

8. Datoteke

Datoteke su važna struktura podataka pa joj zbog toga posvećujemo posebno predavanje. Skoro da nema ni jedne aktivnosti koja se obavlja uz pomoć računara a da se pri tom ne koristi neka datoteka.

Štaje to datoteka?

Datoteka (engleski file – »fajl«) je logički organizovana skupina međusobno povezanih podataka.

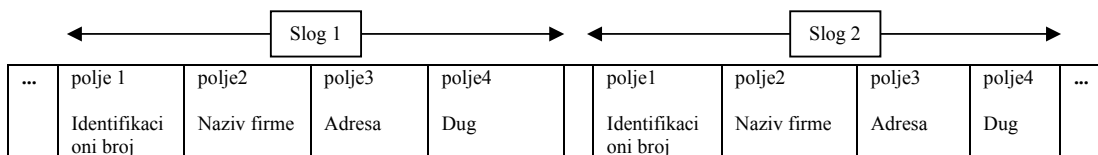
Pod logičkom organizacijom se podrazumeva da je datoteka logički izdvojena na manje delove – slogove koji čine jedinstvene celine podataka. Jedinstvenost se ogleda u tome što se slogovi međusobno razlikuju u bar nekom podatku koji sadrže.

I slogovi sami po sebi jesu strukture podataka, jer se sastoje od više podataka kojima odgovaraju polja (engleski field – »fild«) u slogu.

Pored logičke organizacije koja pokazuje logičke veze medju podacima u datoteci, za potpuno razumevanje koncepta datoteke potrebno je razmatrati i način na koji se podaci iz datoteke memorišu na nekom medijumu (traci, disku, itd...). Metode i tehnike memorisanja datoteka nazivaju se fizičkom organizacijom. Fizička organizacija zapravo pokazuje kako se podaci iz logičke datoteke preslikavaju u fizičke adrese (staze i sektore kod diskova na primer). Mi ćemo se baviti prevashodno logičkom organizacijom datoteka, to jest datoteke ćemo posmatrati sa aspekta programera.

Slogovi

Slogovi su logičke celine međusobno povezanih podataka kojima se modelira onaj entitet koji je predstavljen u datoteci. Tako će recimo datoteka KUPCI sadržati slogove koji sadrže potrebne podatke o kupcima, kao što je prikazano na sledećoj slici:



Kao što se iz predhodnog primera vidi slogovi se sastoje od jednog ili više uzastopnih polja koja sadrže podatke. Polja odgovaraju obeležjima ili atributima koje dodeljujemo entitetu koji se modelira. U predhodnom slučaju entitet su bili kupci, a taj entitet je modeliran atributima: identifikacioni broj, naziv firme, adresa i dug.

Dakle, polja sadrže elementarne podatke i obično su popunjena brojevima i slovima. Moguće su i polja koja sadrže i složenije vidove podataka, kao slike, zvučni zapisi i sl.

Datoteke obično sadrže veliki broj slogova, čiji broj se nekad meri i milionima slogova (za čuvanje biračkih spiskova na primer). Zato je veoma važno da logička i fizička organizacija podataka bude tako uređena da obezbedi efikasno korišćenje i manipulaciju podacima u datotekama.

Koje su osnovne operacije nad datotekama?

Razmotrićmo nekoliko osnovnih operacija koje se sreću pri radu sa datotekama.

Kreiranje datoteke

Kreiranje datoteke je operacija kojom se uspostavlja nova datoteka i priprema za sve naredne operacije. Kreiranjem se obično definišu osnovni podaci o datoteci kao što su: naziv datoteke, struktura slogova – nazivi polja i tipovi podataka u poljima, medijum na kojem se datoteka nalazi (traka, disk, CD, itd.), tip organizacije datoteke (o tipovima organizacije biće reči malo kasnije) itd.

Upisivanje slogova u datoteku

Datoteke se kreiraju sa namerom da u nima budu pohranjeni (memorisani) slogovi. Pohranjivanje slogova u datoteku vrši se operacijama upisa. Da bi se operacija upisa izvršila potrebno je datoteku otvoriti (operacija OPEN), zatim operacijama upisa (operacija WRITE) u datoteku smeštati slogove jedan za drugim i nakraju datoteku zatvoriti (operacija CLOSE) kojom se na kraj datoteke postavlja poseban znak-marker kraja datoteke (end-of-file marker).

