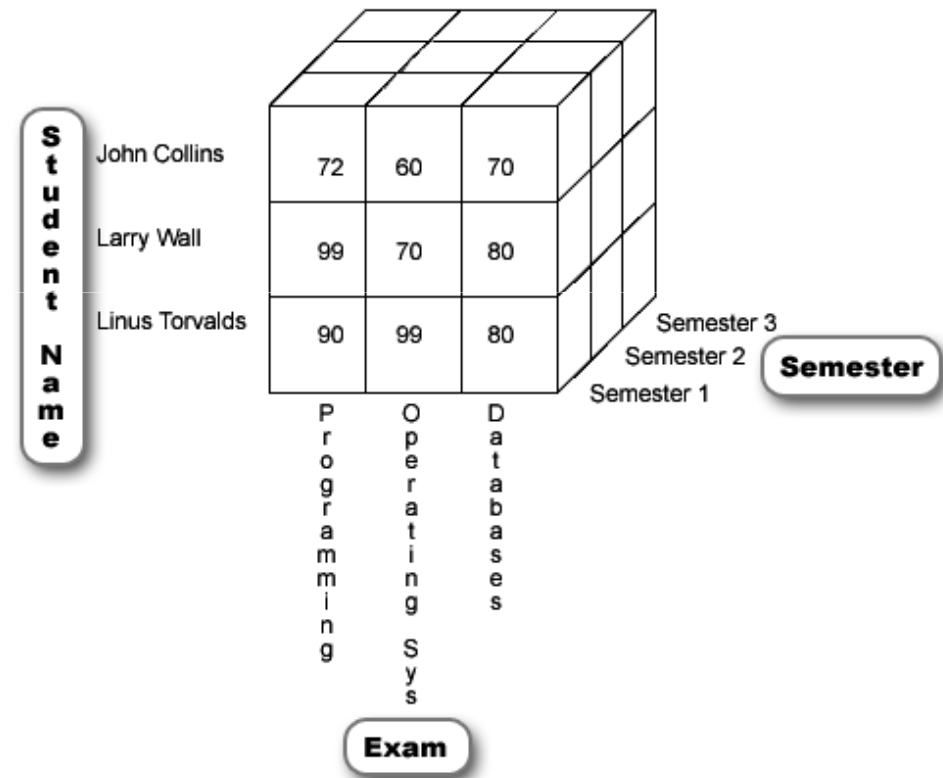
Baze podataka

- Kod višedimenzionalne baze podaci se korisniku predstavljaju u obliku višedimenzionalnog niza, pri čemu je svaki podatak sadržan u jednoj ćeliji niza kojoj se može pristupiti preko višestrukih indeksa.

