

13. előadás

Adattárházak és OLAP

Áttekintés az adattárházakról és az OLAP-ról

Adatbázisrendszerek előadás
2014. december 15.



Célok

Definíciók és fogalmak

Adattárházak
jellemzői

Adattárházak
adatmodelljei

Adattárházak építése

Adattárházak
működése

Ispány Márton, Kósa Márk és Pánovics János
Debreceni Egyetem
Informatikai Kar

Az előadások Elmasry & Navathe: Fundamentals of Database systems alapján készültek.

Az adattárházak célja

- A hagyományos adatbázisok nem csak az adat-hozzáférésre optimalizáltak, hanem emellett az adatok integritását is biztosítják, illetve ezen két szempont között egyensúlyoznak.
- Legtöbbször az adattárház felhasználóknak csak olvasási hozzáférésre van szükségük, azonban ennek a hozzáférésnek gyorsnak kell lenni még nagy méretű adat esetén is.
- Az adattárház elemzésekhez szükséges adatok többsége több adatbázisból jön, továbbá ezek az elemzések ismétlődőek és előrejelezhetőek, így lehetséges speciális szoftverekkel ezeknek a követelményeknek megfelelni.
- Nagy szükség van olyan eszközökre, amelyek információval látják el a döntéshozókat azért, hogy gyorsan és megbízhatóan hozzanak döntéseket historikus adatokra alapozva.
- Ezeket a képességeket **adattárházakkal** és **közvetlen analitikus feldolgozással (online analytic processing - OLAP)** érhetjük el.



Célok

[Definíciók és fogalmak](#)

[Adattárházak jellemzői](#)

[Adattárházak adatmodelljei](#)

[Adattárházak építése](#)

[Adattárházak működése](#)



W.H. Inmon adattárház definíciója

Az adattárház adatok téma-orientált, integrált, nemváltozó, időbélyeggel rendelkező összessége a menedzsment döntéseinek támogatására.

Célok

Definíciók és fogalmak

Adattárházak jellemzői

Adattárházak adatmodelljei

Adattárházak építése

Adattárházak működése

- Az adattárházaknak olyan megkülönböztető jellemzőik vannak, amelyek főként a döntéstámogatási alkalmazásokból következnek. A hagyományos adatbázisok tranzakciósak.
- Adattárházakkal kapcsolatos alkalmazások:
 - Az **OLAP - Online Analytical Processing (közvetlen analitikus feldolgozás)** kifejezést adattárházakból származó komplex adatok elemzésére használjuk.
 - A **DSS - Decision Support Systems (döntéstámogatási rendszerek)**, melyeket EIS - Executive Information Systems (vezetői információs rendszerek)-nek is neveznek a szervezetek vezető döntéshozóit támogatják abban, hogy képesek legyenek összetett és fontos döntések meghozatalára.
 - Az **adattányászat (data mining)** a tudásfeltárás egy fontos eszköze, amely során előre nem várt új tudáshoz jutunk.



- Adattisztítás és újraformázás
- OLAP
- Adatbányászat



Célok

Definíciók és fogalmak

Adattárházak jellemzői

Adattárházak adatmodelljei

Adattárházak építése

Adattárházak működése

- Az adattárházakat főként a megfelelő adatelérésekre optimalizálják. A hagyományos adatbázisok tranzakciósak és egyaránt optimalizáltak az adat-elérési mechanizmusok és a sértetlenség biztosítása tekintetében.
- Az adattárházak nagyobb hangsúlyt helyeznek a historikus adatokra mivel fő céljuk idősorok és trend elemzések támogatása.
- A tranzakciós adatbázisokkal szemben az adattárházak nem változnak abban az értelemben, hogy ha egy adat egyszer oda bekerült, akkor az ott is marad változatlan formában az „idők végezetéig”.
- A tranzakciós adatbázisokban a tranzakció az a mechanizmus, amely megváltoztatja az adatbázist. Ezzel szemben az adattárházakban az információ durván szemcsézett és a frissítési politika alaposan megválasztott, általában inkrementális jellegű.



- Többdimenziós koncepcionális nézet
- Általános dimenziókezelés
- Korlátlan dimenzió és aggregációs szint
- Dimenziók közötti műveletek korlátlansága
- Dinamikus ritka mátrixok kezelése
- Kliens-szerver architektúra
- Többfelhasználós támogatás
- Hozzáférhetőség
- Átláthatóság
- Intuitive adatmanipuláció
- Konzisztens riportoló képesség
- Flexibilis riportolás



Célok

Definíciók és fogalmak

Adattárházak
jellemzői

Adattárházak
adatmodelljei

Adattárházak építése

Adattárházak
működése



- Általában az adattárházak nagyobb méretűek mint a forrás adatbázisok (még ezek együttesénél is).
- A teljes adattömeg kérdéses, leginkább attól függ, hogy az alábbi adattárház típusok közül melyiket választjuk:
 - **Vállalati adattárház**, amely általában egy nagy projekt és nagy idő és erőforrás ráfordítást igényel.
 - **Virtuális adattárház**, amely operatív adatbázisok különböző nézeteit nyújtja, amely nézeteket a hatékony elérés céljából fizikailag is létrehozunk.
 - **Adatpiac**, amely a szervezet egy jól meghatározott részét célozza meg, amelyre viszont erősen fókuszál (pl. marketing osztály stb.).

