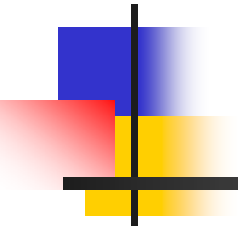
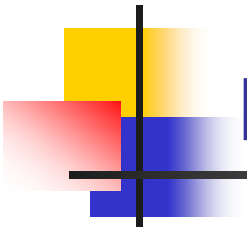


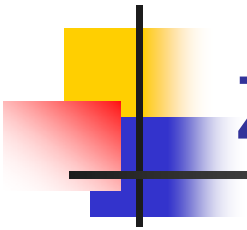
XML baze podataka

Jelena Tomašević
Matematički fakultet, Univerzitet u Beogradu
jtomasevic@matf.bg.ac.yu
www.matf.bg.ac.yu/~jtomasevic



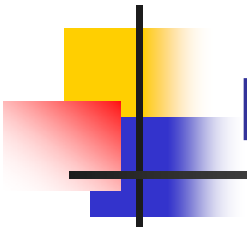
Pregled

- Zašto su dobro organizovani podaci značajni?
- Zašto nam trebaju nove baze podataka?
- Šta su izvorne XML baze podataka?
- Koje su prednosti u smeštanju polustruktuiranih podataka u izvornim XML bazama u odnosu na relacione baze?
- Kako funkcioniše postavljanje upita u XML bazama podataka?



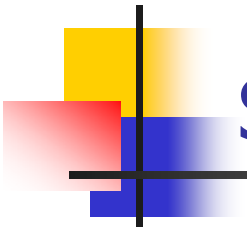
Značaj podataka

- Dobro organizovani podaci nisu dovoljno cenjeni koliko bi trebalo s obzirom na njihov značaj.
- Oni su loše organizovani, teško je upravljati njima i uglavnom su nedovoljno iskorišćeni.
- Rast količine podataka.
- Snažniji računari.
- Očekivanje: dobijanje kvalitetnijih informacija.
- Kvalitetnije informacije – prednost u odnosu na konkurenciju.



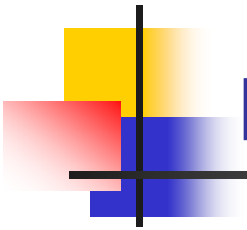
Relacione baze podataka

- Smeštanje struktuiranih podataka.
- Ograničenja modela opisana su shemom.
- Koristi indekse radi poboljšanja performansi upita.
- Standardni upitni jezik - SQL.



Šta su polustruktuirani podaci?

- Podaci koji nemaju zajedničku strukturu.
- Podaci koji mogu sadržati polja koja nisu poznata pre trenutka projektovanja dokumenta.
- Podaci kod kojih se iste vrste podataka mogu predstaviti na različite načine.
- Relacione baze podataka nisu najpodesnije.

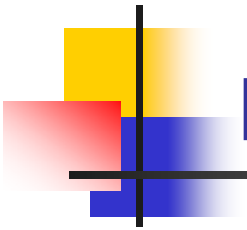


Predstavljanje polustruktuiranih podataka

- XML je jezik za predstavljanje polustruktuiranih podataka.
- XML je skraćenica od eXtensible Markup Language.
- XML je jezik označavanja sličan HTML-u.
- XML je razvijen od strane W3C (World Wide Web Consortium) sa ciljem prevazilaženja nedostataka HTML-a.
- XML nije zamena za HTML. U okviru budućeg razvoja veba, tendencija je da se HTML koristi za prikazivanje a XML za opisivanje podataka.

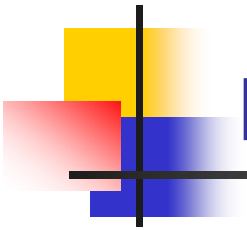
Prednosti XML-a

- Sintaksa etiketa nije fiksirana.
- Ne mora se definisati shema.
- Fleksibilan je i proširiv, omogućava postojanje različitih tipova podataka u okviru jednog dokumenta.
- Opisuje podatke stavljanjem akcenta na to šta podaci jesu a ne kako oni izgledaju.
- Ima format koji je čitljiv za čoveka.
- Ima mogućnost internacionalnog korišćenja zahvaljujući činjenici da koristi Unicode kodnu šemu.
- Nezavistan je od platforme odnosno od softvera i hardvera koji se koristi.
- Postoji veliki broj gotovih aplikacija za procesiranje XML-a koje se mogu koristiti.