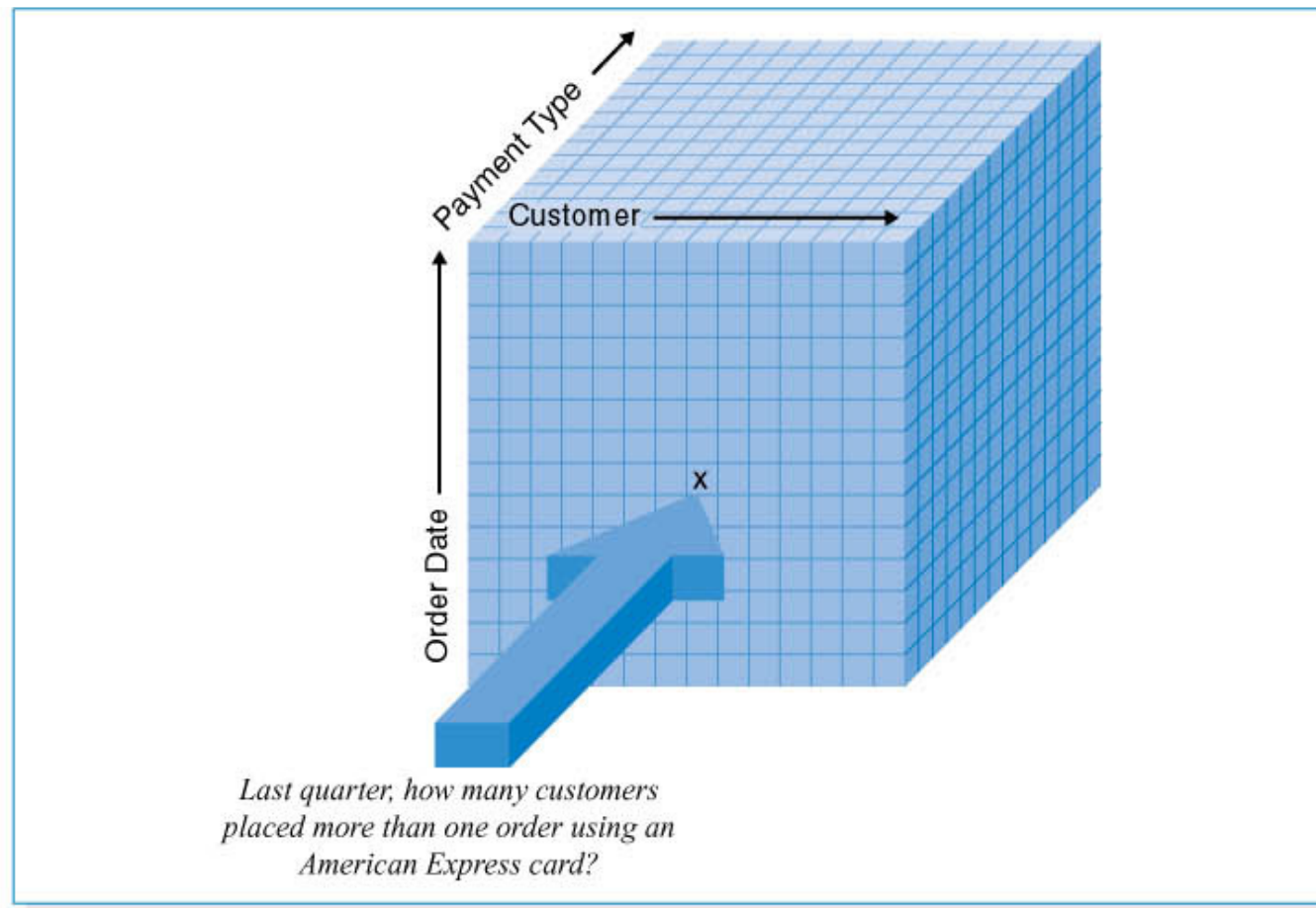


Baze podataka

- Višedimenzionalni niz predstavlja višu organizaciju podataka od relacijskih tablica.
- Pogled na podatke je integriran direktno u strukturu podatka kroz jednu od dimenzija (npr. koliki je prosjek ocjena svih kolegiya u prvom semestru?) za razliku od smještanja u pojedine kolone u tablicama.



