Adattárházak

Az adattárház célja

A tradicionális adatbázisok nem lekérdezésekre vannak optimalizálva. Biztosítaniuk kell az adatok integritását az adatmódosítások esetén.

Az adattárház felhasználói legtöbbször **olvasni** szeretnék az adatokat, és **nagymennyiségű** adat gyors elérésére van szükségük.

Az adattárház elemzésekhez szükséges adatok több adatforrásból származnak. Ezek az elemzések többször ismétlődnek illetve típusaik megjósolhatóak.

Nagy szükség van olyan eszközökre, amelyek ellátják a döntéshozókat olyan historikus adatokon alapuló információval, amelyek alapján gyorsan és megbízható döntést lehet hozni.

Ezeket a funkcionalitásokat támogatja az adattárház és az OLAP (**Online analytical processing)**

Az adattárházak elsősorban a döntéshozást támogatják.

Adattárház definíciója

W. H. Inmon definíciója

- "Az adattárház egy tárgyorientált, integrált, az adatok történetiségét tároló, nem illékony adatrendszer, amelynek fő célja az adatokból történő hatékony információkinyerés biztosítása, elsősorban a döntéshozatali folyamatok támogatása céljából."

Definíció

tárgyorientált, tematikus (subject-oriented):

- Az adattárház a döntéshozók elemzési követelményeire fókuszál a döntéshozatal különböző szintjein, azaz különböző témákra, mint eladás, ügyfélviselkedés.

- Hagyományos adatbázis ezzel szemben az alkalmazások által végrehajtandó funkciókra fókuszál, azaz funkcióorientált. Például egy eladás regisztrálása.

Minden cégnek saját tematikái vannak
biztosítási cég: ügyfél, kötvény, követelés, prémium
gyártó cég: termék, rendelés, szállító, nyers áruk
kiskereskedő: termék, értékesítés, szállító

Definíció

Integrált

- Az adattárház heterogén adatforrásokból dolgozik, de ahhoz, hogy az adatokat át tudja venni, egy szabványos formára kell alakítania, egységbe rendezve egy helyre kell gyűjtenie. Az egységes megközelítés konkrétan egyetlen kulcsstruktúrát és egyetlen adatmegjelenítési módot jelent.

- Az integráltság azt is jelenti, hogy az adattárháznak fel kell tudnia ismerni, hogy ugyanarról az adatról van szó, (bár más jelöléssel és más formátummal érkeznek), és meg kell tudni állapodni egy egységes jelölésben.

Definíció

nem illékony, vagyis tartós (non-volatile)

- Az adattárházba bevitt adatok csak akkor tűnnek el, ha explicit módon töröljük őket. A bekerült adatok tehát tartósan meg is maradnak (akár 5-10 évig).

- Megvan minden pillanatfelvétel (egy idő bélyeggel ellátva) visszamenőleg egészen addig, míg bizonyos adatokat nem archiválunk, vagy végképp eltávolítunk az adattárházból. (Éppen ezért nem szokás gyorsan érvénytelenné váló adatokat bevinni az adattárházba.)

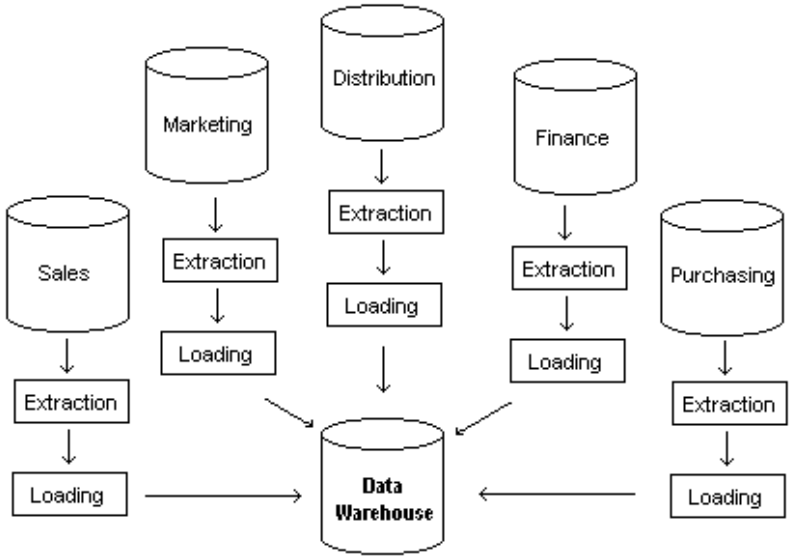
Definíció

idő függő (time-variant)

