Adatbázisrendszerek
10. előadás: Adattárházak és OLAP
Áttekintés az adattárházakról és az OLAP-ról

2021. április 26.

DEBRECENI
EGYETEM

Az adattárházak célja

- A számítási kapacitások állandó növekedése és az analitikai eszközök és módszerek egyre összetettebbé (szofisztikáltabbá) válása eredményezte azt a fejlődést, amely az adattárházakban kulminálódott.
- A hagyományos adatbázisok nem csak az adatok hozzáférésére 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 mennyiségű adat esetén is.

Az adattárházak célja

- 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 históríkus adatokra alapozva.
- Ezeket a képességeket adattárházakkal és közvetlen analitikus feldolgozással (online analytic processing - OLAP) érhetjük el.

Az adattárház definíciója

W.H. Inmon adattárház definíciója

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

Alapfogalmak

- 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 analítikus 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 adatbányászat (data mining) a tudásfeltárás egy fontos eszköze, amely során előre nem várt új tudáshoz jutunk.

Az adattárház koncepcionális szerkezete

- Adattisztítás és újraformázás
- OLAP
- Adatbányászat

Összehasonlás tradicionális adatbázisokkal

- Az adattárházakat főként a gyors adatelérésre optimalizálják. A hagyományos adatbázisok tranzakciósak és egyaránt optimalizáltak az adatelérési mechanizmusok és a konzisztencia biztosítása tekintetében.
- Az adattárházak nagyobb hangsúlyt helyeznek a históriku 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.

Az adattárházak jellemzői

- 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

Adattárházak osztályozása

- Általában az adattárházak egy vagy két nagyságrenddel 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.).

Adattárházak adatmodelljei

- 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.

Adattárházak adatmodelljei

- Két- illetve többdimenziós adatszerkezetek
  - Kétdimenziós: táblázat, adattábla
  - Többdimenziós: adatkocka (hiperkocka)
- A többdimenziós modellek előnyei:
  - egyes dimenziók előtérbe helyezése forgatással (pivoting)
  - könnyen hagyja magát hierarchikusan szemlélni az ún. felgönygyölítés (roll-up) és lefúrás (drill-down) műveletekkel.
  - az adatok közvetlenül lekérdezhetőek bármilyen dimenzió kombinációban összetett adatbázis lekérdezések útján.

Többdimenziós sémák

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 olymódon továbbfejlesztett változata, amely dimenzió-táblák egy hierarchiáját tartalmazza.

Többdimenziós sémák

- 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. 1-et írunk a j-edik pozicióba, ha a j-edik rekord rendelkezik az adott értékkel, egyébként pedig 0-t. Elsősorban alacsony számosságú tartományokon működik jól. Pl.: 100 E rekord és 4 attribútumérték esetén 4 db 100 E hosszú bitvektor (12.5Kb, összesen 50Kb) jön létre. A join idexelés a hagyományos elsődleges-külső kulcs kapcsolatot valósítja meg hatékonyan a dimenzió-táblák és a tény-tábla között.

Adattárházak építése

- Az adattárház építőinek széles áttekintéssel kell bírniuk az adatárház 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.

Az adatok összegyűjtése

- 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.

Adattárházak építése

- Tároljuk le az adatokat az adattárház adatmodelljének megfelelően.
- Hozzuk létre és tartsuk karban a szükséges adatszerkezeteket.
- Hozzuk létre és tartsuk 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.

Adattárházak működése

- gönygyölítés (roll-up)
- lefúrás (drill-down)
- pivot
- slice - dice
- rendezés
- szelekció
- származtatott attribútumok