Smery v databázovom výskume. (Veda aj kuchárky)

Ján Šturc
KI-FMFI-UK
Január, 2005

Tri pohľady

1. Matematický prístup.
   - Lepšia a úplnejšia formalizácia relačného modelu
2. Na čo sa orientuje praktický výskum?
   - Viac a väčšie dáta.
   - Optimalizácia dotazov.
3. Nové „módne” smery.

Relačný model

- Prvá a zatiaľ jediná matematická formalizácia databázovej teórie.
  - Vzťah k predikátovému kalkulu a matematickej teórii modelov.
- Čo zostalo od Codda (1972) nedoriešené ?
  - Nekonečné matematické relácie
    - Napr.: aritmetické funkcie
  - Agregačné funkcie

Je to vôbec potrebné

- Asi mnohí si myslia, že nie.
  - To si matematici hrajú na svojom piesočku.
- Myslím si, že áno.
  - Dopad na optimalizáciu napr. :
    - $\sigma_{z=x+y}(R(x,y)\times S(z))$ a $R(x,y,z)\bowtie(z=x+y)\bowtie S(z)$
    - Optimalizácia zložených dotazov s agregačnými funkciami napr.: poradie agregácie a joinu.

Relačné jazyky

- Predikátový kalkul
  - Zápis mnohých dotazov je zbytočne komplikovaný.
  - Okrem matematikov je pre uživateľa ťažký.
- SQL je dosť neprirodzený, ale je to norma
  - Správnosť mnohých dotazov sa dá roznať až počas behu programu.
  - Rozšírenia sú pridané neorganicky.

Nové smery

1. Integrácia informácii.
2. Olap a dátové kocky.
3. Spracovanie prúdov.
4. Semištruktúrované dáta a XML.
5. Partnerské a grid databázy.
6. Dolovanie z dát.

Čo je rozhodujúce?

- Spracovať podľa možnosti, čo najväčšie množstvo dát. (vraj je to zrejmé)
- Používať jazyky veľmi vysokej úrovne.
  - S veľkým objemom dát treba narábať uniformne.
- Optimalizácia dotazov.
  - Príklad APL (nerozšírilo sa), SQL (všade).

Nové smery

1. **Integrácia informácii**.
2. Olap a dátové kocky.
3. Spracovanie prúdov.
4. Semištruktúrované dáta a XML.
5. Partnerské a grid databázy.
6. Dolovanie z dát.

Integrácia informácii

- Rôzne zdroje dát rovnakej povahy treba vidieť ako jeden celok.
- Príklad: katalógy (výrobky viacerých výrobcov), digitálne knižnice, vedecko výskumné dáta, podnikové informácie o zdrojoch, prostriedkoch, výrobe a distribúcii a pod..