- A forrásrendszerek adatai nagyrészt az aktuális állapotra vonatkoznak, a jelen pillanatra.
- Az adattárház adatai történeti adatokat (historical data), több éves tevékenységeket fognak át. Az adatokat az időpontok és időintervallumok szerint tárolják és kezelik, a forrásrendszerek változását nyomon követve.
- Az igazi különbség az időbeliséget hordozó, és a hagyományos adatbázis rekordok között az, hogy míg a hagyományos rekordban módosítunk, törlünk, addig az új rekord adattárházba való betöltésekor az a pillanat is feljegyzésre kerül, amikor ez megtörtént. Az adattárházba egyszer bekerült adat általában már nem módosul, ugyanis ugyanannak az adatnak a módosítása időben később történik, így az adattárházba is egy másik időbélyeggel kerül be, egy új rekordba.

Adattárház

Több forrásból gyűjtött, egységesített sémával tárolt információk összessége.



Adattárházak által támogatott alkalmazások

- **OLAP** (Online Analytical Processing) az adattárházban lévő komplex adatok elemzése.
- **DSS** (Decision Support Systems) más néven EIS (Executive Information Systems) a szervezet vezető döntéshozóit támogatja, hogy összetett és fontos döntéseket hozhassak meg.
- **Data Mining** (adatbányászat) tudásfeltárásra használják, azaz előre nem látott (ad hoc) új tudást keresését jelenti.

Hagyományos adatbázis ↔ Adattárház

Hagyományos adatbázis

- OLTP (Online Transactional Processing)
  insert, update, delete, select
- Olyan lekérdezésekre optimalizálták, amelyek csak egy pár sort érintenek, és olyan tranzakciókra optimalizálták, amelyek pár sort módosítanak az adatbázisban.

Adattárház

- Arra tervezték, hogy hatékonyan támogassa a kinyerést (extract), a feldolgozást és a megjelenítést. A célja az elemzés és a döntéshozás.
- Nagyobb hangsúlyt helyez a történeti adatokra, mert a fő célja az idősorok és a trendek elemzése. A frissítés általában inkrementális.
- Több adatforrásból származó nagymennyiségű adatot tartalmaz. Az adatforrások független rendszerek lehetnek, más-más platformon, az adatmodelljük más-más lehet, és akár egyszerű állományok is lehetnek

Szempont	Hagyományos adatbázis	Adattárház
Felhasználó típusa	Irodai dolgozó	Vezető, döntéshozó
Használat	Megjósolható, ismétlődő	Ad hoc, nem strukturált
Adattartalom	Aktuális, részletes	Történeti, összegzett
Adatok szervezése	A működésnek megfelelően	Az elemzési problémáknak megfelelően
Adatstruktúra	Tranzakciókra optimalizált	Összetett lekérdezésekre optimalizált
Használat gyakorisága	Magas	Közepestől az alacsonyig
Elérés módja	Olvasás, insert, update, delete	Olvasás és hozzáfűzés (append)
Egy elérés hány rekordot érint	Keveset (10 sor)	Sokat (millió sort)
Válasz idő	Rövid	Lehet hosszú
Tervezés és megvalósítás	A teljes rendszert egy időben	Inkrementális