Nedostatci XML-a

- Ne mora se definisati shema.
- Fleksibilan je i proširiv.
- Manipulisanje podacima često sporije.
- Optimizacija je kompleksnija zahvaljujući bogatstvu i velikoj izražajnoj moći upitnih jezika koje koristi.



Da li je XML baza podataka?

- Ako se pod bazom podataka podrazumeva bilo kakva kolekcija podataka, XML u tom striktnom smislu, može se smatrati bazom podataka.
- Da li XML i tehnologije koje ga okružuju mogu pretstavljati sistem za upravljanje bazama podataka?
  - XML može obezbediti skladište podataka (XML dokumenti), sheme (DTD, XML sheme), upitne jezike (XQuery, XPath, XQL, XML-QL), interfejse.
  - Nedostaje: efikasno skladištenje, indeksi, bezbednost, transakcije i integritet podataka, pristup od strane više korisnika, upiti nad više dokumentata odjednom.

XML dokumenti

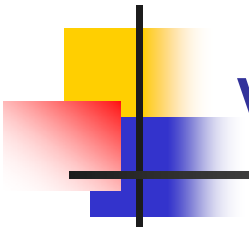
- XML dokumenti upadaju u dve široke kategorije:
  - **Data-centric**
    - Karakterišu se regularnom strukturom, fino zrnastim podacima bez mešanog sadržaja.
    - Oni su projektovani tako da se koriste uglavnom za obradu od strane mašina pre nego za ljudsku upotrebu.
    - Ovakvi podaci su najčešće smešteni u nekoj relacionoj bazi podataka i javlja se potreba za transferom podataka iz relacione baze u XML dokument, iz XML dokumenta u relacionu bazu podataka ili u oba smera.
  - **Document-centric**
    - Karakterišu se manje regularnom ili neregularnom strukturom, krupno zrnastim podacima i ima dosta mešanog sadržaja.
    - Dokumenti projektovani uglavnom za ljudsku upotrebu.
    - Redosled elemenata često nije od značaja.
    - Najčešće su ručno pisani u XML-u.

Primer data-centric dokumenta

```
<meni datum='5.10.2007'>
        <jelo>
            <ime>Domaca pileca supa</ime>
            <cena>100.00</cena>
            <kalorije>650</kalorije>
        </jelo>
        <jelo>
            <ime>Francuski tost</ime>
            <cena>30.50 din</cena>
            <kalorije>300</kalorije>
        </jelo>
        <jelo>
            <ime>Bakin dorucak</ime>
            <cena>200.00 din</cena>
            <kalorije>350</kalorije>
        </jelo>
</meni>
```

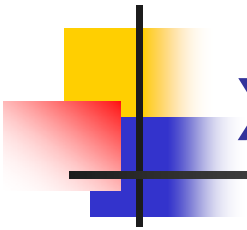
Primer document-centric dokumenta

<Proizvod>
<Ime>KIRKOLINA – čaj za mršavljenje</Ime>
<Proizvodjac>Kirka-Pharma</ Proizvodjac>
<Opis>
<Paragraf> Predstavlja mešavinu lekovitog bilja koje kombinovanim dejstvom regulišu promet materija u organizmu, ubrzavaju sagorevanje masnih naslaga i utiču na smanjenje telesne težine. Krušina, sena, zova stimulišu metabolizam, podstiču probavu, smanjuju nadutost. Breza, pirevina, rastavić eliminišu nakupljene toksične materije, poboljšavaju cirkulaciju. Matičnjak oslobađa od stresa koji je često uzrok nekontrolisanog konzumiranja hrane. Žalfija kao izuzetni antiseptik štiti od mogućih infekcija i utiče na jačanje organizma. </Paragraf>
<Paragraf>Možete:</Paragraf>
<List><Item><Link URL="Naruci.html">Naručiti svoj čaj za mršavljenje</Link></Item>
<Item><Link URL="Kirkolina.htm">Pročitati više o ovom proizvodu</Link></Item>
<Item><Link URL="Katalog.zip">Skinuti katalog naših proizvoda</Link></Item></List>
<Paragraf>Ovaj čaj košta samo 500 dinara.</Paragraf>
</Opis>
</Proizvod>