Učitavanje slogova iz datoteke

Slogovi koji su predhodno upisani u datoteku mogu biti naknadno učitavani. Pod učitavanjem se ovde podrazumeva proces prenošenja podataka iz sloga na medijumu na kome se datoteka nalazi (traka, disk, CD) u centralnu (RAM) memoriju računara. Operacija učitavanja (READ) zahteva takođe da datoteka bude predhodno otvorena. Zavisno od načina logičke i fizičke organizacije datoteke moguće je slogove čitati u različitom redosledu. O tome će biti govora kasnije. Za sada ćemo smatrati da se slogovi čitaju jedan za drugim onako kako su i bili upisivani. Na kraju čitanja datoteku treba ponovo zatvoriti.

Brisanje slogova

Ponekad je potrebno da se neki već upisani slog (ili više njih) odstrani iz datoteke. Ovu operaciju nazivamo brisanjem sloga (operacija DELETE). Za izvršavanje ove operacije potrebno je otvoriti datoteku, pronaći (locirati) slog koji se briže, izvršiti brisanje i zatvoriti datoteku. Kako se slogovi lociraju? Pa tako što se učitavaju redom slogovi i proverava identifikaciono polje da li odgovara slogu koji treba brisati.

Možda je ovo trenutak da definišemo identifikator sloga (datoteke). U gornjem primeru smo u datoteci KUPCI imali jedno polje – identifikacioni broj, koje je služilo jedinstvenu identifikaciju sloga. To znači da datoteka KUPCI ne može da sadrži dva sloga sa istim identifikacionim brojem. Jedno ili više polja datoteke čiji sadržaj se ne sme ponoviti u drugim slogovima nazivaju se potencijalnim ključevima datoteke.

Ključevi su dakle jedinstveni za svaki slog i mogu služiti za njihovo pretraživanje i lociranje . O upotrebi ključeva biće više reči kada se budu razmatrale različite vrste organizacije datoteka.

Dodavanje slogova

Kada u već postojeću datoteku koja sadrži slogove treba dodati slog sa podacim, tada se operacija dodavanja vrši upisivanjem novog sloga, ato znači da datoteku treba otvoriti, locirati mesto gde će novi slog biti dodat, upisati slog i zatvoriti datoteku. Ponovo imamo pitanje lociranja mesta gde će slog biti upisan. Mesto zapisa novog sloga zavisi od vrste organizacije datoteke, što ćemo razmatrati uskoro.

Promena sadržaja sloga

Ponekad je potrebno izmeniti sadržaj jednog ili više polja u nekom slogu datoteke. Ta operacija je delikatna sa više aspekata. Poseban problem predstavlja promena onih polja koji sačinjavaju ključ datoteke, jer to može da izazove neželjene posledice (pojavljivanje slogova sa istim ključem, narušavanje uređenosti datoteke, i sl.). Sama operacija bi se promene sadržaja bi se mogla odvijati na sledeći način: učitati slog koji se menja, izvršiti zamenu polja, bisati slog iz datoteke, dodati sloga sa izmenjenim poljima u datoteku. Ovakav scenario se najčešće i primenjuje pri promeni (ažuriranju) slogova datoteke.

Organizacija datoteka

Postoji više različitih tipova logičke organizacije datoteka. Svaki od tih tipova ima svoje prednosti i nedostatke i svaki od njih je pogodniji od drugih za neke specifične primene. Pravilan izbor organizacije datoteka presudno utiče na ukupnu efikasnost obrade podataka.

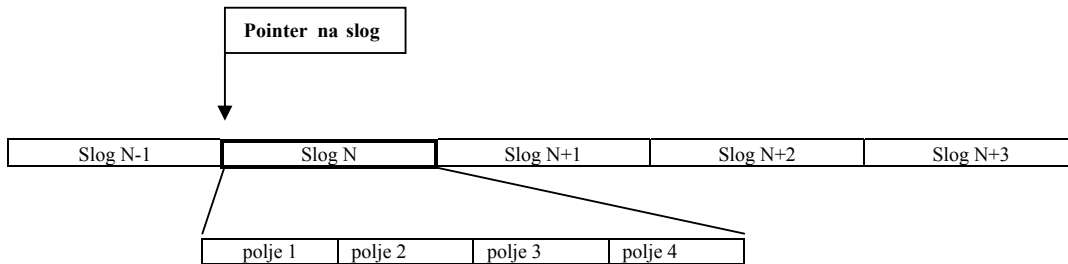
Slede četiri glavne organizacije datoteka koje će biti i ukratko opisane:

- 1) Datoteke sa serijskim pristupom (Serijske datoteke)
- 2) Datoteke sa sekvencijalnim pristupom (Sekvencijalne datoteke)
- 3) Datoteke sa direktnim pristupom (Direktne datoteke)
- 4) Datoteke sa indeks-sekvencijalnim pristupom (Indeks-sekvencijalne datoteke)

Moguće su i druge varijacije ovih osnovnih tipova datoteka, ali mi se njima nećemo baviti.

Serijske datoteke

Osnovna karakteristika serijskih datoteka je da se slogovi u datoteci upisuju jedan za drugim bez nekog logičkog redosleda (najčešće onako kako podaci pristižu u vremenskoj seriji). Takva organizacija se najčešće primenjuje kada se podaci prikupljaju sa raznih lokacija (izvora) tokom nekog vremenskog perioda. Oni se tada najčešće memorišu (upisuju u datoteku) sa namerom da budu obrađeni u nekom kasnijem trenutku (kada pristignu svi očekivani podaci). Sledeća slika ilustruje serijsku organozaciju.



Serijske datoteke se obično obrađuju tako što se slogovi čitaju redom od prvog do poslednjeg. One takođe mogu biti korišćene kao priprema za formiranje uređenih datoteka čiji opis sledi.

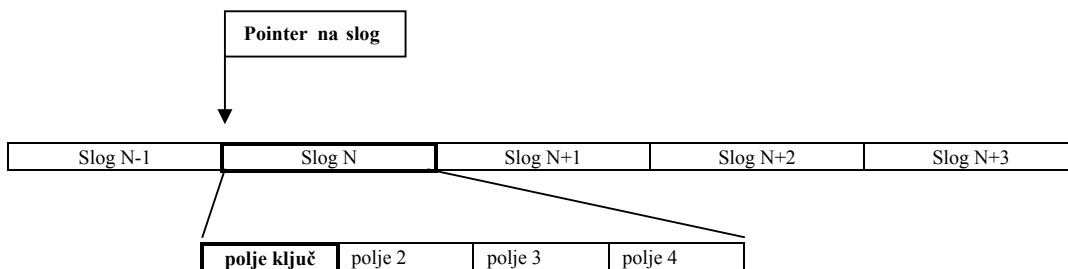
Sekvencijalne datoteke

Kada želimo da datoteka ima neki logičan redosled slogova onda nam sekvencijalne datoteke pružaju jednu takvu mogućnost. Na primer, bankovne račune u datoteci možemo smeštati u rastućem redosledu broja računa, podatke o studentima možemo u slogovima koji su poređani u po azbučnom redosledu imena studenata, ili pak po rastućem redosledu broja indeksa.

Upravo na primeru studenata vidimo značaj ključa datoteke. Dok u školi možemo imati više studenata sa istim imenom i prezimenom, broj indeksa je jedinstven za svakog studenta i predstavlja ključ datoteke.

Sekvencijalne datoteke su jako efikasne za takozvanu beč (batch) obradu podataka, kao što je slučaj kod obračuna računa za komunalne usluge (struja, voda, itd.). Pri beč obradi se pristupa svim slogovima datoteke u redosledu koji je određen ključem, vrši se odgovarajuća obrada podataka iz sloga (obračun potrošnje vode ili stuje) i štampaju izveštaji (računi, fakture, statistički pokazatelji, itd.)

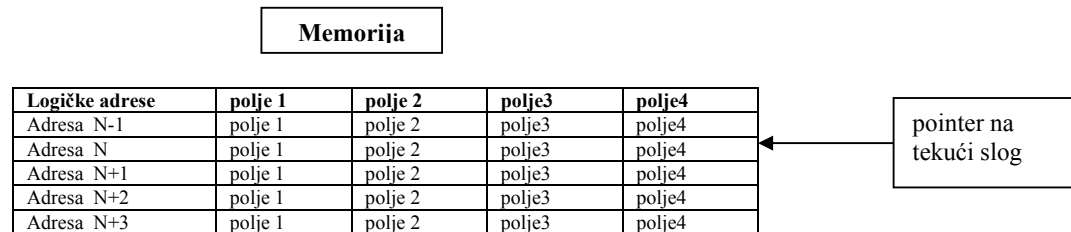
Sledeća slika ilustruje sekvencijalnu organizaciju podataka.