Baze podataka



Usporedba formata



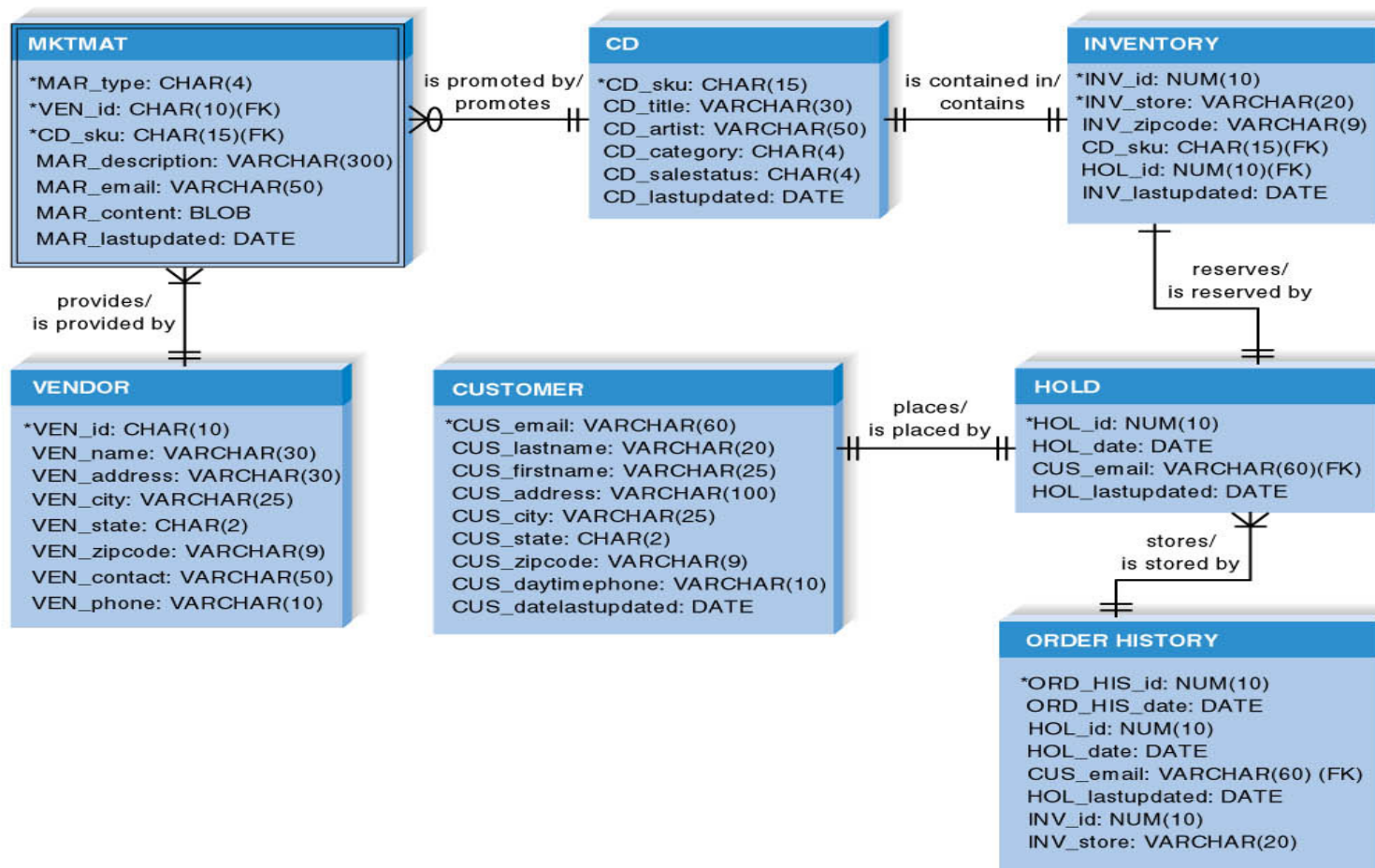
	Files	Legacy DBMS	Relational DBMS	Object-Oriented DBMS	Multi-dimensional DBMS
Major strengths	Files can be designed for fast performance; good for short-term data storage	Very mature products	Leader in the data-base market; can handle diverse data needs	Able to handle complex data	Configured to answer decision support questions quickly
Major weaknesses	Redundant data; data must be updated using programs	Not able to store data as efficiently; limited future	Cannot handle complex data	Technology is still maturing; skills are hard to find	Highly specialized use; skills are hard to find
Data types supported	Simple	<i>Not recommended for new systems</i>	Simple	Complex (e.g., video, audio, images)	Aggregated
Types of application systems supported	Transaction processing	<i>Not recommended for new systems</i>	Transaction processing and decision making	Transaction processing	Decision making
Existing data formats	Organization dependent	Organization dependent	Organization dependent	Organization dependent	Organization dependent
Future needs	Limited future prospects	Poor future prospects	Good future prospects	Uncertain future prospects	Uncertain future prospects
DBMS = database management system.					

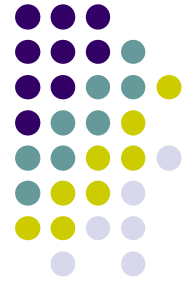
Prelazak sa logičkog modela podataka na fizički model



- ERD dijagram sadrži iste elemente i za logički i za fizički model.
- Fizički ERD dijagram sadrži reference o tome kako su podaci točno fizički pohranjeni (npr. int vrijednost u bazi podataka maksimalne vrijednosti 10000).
- Postupak prelaska sa logičkog na fizički model podataka sastoji se od pet koraka:
 1. pretvaranje entiteta u tablice ili datoteke
 2. pretvaranje atributa entiteta u polja
 3. dodavanje primarnih ključeva
 4. dodavanje sekundarnih ključeva
 5. dodavanje sistemskih podataka
- Nakon dobivanja fizičkog modela podataka potrebno je napraviti reviziju CRUD matrice.

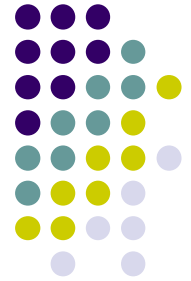
Prelazak sa logičkog modela podataka na fizički model





Optimizacija baze podataka

- Optimizacija baze podataka povećava efikasnosti pohrane podataka, smanjuje brzinu pristupa podacima u bazi, te veličinu baze.
- Osnovni korak u optimizaciji baze je smanjenje redudantnih podataka te smanjenje polja sa *nul* vrijednošću.
- Normalizacija je postupak koji osigurava najmanje redudantnih podataka i polja sa nul vrijednošću.