Vrste XML baza podataka

- XML-proširene baze podataka.
  - Koristi postojeći sistem za upravljanje bazama podataka.
  - Preslikava XML podatke u sopstveni model.
    - Čuvaju se hijararhija i podaci.
    - Gubi se identitet dokumenta, redosled čvorova na istom nivou,...
  - Data-centric dokumenta
- Izvorne XML baze podataka.
  - Baze podataka koje smeštaju XML u "izvornom" obliku održavajući prirodnu drvoliku strukturu ovih dokumenata.
  - Document-centric dokumenta.



XML-proširene baze podataka

- Problemi koji se mogu javiti pri prenosu podataka između XML dokumenata i baze.
  - Tipovi podataka.
  - Nedostajuće vrednosti.
  - UNICODE podrška.
  - Procesirajuće instrukcije i komentari.
- Nije omogućeno kružno putovanje dokumenta (round trip).
- Polustruktuirani podaci smešteni u relacionim bazama imaće kao rezultat veliki broj nedostajućih vrednosti ili veliki broj tabela.

Više nedostajućih vrednosti ili više tabela

```
<lista>
  <osoba>
    <ime>Nikola</ime>
    <godine>15</godine>
    <majka>Ljubica< /majka >
    <otac>Vojislav</otac >
  </osoba>
  <osoba>
    <ime>Stefan</ime>
    <roditelj>Mirko</roditelj>
  </osoba>
  <osoba>
    <ime>Milan</ime>
  </osoba>
</lista>
```

ime	godine	majka	otac	roditelj
Nikola	15	Ljubica	Vojislav	null
Stefan	null	null	null	Mirko
Milan	null	null	null	null

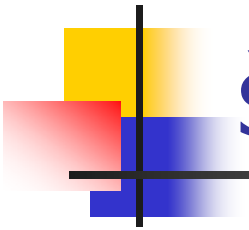
id	ime
1	Nikola
2	Stefan
3	Milan

id	majka
1	Ljubica

id	godine
1	15

id	otac
1	Vojislav

id	roditelj
2	Mirko



Šta su izvorne XML baze podataka?

Osnovne osobine izvornih XML baza podataka:

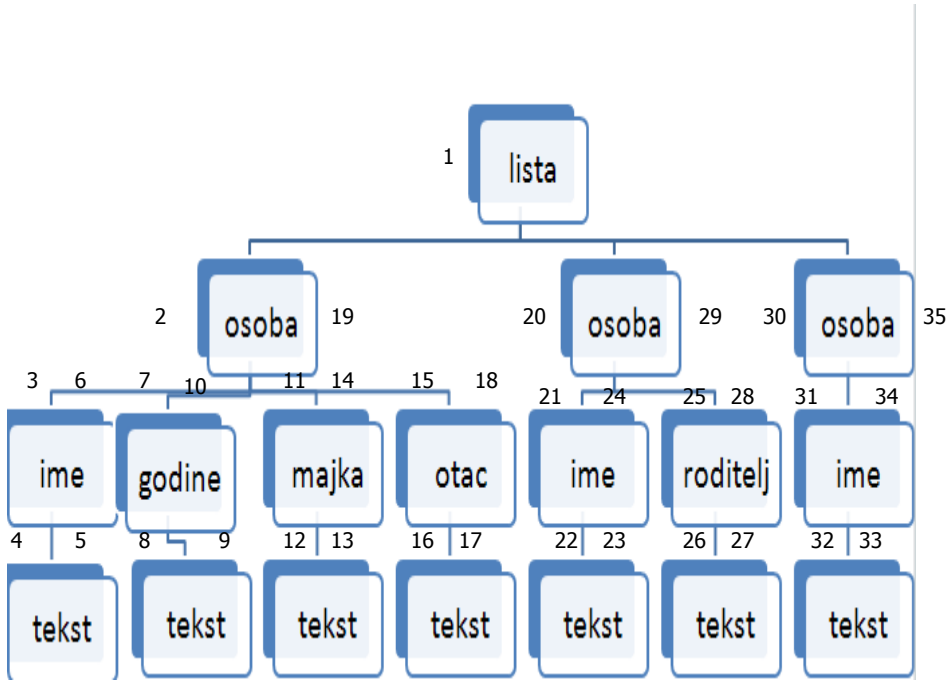
- XML dokument je osnovna logička jedinica, kao što je to vrsta u tabeli kod relacionih baza.
- Minimalno, model mora uključiti elemente, atribute, tekstualne podatke (PCDATA) i redosled dokumenta.
- Nema zahteva za postojanjem bilo kakvog specifičnog fizičkog modela skladištenja.