Sekvencijalna organizacija nije pogodna za pojedinačnu obradu slogova. Na primer, ako nam iz neke datoteke koja ima N slogova treba samo jedan slog da bi nad njim izvršili neku operaciju (brisanje, izmena podataka, obrada podataka iz sloga), onda je u proseku potrebno $N/2$ učitavanja slogova da bi učitali željeni slog. Za veliko N (npr. milion) to bi moglo da rezultira u veoma neefikasnom radu.

Direktne datoteke

Direktne datoteke upravo rešavaju problem individualnog pristupa slogovima. Kod direktne organizacije slogovima se pristupa direktno, bez čitanja predhodnika traženom slogu. To se postiže na taj način što se uspostavlja veza između ključa i logičke (fizičke) adrese u memoriji na kojoj se slog sa datim ključem nalazi. Ova vrsta organizacije može se ilustrovati sledećom slikom.



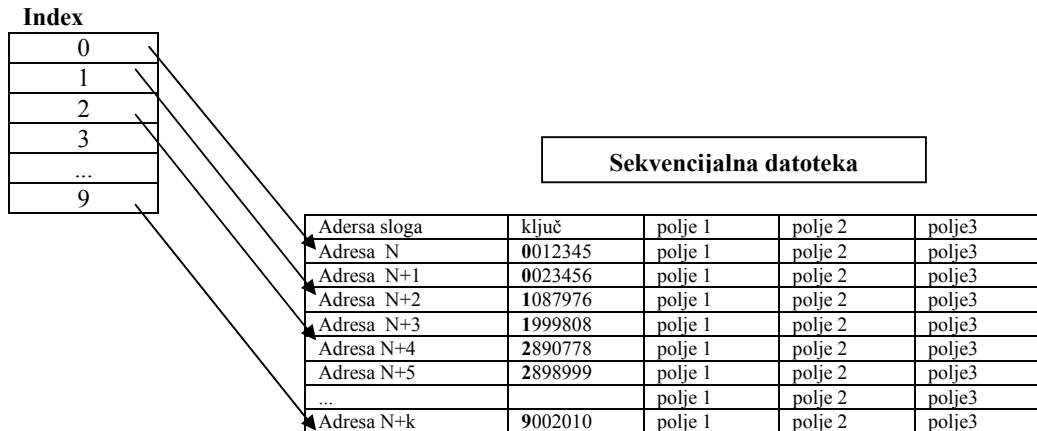
Uspostavljanje veze između ključa i adrese se može vršiti na više načina. Najčešće se koristi hešing metoda o kojoj je već bilo reči pri definisanju heš tabela u predhodnom predavanju. Direktna organizacija zahteva i da uređaj za memorisanje podataka ima mogućnost direktnog pristupa (diskovi imaju tu mogućnost, dok na trakama možemo čuvati samo serijske i sekvencijalne datoteke).

Za razliku od sekvencijalnih datoteka, gde su slogovi poređani u nekom logičkom redosledu, kod direktnih datoteka slogovi su rasuti po memoriji bez ikakvog logičkog reda (već onako kako transformacija ključa u adresu zahteva). Zato su nepogodne za već obradu.

Indeks-sekvencijalne datoteke

Indeks sekvencijalne datoteke imaju za cilj da pomire prednosti i nedostatke sekvencijalnih i direktnih datoteka. Drugim rečima, kod indeks-sekvencijalne organizacije slogovima se može efikasno pristupiti i sekvencijalno i direktno. U cilju postizanja takve organizacije uvode se pomoćne datoteke koje sadrže indekse za lakše pronalaženje slogova (setite se indeksa na kraju knjige koji olakšavaju pronalaženje karakterističnih reči koje se u knjizi pojavljuju).

Indeks-sekvencijalna datoteka se zapravo sastoji od jedne sekvencijalne (po ključu uređenje datoteke) i jedne pomoćne datoteke indeksa, kao što to ilustruje sledeća slika.