Célok

Definíciók és fogalmak

Adattárházak jellemzői

Adattárházak adatmodelljei

Adattárházak építése

Adattárházak működése



- A hagyományos adatbázisok általában kétdimenziós adatokkal foglalkoznak (adattábla, adatmátrix, reláció). Azonban a többdimenziós adattároló modellekben a lekérdezés hatékonysága jobb.
- Az adattárházak képesek kihasználni ennek a tulajdonságnak az előnyeit, mivel ők
 - nemváltozóak,
 - a végrehajtandó elemzés jól előrejelezhető.

Célok

Definíciók és fogalmak

Adattárházak jellemzői

Adattárházak adatmodelljei

Adattárházak építése

Adattárházak működése



- Példa két- illetve többdimenziós adatszerkezetre
- A többdimenziós modellek előnyei:
 - könnyen hagyja magát hierarchikusan szemlélni az ún. felgöngyölítés (roll-up) és lefűrés (drill-down) műveletekkel.
 - az adatok közvetlenül lekérdezhetőek bármilyen dimenzió kombinációban összetett adatbázis lekérdezések útján.

Célok

Definíciók és fogalmak

Adattárházak jellemzői

Adattárházak adatmodelljei

Adattárházak építése

Adattárházak működése

A többdimenziós sémákat az alábbiak felhasználásával specifikálhatjuk:

- **Dimenzió-tábla**, amely a dimenziók attribútumainak rekordjaiból áll.
- **Tény-tábla**, amelynek minden rekordja egy rögzített tény adat. Ez a tény mért vagy megfigyelt változókból áll és a dimenzió táblákra mutató pointerekkel azonosítjuk őket. A tény-tábla tartalmazza az adatokat és a dimenziókat az adatokbeli rekordok azonosítására.

Az általánosan használt többdimenziós sémák a következők:

- **Csillag séma**, amely egy tény-táblát és minden dimenzióhoz egy egyszerű táblát tartalmaz.
- **Hópehely séma**, amely a csillag-séma egy oly módon továbbfejlesztett változata, hogy dimenzió-táblák egy hierarchiáját tartalmazza.



Célok

Definíciók és fogalmak

Adattárházak jellemzői

Adattárházak adatmodelljei

Adattárházak építése

Adattárházak működése

- **Tény konstelláció** olyan táblák halmaza, amelyek ugyanazon dimenziók között osztoznak. A tény konstellációk behatárolják az adattárházbeli lehetséges lekérdezéseket.
- **Indexelés.** Az adattárházak indexelést használnak a nagy hatékonyságú elérés támogatására. A bitmap indexelés módszere bitvektorokat használ az indexelendő tartomány minden értékére. Elsősorban alacsony számosságú tartományokon működik jól.



Célok

Definíciók és fogalmak

Adattárházak jellemzői

Adattárházak adatmodelljei

Adattárházak építése

Adattárházak működése



- Az adattárház építőinek széles áttekintéssel kell bírniuk az adattárház későbbi használatáról.
- A tervezésnek támogatnia kell az ad-hoc lekérdezéseket.
- Alkalmas sémát kell választani ahhoz, hogy visszaadjuk az előrejelzett használatot.
- Az adattárház-építés lépései:
 - Az adatok összegyűjtése az adattárház számára.
 - Annak biztosítása, hogy az adattárolás hatékonyan találkozik a lekérdezési követelményekkel.
 - Teljes áttekintés nyújtása arról a környezetről, ahol az adattárház majd működni fog.

Célok

Definíciók és fogalmak

Adattárházak jellemzői

Adattárházak adatmodelljei

Adattárházak építése

Adattárházak működése

- Az adatokat több, heterogén forrásból kell kinyerni.
- Az adatokat formázni kell az adattárház számára a konzisztencia biztosításához.
- Az adatokat tisztítani kell az érvényesség biztosításához. Nehéz automatikus eljárást találni. Visszacsatolás, az adatok frissítése tisztított adatokkal.
- Az adatokat az adattárház adatmodelljéhez kell illeszteni.
- Az adatokat be kell tölteni az adattárházba. Fontos a frissítési politika helyes megtervezése.



Célok

Definíciók és fogalmak

Adattárházak jellemzői

Adattárházak adatmodelljei

Adattárházak építése

Adattárházak működése

- Tároljuk le az adatokat az adattárház adatmodelljének megfelelően.
- Hozzuk létre és tartjuk karban a szükséges adatszerkezeteket.
- Hozzuk létre és tartjuk karban a megfelelő elérési utakat.
- Gondoskodjunk az időben változó adatokról amint új adatokat adunk az adattárházhoz.
- Támogassuk az adattárház adatok naprakészre hozását.
- Frissítsük az adatokat.
- Tisztítsuk az adatokat.
- A használat megtervezése.
- Az adatmodell illeszkedése.
- A használható adatforrások jellemzői.
- A metaadat komponens tervezése.
- Moduláris komponens tervezése.
- A menedzselhetőség és a változás megtervezése.
- Elosztott és paralel architektúrák figyelembe vétele.



- gönygyöltés (roll-up)
- lefűrás (drill-down)
- pivot
- slice - dice
- rendezés
- szelekció
- származtatott attribútumok



Célok

Definíciók és fogalmak

Adattárházak
jellemzői

Adattárházak
adatmodelljei

Adattárházak építése

Adattárházak
működése