Arhitektura izvornih XML baza

- Arhitektura bazirana na tekstu.
  - XML podaci se čuvaju kao tekst.
  - Izuzetno dobro indeksiranje koje omogućava brzo referenciranje bilo kog XML dokumenta ili nekog njegovog dela.
  - Izuzetne performanse kada se podacima pristupa u skladu sa predefinisanom hijerarhijom.
- Arhitektura bazirana na modelu.
  - Interni objektni model na osnovu XML dokumenta.
  - Najčešće su izgrađene tako da koriste neku relacionu bazu kao sredstvo za fizičko skladištenje podataka. U tom slučaju performanse ovih baza u velikoj meri zavise od relacionih baza koje rade u pozadini.

Arhitektura bazirana na modelu - primer

```
<lista>
  <osoba>
    <ime>Nikola</ime>
    <godine>15</godine>
    <majka>Ljubica< /majka >
    <otac>Vojislav</otac >
  </osoba>
  <osoba>
    <ime>Stefan</ime>
    <roditelj>Mirko</roditelj>
  </osoba>
  <osoba>
    <ime>Milan</ime>
  </osoba>
</lista>
```



Arhitektura bazirana na modelu - primer

dokumenti

dok_id	koren_id
1	1

elementi

element_id	dok_id	dubina	roditelj_id	prethodni_brat	sledeci_brat	prvo_dete
1	1	1	null	null	null	2
2	1	2	1	null	20	3
3	1	3	2	null	7	4
4	1	4	3	null	8	null
...	...	...	...	...	...	...

ime elementa

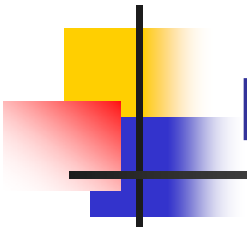
element_id	ime
1	lista
2	osoba
3	ime
...	...

vrednost elementa

element_id	vrednost
4	Nikola
8	15
...	...

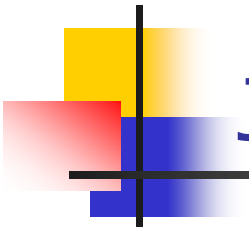
Osobine izvornih XML baza podataka

- Kolekcije dokumenata.
- Ažuriranje i brisanje baze.
- Transakcije, zaključavanje i konkurencija.
- Programski interfejsi (Application Programming Interfaces — APIs). Najpoznatiji su XML:DB API i XQJ (XQuery API for Java).
- Kružno putovanje (Round-Tripping).
- Indeksi.
  - Vrednosni indeksi - "Pronaći sve elemente čija je vrednost Santa Cruz"
  - Strukturalni indeksi -"Pronaći sve City elemente čija je vrednost Santa Cruz"
  - Full-text indeksi - "Pronaći sva dokumenta koja sadrže reč Santa Cruz"



Kako izabrati najbolje rešenje

- Pri izboru baze podataka prvo pitanje na koje treba dati odgovor jeste ”Koji je razlog zbog koga se želi koristiti baza podataka kao i na kakav način se ta baza želi koristiti?”
- Možda najznačajniji faktor u izboru baze podataka je da li će se u nju smeštati podaci ili dokumenti.
- Ako se podacima pristupa u skladu sa predefinisanom hijararhijom ili ako nije definisana shema dokumenta, izvorne baze podataka imaju prednost.
- Ako se očekuju česta ažuriranja onda izvorne baze nisu najbolje rešenje.
- Neke upite je lakše postaviti nad izvornim XML bazama nego nad relacionim bazama.