Ovakvom organizacijom se obezbeđuje da se slogovima datoteke može pristupati sekvencijalno jer su slogovi upisani u sekvencijalnu datoteku u recimo rastućem reodsledu ključa (kao na slici). S druge strane, kad želimo da pristupimo jednom slogu sa određenim ključem, tada najpre pogledamo kojem indeksu tak ključ odgovara pa sledeći pointer iz indeksne datoteke pristupamo grupi slogova u kojoj se može naći slog sa traženim ključem. Unutar te grupe pristupamo slogovima redom, pa ili pronađemo traženi slog ili zaključimo da takav slog (sa zadatim ključem) nije u datoteci.

Datoteke su obično u novije vreme zapravo najčešće delovi baze podataka. Drugim rečima baze podataka se obično sastoje od više datoteka uređenih na jedan od predhodno opisanih načina. Zato i softver za podršku radu sa bazama podataka, tzv DBMS (database management software) kao što je Oracle, SQL ili neki drugi, u sebi sadrže sve potrebne elemente za kreiranje, brisanje, modifikaciju kao i obradu podataka u datotekama koje čine bazu. Programski jezici obično ne sadrže takve mogućnosti kao inherentne karakteristike jezika, već obično uz takve jezike idu biblioteke potprograma (klasa, funkcija, procedura) koje programerima omogućavaju manipulaciju podacima u bazama podataka. To su tzv. ODBC (Open Database Connectivity) drajveri.

Kako su datoteke (u bazama podataka) osnovni resurs koji sadrži velike i značajne podatke u svakoj poslovnoj strukturi, to njihova organizacija, ažurnost i obrada značajno doprinose ukupnoj produktivnosti poslovnog sistema. Danas, kada poslovni sistemi prelaze ne samo granice zemalja već i kontinenta, nije redak slučaj da se datoteke jedne kompanije nalaze na više lokacija širom sveta i da obuhvataju velike količine informacija (reda tetra bajta).

Zbog značaja koje podaci imaju u poslovanju svake kompanije posebna pažnja se poklanja integritetu, sigurnosti i zaštiti podataka (datoteka).

Pod integritetom se podrazumeva sprečavanje grešaka na unosu, grešake u proceduri, programskih grešaka, virusa, grešake u prenosu. To se postiže permanentnom validacijom i verifikacijom podataka kojom se utvrđuje integritet, kao i raznim metodama za uspostavljanje narušenog integriteta.

Sigurnost datoteka se bavi pitanjima slučajnih oštećenja, namernih oštećenja, ilegalnog pristupa i sl. Povećanje sigurnosti podataka postiže se različitim merama kao što su ograničenje pristupa podacima putem lozinki, strogo poštovanje procedura za promenu podataka, čuvanje kopija (back-up), firewall-ove, pa sve do fizičke zaštite pristupa prostoru i opremi na kojoj se čuvaju podaci.

Zaštita datoteka uključuje takođe i kriptografiju i antivirusne mere.

Detaljan opis nekih od najčešćih procedura kao što su sortiranje i merđovanje datoteka, kao i primere obrade možete naći u knjizi »Projektovanje programa«, N.Marković, BPS 2001.

Pitanja

1. Definišite sledeće izraze povezane sa datotekama: Character, Field, Record, File.
2. Po čemu se datoteka razlikuje od drugi struktura podataka?
3. Šta je to slog?
4. Šta je polje?
5. Koje su četiri osnovne organizacije podataka?
6. Šta je serijska datoteka?
7. Šta je sekvencijalna datoteka?
8. Šta je ključ datoteke?
9. Koji tipovi datoteke se mogu organizovati na magnetnoj traci?
10. Šta je direktna datoteka?
11. Šta je indeks-sekvencijalna datoteka?
12. Koje se tipične operacije izvršavaju nad datotekama?
13. Kako se kreira datoteka?
14. Šta znači ažurirati datoteku?
15. Kakvi se problemi mogu javiti pri ažuriranju sekvencijalne datoteke?
16. Šta znači sortiranje datoteke?
17. Šta je merđovanje datoteka?
18. Šta su master i transakcione datoteke?
19. Kako se može sprečavati neautorizovan pristup podacima?
20. Kako se može detektovati neautorizovan pristup podacima?
21. Kako se podaci mogu zaštititi od uništenja?
22. Kako kriptografija pomaže zaštitu podataka?
23. Kako virusi mogu uticati na podatke?