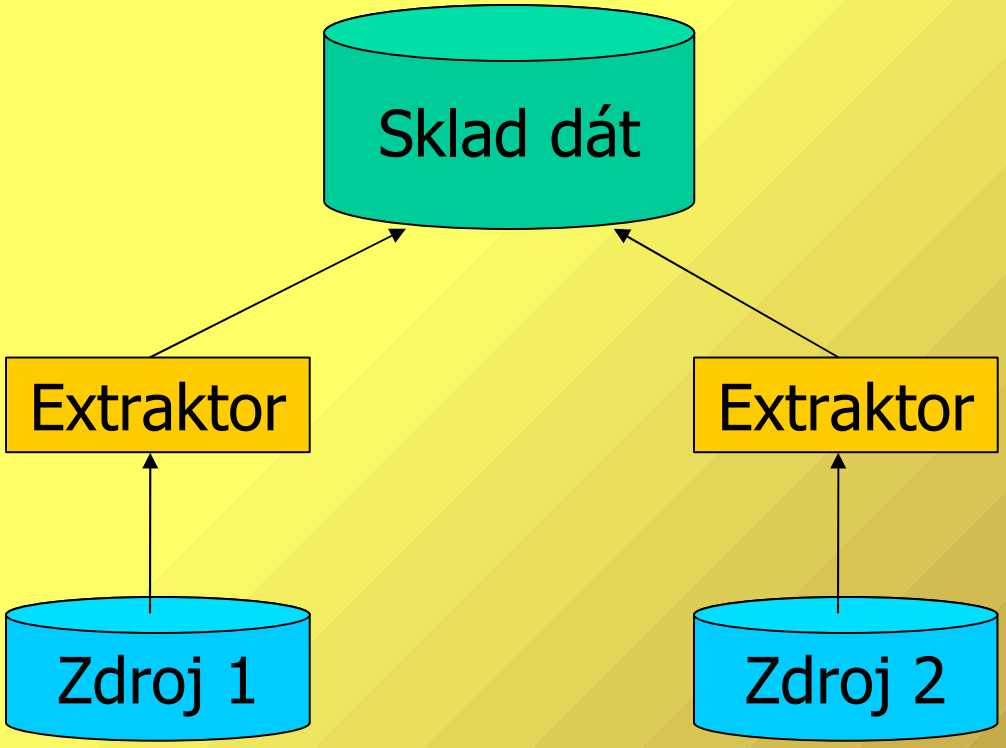
Globálna a lokálne schémy

- Každá databáza má vlastnú *lokálnu schému* = spôsob reprezentácie a organizácie dát.
- Integrácia vyžaduje *globálnu schému* a mechanismus prekladu medzi globálnou a všetkými lokálnymi schémami.