Jezici za postavljanje upita

- Veliki broj jezika je kreiran za postavljanje upita nad XML dokumentima uključujući XML-QL, XPath, XQL, XQuery.
- XPath je W3C preporuka a sa pojavom XQuery postaje još popularniji. Koriste se za dobijanje i manipulisanje podacima iz XML baza podataka.
- Veliki broj ugrađenih funkcija.
- Korisnik ima mogućnost definisanja sopstvenih funkcija.
- Indeksi su potrebni radi efikasnijeg izvršavanja upita nad velikim kolekcijama dokumenata.

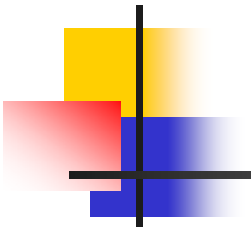
XPath

- Xpath koristi iskaze putanja za kretanje kroz logičku, hijararhijsku strukturu XML dokumenta.
- Dizajniran je da radi sa jednim XML dokumentom. Vrednost vraćena XML upitom je skup čvorova.
- Primer: knjiga//odeljak/naslov

Za sve putanje koji počinju od elementa knjiga, ispitati da li se u okviru njih nalazi element odeljak i vratiti kao rezultat element naslov koji je dete elementa odeljak.

XPath - primeri

- Selektovati sve elemente starost u dokumentu.
  //starost
- Selektovati sve elemente koji su deca korenog elementa student
  /student/*
- Selektovati sve studbr atribute elemenata student u dokumentu.
  /student[@studbr]
- Selektovati sve elemente starost.
  //*[name()='starost']
- Selektovati sve pretke od svih elemenata starost koji su deca od elementa student.
  /student/starost/ancestor::*



XQuery — XML Query Language

- XQuery je jezik koji je projektovan da bude mali, da se lako implementira i da bude lako razumljiv jezik.
- On je nastao sa idejom da obezbedi upitni jezik koji ima istu širinu funkcionalnosti kao SQL nad relacionim bazama podataka.
- To je funkcionalni jezik u kome je svaki upit iskaz.
- Iskazi u XQuery-u upadaju u 6 širokih tipova:
  - Iskazi putanje.
  - Konstruktori elemenata.
  - FLWR iskazi.
  - Uslovni iskazi.
  - Kvantifikovani iskazi.
  - Iskazi koji u sebi uključuju korisnički definisane funkcije.

XQuery – iskazi putanje

- XQuery obezbeđuje iskaze putanja koje su nadskup od onih u XPath-u.
  - Iz dokumenta koji sadrži zapslene i njihovu mesečnu zaradu, izdvojiti godišnju zaradu za zaposlenog sa imenom Marko.

    //zaposleni[ime="Marko"]/zarada * 12

  - U dokumentu ”zoo.xml” pronaći sve slike u poglavljima od 2 do 5.

    document("zoo.xml")//poglavlje[2 TO 5]//slika

XQuery – konstruktori elemenata

- Ponekad je neophodno za upit da kreira ili generiše elemente. Takvi elementi se mogu generisati direktno u upitu u okviru iskaza nazvanog konstruktori elemenata.
  - Generisati elemente <zaposleni> koji imaju zapid atribute. Vrednost atributa i sadržaj elementa su specifikovani promenljivom $id koja je dodeljena u nekom drugom delu upita.

```
<zaposleni zapid = {$id}>
        {$ime}
        {$posao}
</zaposleni>
```

Xquery – FLWR iskazi

- FLWR se izgovara kao ”flower”.
- Ovaj iskaz je upit koji se sastoji od FOR, LET, WHERE I RETURN klauze.
  - Izlistati sve izdavače koji su izdali više od 100 knjiga.

```
<veliki_izdavaci>
{
      FOR $p IN distinct(document("bib.xml")//izdavac)
      LET $b := document("bib.xml")//knjiga[izdavac = $p]
      WHERE count($b) > 100
      RETURN $p
 }
 </ veliki_izdavaci >
```