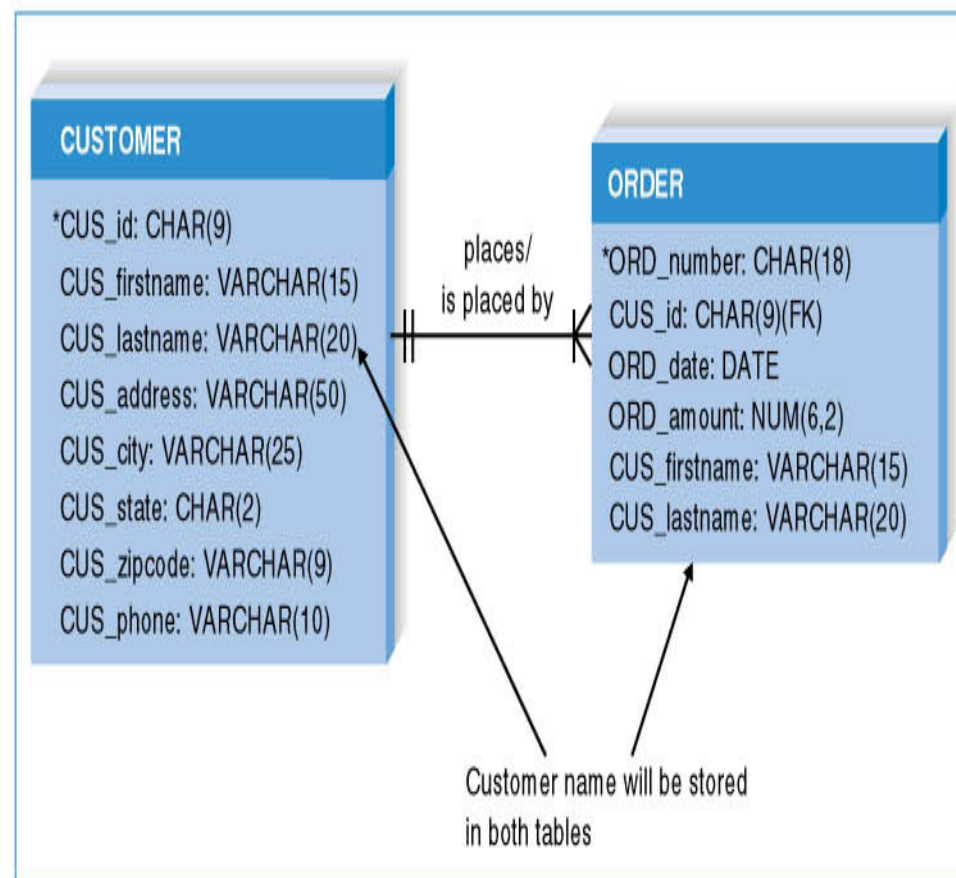
Optimizacija baze podataka

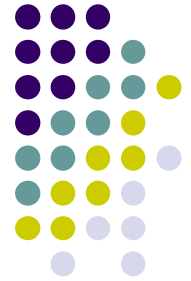
- Nakon optimiziranja dizajna modela podataka radi efikasnosti spremanja podataka, krajnji rezultat su podaci koji su rašireni preko više tablica.
- Postoji više tehnika koje razvojni tim može koristiti kako bi pokušao ubrzati pristup podacima:
 - Denormalizacija
 - Uskupljavanje (*clustering*)
 - Indeksiranje



Denormalizacija

- Postupkom denormalizacije se namjerno unose redudancije u podatke koje ubrzavaju dohvat podataka. Npr. ako imamo česte upite prema narudžbama, a podatke o naručitelju držimo u drugoj tablici (prema normalizaciji) bolje je kopirati podatke iz tablice naručitelja u tablicu narudžbi da ne trebamo “čupati” podatke iz obje tablice.