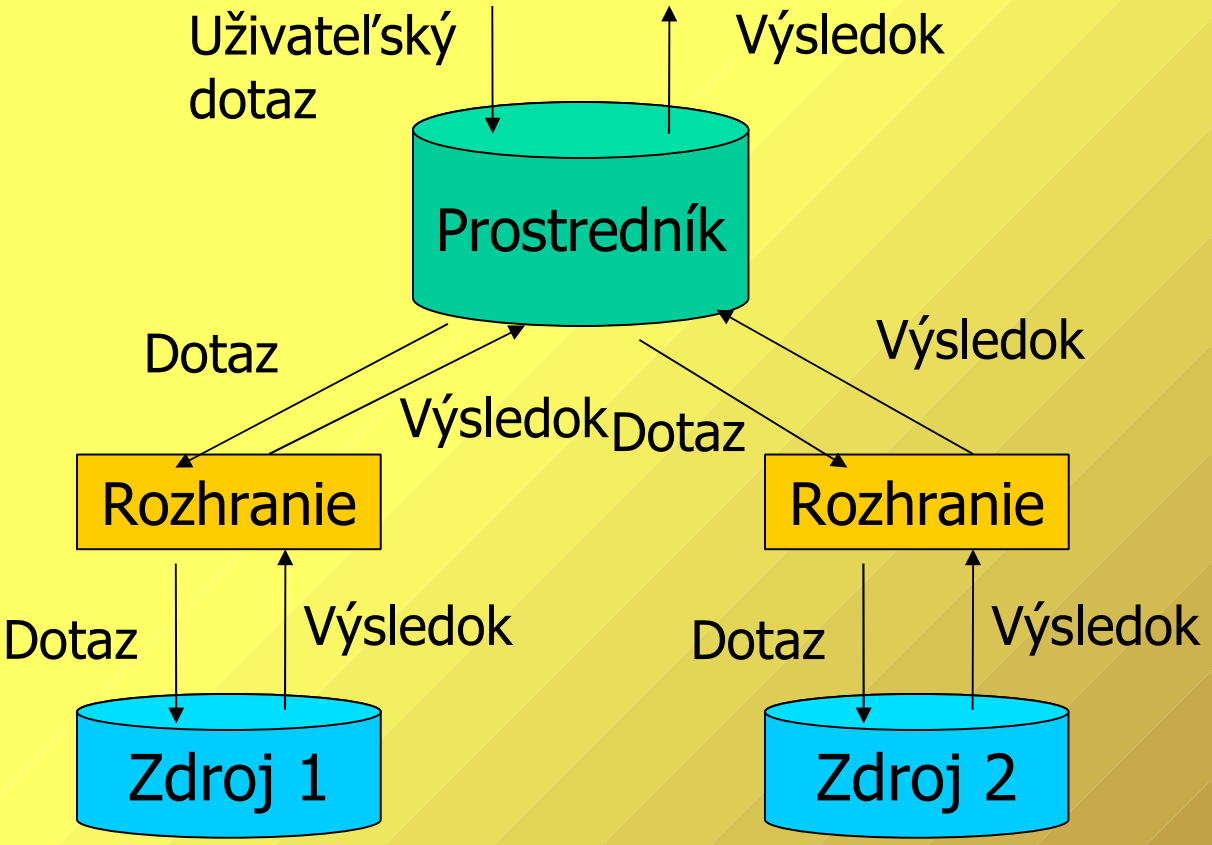
Dva prístupy

1. *Skladovanie - off line*
   - Periodicky zbierať dáta do globálnej databázy - „dátového skladu” .
   - Dotazy sa spracujú v sklade, zdrojové databázy spracovávajú transakcie nezávisle.
2. *Sprostredkovanie - on line*
   - Dotazy sa spracovávajú prekladom z globálnej schémy do lokálnej.

Sklad dát



Sprostredkovanie



Dva spôsoby sprostredkovania

1. *Dotazovo-centrický* : Prostredník transformuje dotazy na poddotazy k zdrojovým databázam a z čiastočných odpovedí skladá výsledok.
2. *Pohľadovo-centrický* : Zdroje sú definované ako pohľady na globálne relácie; mediator hľadá spôsoby ako poskladať výsledok z týchto pohľadov.

Veľmi jednoduchý príklad

- Firma Dell potrebuje kúpiť zbernicu a disk, ktoré používajú ten istý protokol.
- Globálna schéma:
  Zbernice(výrobca, model, protokol)
  Disky(výrobca, model, protokol).
- Lokálne schemy: sú podobné, ale iba dvojatribútové, *výrobca* je samozrejmý.

Príklad: Dotazovo-centricky

- Prostredník môže začať dotazom na dvojice *<model, protokol>* z relácii *zbernica* u každého vyrobcu.
  - Rozhranie k ním pridá *výrobcu*.
- Potom pre vrátený protokol, sa prostredník u každého výrobcu spýta na disk s týmto protokolom.
  - Rozhranie znovu pridá *výrobcu*.

Príklad: Pohľadovo-centricky

- Jednotlivé databázy sú definované prostredníctvom globálnych predikátov.
  - Napr.: databáza IBM je definovaná *IBMdisky(M, P) = Disky('IBM', M, P)* *IBMzbernice(M, P) = Zbernice('IBM', M, P).*
- Prostredník vytvorí odpoveď ako zjednotenie cez všetky dvojice výrobcov joinov zberníc a diskov podľa protokolu.
  - Problém je dostatočne všeobecná metóda (teória) ako skladať výsledok.

Kde je problém?

- Optimalizácia a zase optimalizácia.
  - V query-centrickom prístupe: výber plánu vyhodnotenia dotazu
    - Napr. poradie vyhodnocovania poddotazov.
  - V view-centrickom systéme: teória všeobecného vytvorenia dotazov a skladania odpovedí aj v prípadoch, keď jednotlivé pohľady sú „trošku” nehomogénne.

On-line analytické spracovanie a dátové kocky

- Podnikové štatistické a analytické dáta prerastajú operačné dáta.
  - Posun horizontu, paradox výpočtovej sily.
  - Zložité agregácie, napr.: všetky, deň, týždeň, mesiac rok.
- Konflikty medzi operačnými transakciami a analytickými „ságami”.
  - Sériovateľnosť vers. „very dirty read”.

Riešenia

- Samostatná databáza pre analytické spracovanie.
  - Podobné na data warehohuse s možnosťou čiastočnej predagregácie
- Hlavný problém je množstvo dát a optimalizácia.
  - Triedenie
  - Hašovanie