Xquery – uslovni iskazi

- Uslovni iskazi ocenjuju test iskaze i onda vraćaju jedan od dva rezultujuća iskaza. Ako je vrednost test iskaza tačno onda se vraća kao rezultat vrednost prvog rezultujućeg iskaza, u suprotnom, vraća se vrednost drugog.
  - Napraviti listu svih knjiga uređenih po naslovu. Za beletristiku, uključiti izdavača a za sve ostale autora.

```
FOR $k IN //knjiga  RETURN
<knjiga>
{$k/naslov, IF ($k[@zanr = "Beletristika"]) THEN $k/izdavac ELSE $k/autor}
</knjiga>
SORTBY (naslov)
```

Xquery – kvantifikovani iskazi

- SOME klauza i EVERY klauza - ekvivalentne kvantivikatorima koji se koriste u matematici i logici.
  - Pronalaći naslove svih knjiga u kojima su ”jedrenje” i ”surfovanje” pomenuti u nekom paragrafu.

```
FOR $k IN //knjiga
WHERE SOME $p IN $b//paragraf SATISFIES
(contains($p, "jedrenje") AND contains($p, "surfovanje"))
RETURN $k/naslov
```

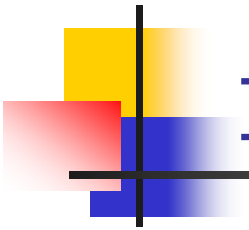
  - Pronalaći naslove knjiga u kojima se ”jedrenje” pominje u svakom paragrafu.

```
FOR $k IN //knjiga
WHERE EVERY $p IN $k//paragraf SATISFIES
contains($p, "jedrenje")
RETURN $k/naslov
```

Xquery – iskazi koji u sebi uključuju korisnički definisane funkcije

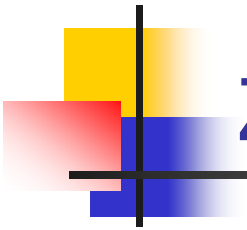
- Osim toga što je podržana centralna biblioteka funkcija sličnih onima u XPath-u, XQuery takođe daje mogućnost korisnicima da definišu funkcije koje će proširiti ovu biblioteku.
  - Napisati funkciju koja za knjigu za koju su date informacije o ceni i popustu (u procentima), izračunati cenu sa popustom.

```
declare function local:minCena($cena as xs:decimal?, $popust as
xs:decimal?) AS xs:decimal?
{
let $pop := ($cena * $popust) div 100
return ($cena - $pop)
};
(: Primer poziva ove funkcija je::)
<minCena>{local:minCena($knjiga/cena,$knjiga/popust)}</minCena>
```



Izvorne XML baze podataka

- Najpoznatiji sistemi za upravljanje izvornim XML bazama podataka su:
  - eXist
    - Open source sistem za upravljanje izvornim XML bazama podataka, koji jednostavno može biti integrisan u druge aplikacije koje koriste i obrađuju XML.
    - Baza podataka je potpuno napisana u Javi.
  - Berkeley DB XML
  - Oracle XML DB
  - MarkLogic Server, izvorna XML baza podataka koja koristi XQuery.



Zaključak

- Podaci nisu dovoljno vrednovani s obzirom na značaj koji imaju.
- Relacione baze podataka su pogodne za smeštanje dobro struktuiranih podataka ali ne i za nestruktuirane i polustruktuirane podatke.
- Izvorne XML baze podataka su pogodne za polustruktuirane podatke ali su one manje razvijene od relacionih baza.
- Izvorne XML baze variraju u načinima modeliranja i smeštanja podataka.
- Postoje dve osnovne arhitekture izvornih XML baza podataka: bazirana na tekstu i bazirana na modelu.
- Razlog i način korišćenja baze podataka je osnova za određivanje vrste baze koju treba koristiti.

Literatura

- Bourett, Ronald. XML and Databases. Available at: http://www.rpbourret.com/xml/XMLAndDatabases.htm
- Dare Obasanjo. An Exploration Of XML In Database Management Systems. Available at: http://www.uni-weimar.de/~bauinf/lehre/Bi_3/Vorlesung/Obasanjo_XMLandDatabases.pdf
- Gordana Pavlović-Lažetić. NATIVE XML DATABASES vs. RELATIONAL DATABASES IN DEALING WITH XML DOCUMENTS.