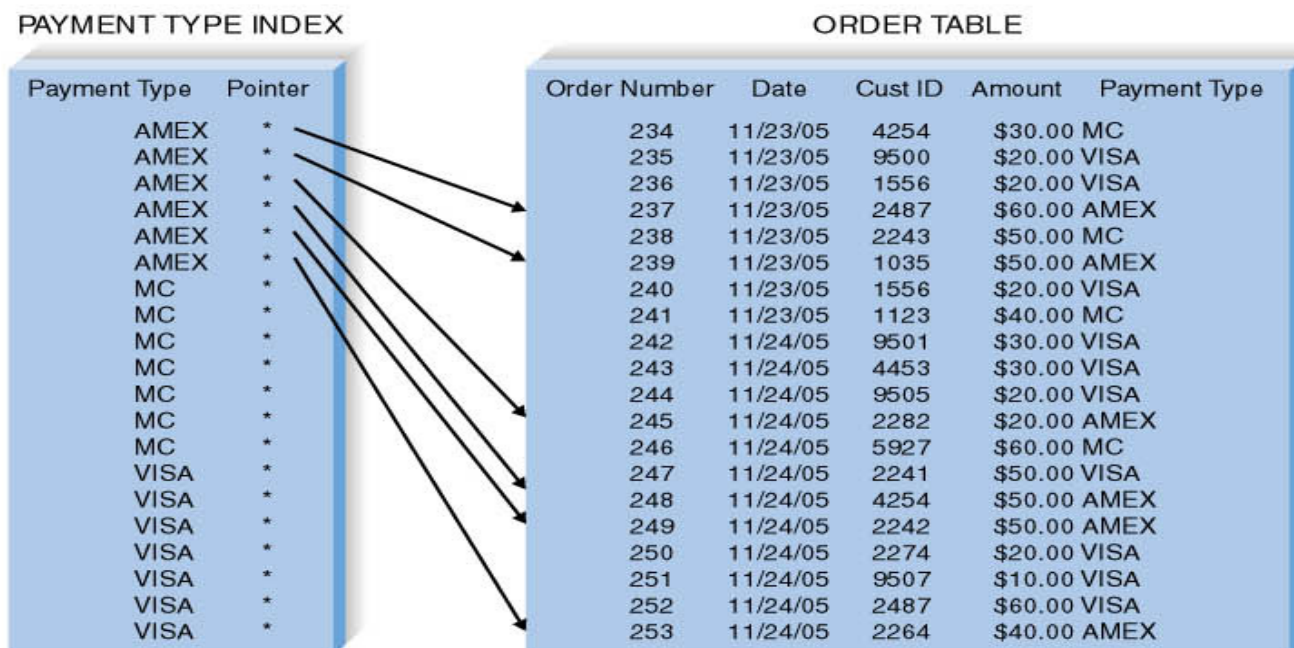
Uskupljavanje

- Reduciranje broja pristupa bazi fizičkim postavljanjem sličnih zapisa jednih blizu drugih:
 - Intrafile clustering – slični zapisi u tablici su spremljeni zajedno
 - Interfile clustering – kombiniranje zapisa koji se tipično dohvaćaju zajedno iz više od jedne tablice



Indeksiranje

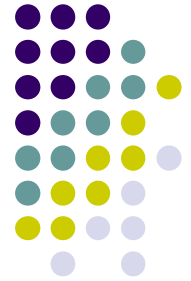
- Relacijske baze omogućavaju indeksiranje podataka.
- Indeksi spadaju u *overhead* relacijske baze podataka. To su dodatni podaci koji se spremaju za promatranu tablicu, ali višestruku ubrzavaju rad sa tablicom.



Indeksiranje



- Preporuke za indeksiranje:
 - Oskudno korištenje indeksa kod transakcijskih sustava.
 - Znatno korištenje indeksa u sustavima za podršku odlučivanja.
 - Za svaku tablicu kreiraj jedinstveni indeks koji je zasnovan na primarnom ključu.
 - Za svaku tablicu kreiraj indeks koji je zasnovan na stranom ključu radi poboljšanja performansi udruživanja.
 - Kreiraj indeks za polja koja se učestalo koriste za grupiranje, sortiranje ili kriterije.



Procjena veličine baze

- Procjena veličine baze podataka radi se na osnovi fizičkog modela podataka.
- Za svaku se tablicu može izračunati točno koliko mjesta zauzima pojedini slog. (npr. int polje zauzima 20 bajtova, varchar(15) 15 bajtova,...). Pored ove vrijednosti potrebno je uzeti u obzir i *overhead* (obično 30% od veličine sloga).
- Kada se izračuna početna veličina baze podataka potrebno je procijeniti i rast broja zapisa u vremenu kako bi se mogla procijeniti veličina baze za 3 mjeseca, 1 godinu, ...



Procjena veličine baze

Field	Average Size (Characters)
Order number	8
Date	7
Cust ID	4
Last name	13
First name	9
State	2
Amount	4
Tax rate	2
Record size	49
Overhead	30%
Total record size	63.7
Initial table size	50,000
Initial table volume	3,185,000
Growth rate/month	1,000
Table volume @ 3 years	5,478,200