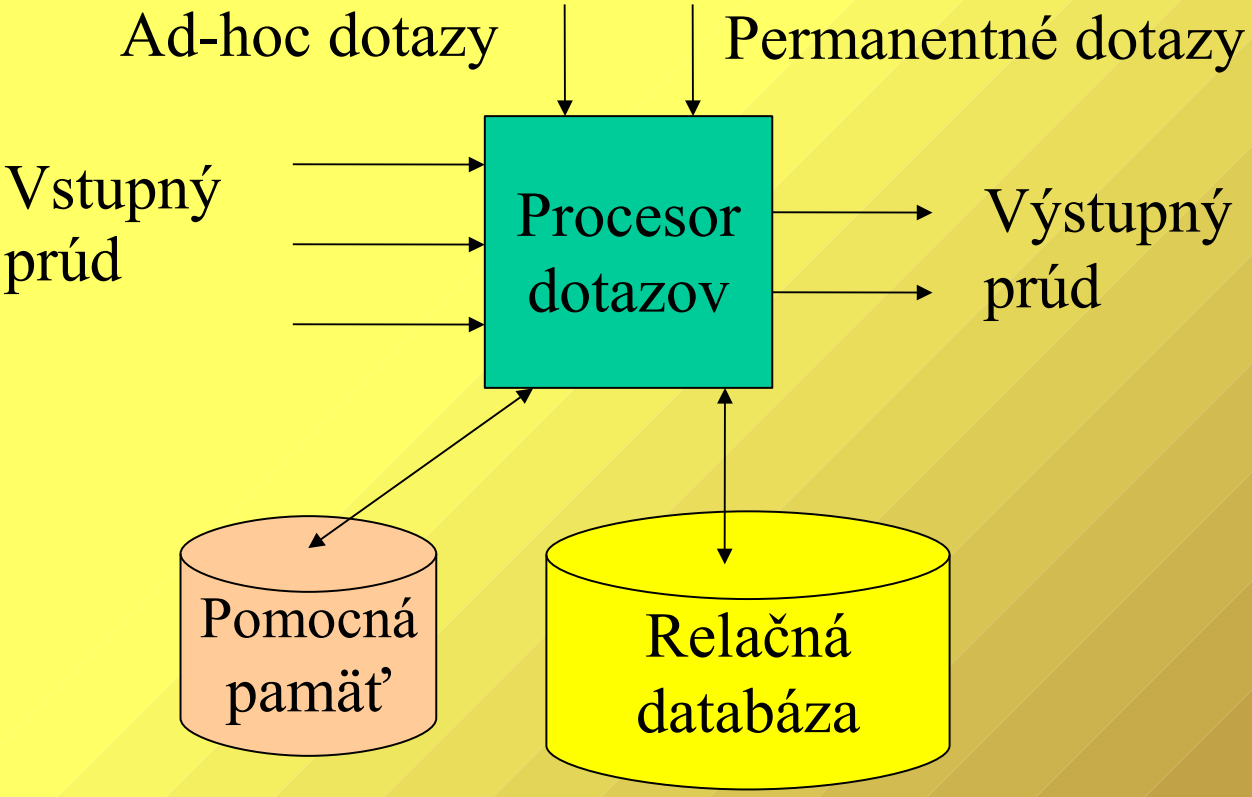
Nové smery

1. Integrácia informácii.
2. Olap & dátové kocky.
3. **Spracovanie prúdov**.
4. Semištruktúrované dáta a XML.
5. Partnerské a gridové databázy.
6. Dolovanie z dát.

Systémy na prácu s prúdmi

- Pridávame nový dátový typ prúd - *stream* = nekonečná postupnosť n-tíc, ktoré prichádzajú na nejaký port jedna za druhou. Treba ich spracovávať on-line.
- Aplikácie: Účtovanie telefonných hovorov, detekcia útokov, monitorovanie návštev stránok, riadenie procesov a pod.

Architektúra prúdového SRBD



Stanfordský prístup (Widom, Motwani)

- Hlavným pojmom je *window* - relácia ktorá sa vytvára z prúdu podľa nejakého pravidla.
  - Príklady: „posledných 10 n-tíc”, „všetky n-tice za posledných 24 hodín”. Pravidlo môže obsahovať aj filter.
- Dotazový jazyk je SQL-like s nástrojom na konverziu prúdu do okien a relácií.

Príklad:

```
SELECT …
FROM Stream1 [last 10] as Window1,…
WHERE Window1.a = 5 AND …
```

Výskumné problémy

- Znovu, rozhodujúca je optimalizácia.
  - V novom jazyku s novými dátovými typmi zaužívané metódy nemusia fungovať.
- Ani sémantika nie je celkom jasná.
  - Príklad: Keď spojíme dve okná vytvorené na základe rôznych časových ohraničení, čo znamená výsledok vzhľadom k pôvodnému prúdu?
    - Je to dôležité, keď chceme aplikovať algebraické zákony na úpravu výrazov.

MIT- prístup (Stonebraker a kol.)

- Definuje operácie priamo na prúdoch.
- Dotaz je postupnosť operácii.
- Dotazový jazyk podobný algebre.
- Optimalizácia je stále rozhodujúca. Teraz má algebraickú povahu.

Partnerské systémy a Gridy

- Partnerské systémy umožňujú zdielať dáta a procesy na aplikačnej úrovni.
- Grid je podpora pre partnerské systémy na úrovni operačného systému.

Aplikácie pre partnerské systémy

1. Zdielanie súborov: Napster, Kazaa a pod.
2. Špecifické vedecké výpočty: Seti@home, Folding@home, kódy, genom.
3. Distribuované databázy, digitálne knižnice.
4. Replikácie dát v intranete pre väčšiu dostupnosť.

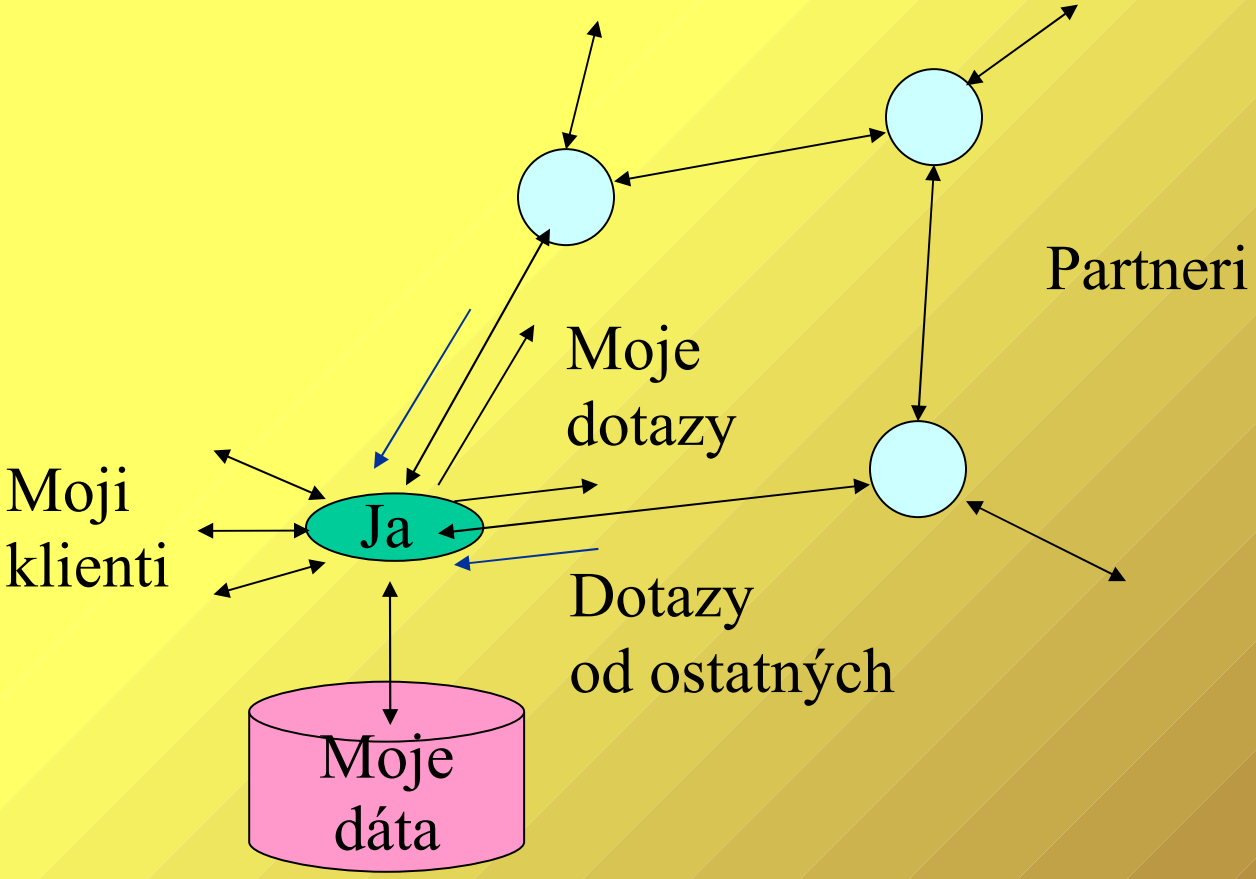
Ďalšie použitie Gridu

1. Vedecké aplikácie obvykle riešené na pracovných staniciach v sieti.
2. Využitie, predaj nevyužitého strojového času.
3. Globálne zdroje, napr.: umietstnenie dát na internete miesto na vlastnom disku.

Partnerské (peer to peer) databázy

- Dáta sú distribuované na nezávislých zdrojoch.
- Ako integrácia informácii, ale z ďaleko slabšími väzbami medzi jednotlivými zdrojmi dát.
- Aplikácie: zdielani knižníc a dátových zdrojov, ochrana pred výpadkom replikáciou dát a pod.

Architektúra partnerských SRBD



Otázky pre výskum

1. Protokoly, stratégie pre správu pamäte.
   - Ak akceptujem požiadavku na okopírovanie dát, ako dlho zostanú tieto?
   - Stratégie ponúkania a predaja pamäťovvého priestoru.
2. Dotazy a stratégie vyhľadávania.
   - Ako „ďaleko” hľadať ?
   - Čo robiť so súperiacími alebo nekompatibílnymi požiadavkami?

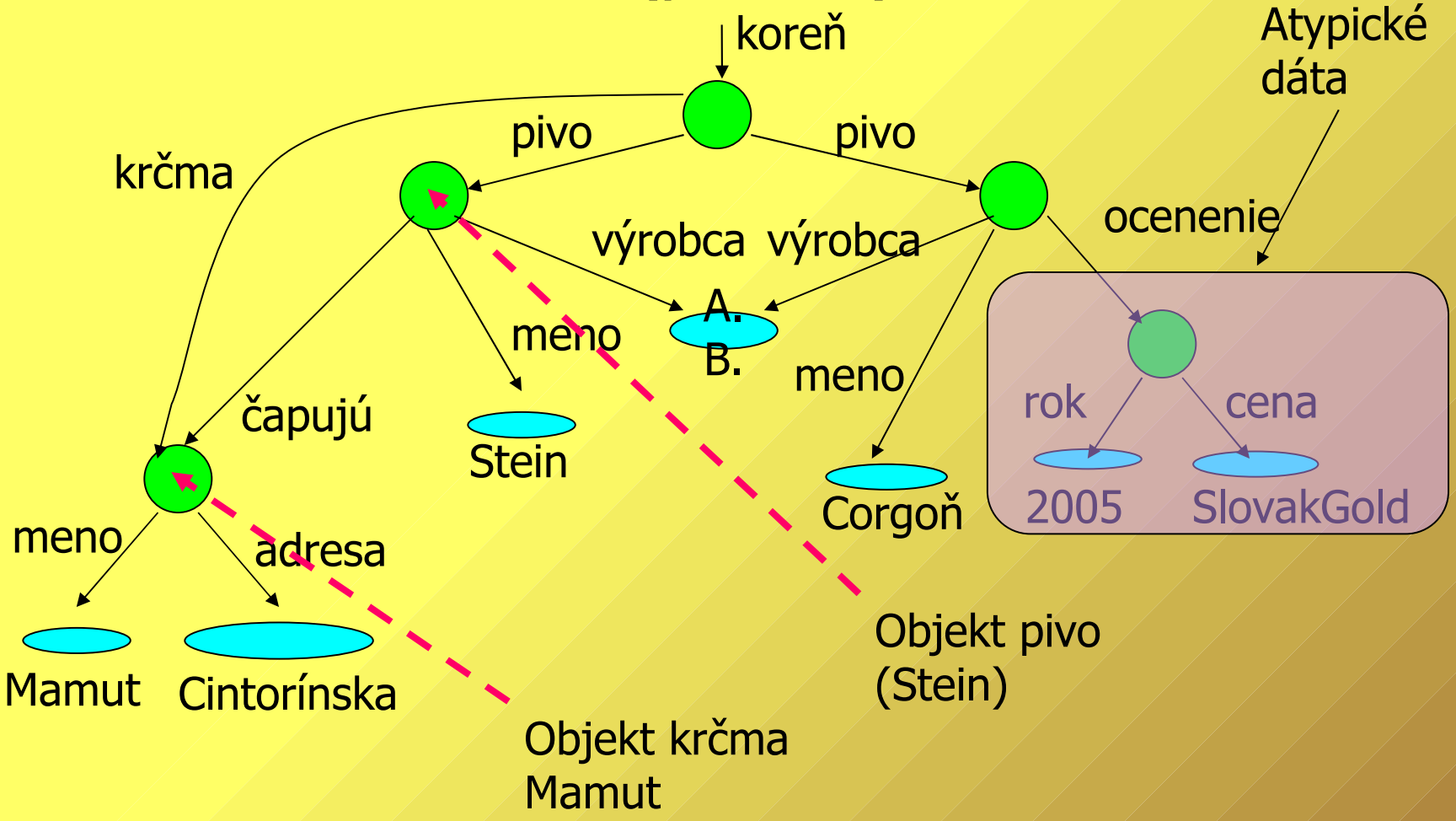
Stav výskumu partnerských db

- Počiatočné úspechy: SETI@home, folding@home.
- Napster a iní urobili veľký pokrok v architektúre a technikách, ale v istých kruhoch aj „zlé meno” pre túto oblasť.
- Optimalizácia a analýza data-retrieval algoritmov je iba v počiatku.

Semištruktúrované dáta (XML databázy)

- Modelom dát sú stromy prípadne grafy namiesto relácii. Neexistuje globálna schéma.
- Podobné ako integrácia informácii, ale lokálne schémy sa stále menia.
- W3C-XML: XML, XSL, XPATH, XQUERY atd.

Graf semištruktúrovaných dát (príklad)



Aplikácie XML

- Zdielanie dát v štandardnom formáte.
- Dá sa použiť ako globálna schéma pri integrácii informácii na podnikovej úrovni.
- Pamätanie dát, ktoré nemajú globálnu schému.

Dotazy v XML databázach

- XQUERY je nový štandard dotazového jazyka na XML dokumenty.
  - Podobne ako SQL je to jazyk vysokej úrovne.
- Výskum v oblasti optimalizácie dotazov a normalizácie XML db je iba v počiatkoch.
  - Treba použiť nové techniky, ktoré sa nepodobajú tým, čo boli s úspechom použité pre SQL.

Teoretické problémy: simulácia a bisimulácia

Definicia: Nech $G_1 = \langle V_1, E_1, L_1 \rangle$ a $G_2 = \langle V_2, E_2, L_2 \rangle$ sú dva hranovo ohodnotené grafy . Simulácia je relácia G_1 do G_2 je relácia $R \subseteq V_1 \times V_2$, ktorá má vlastnosť: Ak $\langle x_1, x_2 \rangle \in R$ a $\langle x_1, y_1 : a \rangle \in E_1$, potom existuje y_2 také, že $\langle x_2, y_2 : a \rangle \in E_2$ a $\langle y_1, y_2 \rangle \in R$.

Zaujímajú nás maximálne simulácie.
Ak R je funkcia, potom je to grafový homomorfizmus.

Bisimulácia

Ak R a R^{-1} sú simulacie G_1 do G_2 a naopak, potom R nazývame bisimuláciou.

Ak G_1 a G_2 sú dva koreňové. Hovorímeme že grafy G_1 a G_2 sú bisimilar, ak existuje bisimulácia R taká, že $<root(G_1),root(G_2)> \in R$

Píšeme $G_1 \simeq G_2$.

Zaujimajú nás algoritmy pre maximálne bisimulácie a simulácie.

Nové smery

1. Integrácia informácii.
2. Olap a dátové kocky.
3. Spracovanie prúdov.
4. Semištruktúrované dáta a XML.
5. Partnerské a grid databázy.
6. **Dolovanie z dát.**

Dolovanie z dát

1. Štatistika a štatistické metódy.
2. Induktívne inferencie.
3. Aproximácia a interpolácia, analýza časových radov.
4. Klasifikačné metódy.
   - „Cluster analysis“
   - Bayesovské siete
   - Rozhodovacie stromy

Dolovanie z dát
(pokračovanie)

5. Pattern, string, sequence matching.
6. Web mining (google, clusty, yahoo, ...).
7. Analýza nákupného košíka.
8. Asociačné pravidlá.
9. Episode mining.