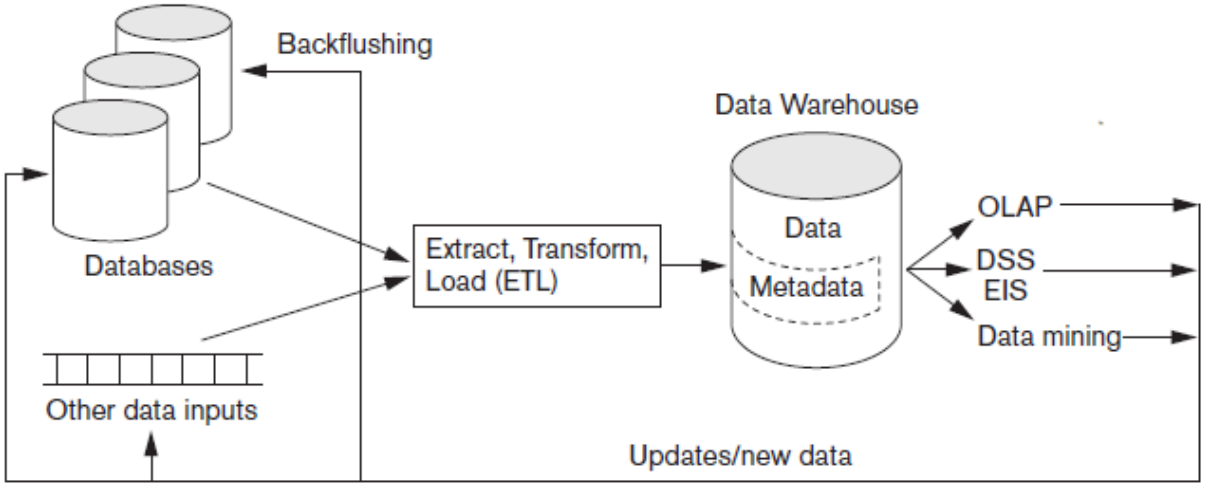
Hagyományos adatbázis ↔ Adattárház

Szempont	Hagyományos adatbázis	Adattárház
Adatmodell	Relációs, normalizált	Csillagséma, hópehelyséma
Orientáció	Tranzakció	Elemzés
Funkció	Napi műveletek	Döntés támogatás
Felhasználók száma	Sok (ezer)	Kevés (száz)
Adatbázis mérete	Néhány GB	100 GB-től több TB-ig
Párhuzamossági szint	Magas	Alacsony
Módosítás gyakorisága	Magas	Nincs módosítás
Adat redundancia	Alacsony	Magas

Az adattárházak koncepcionális felépítése

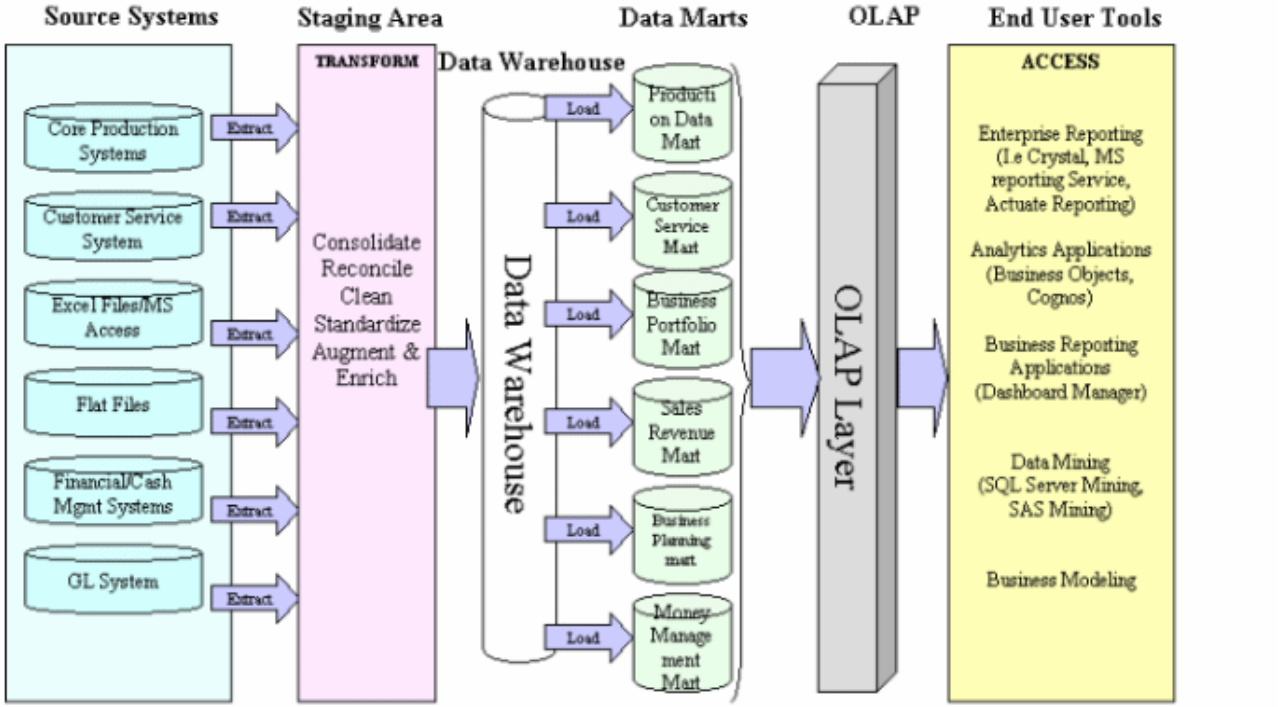
Az adattárház kezelésének folyamata magában foglalja

- Az adatok tisztítását és újraformázását (ETL tool – extract, transform, load)
- OLAP, Data Mining, DSS (Decision Support System), EIS (Executive Information System) új releváns infókat generálhatnak (pl.szabályok)



Tipikus adattárház keretrendszer

Az adatpiac (data mart) az adattárház része, mely a kiválasztott tárgyakra fókuszál. Hatóköre osztályszintű, míg az adattárházé szervezeti szintű.



ETL

Extract

- Adatokat gyűjt több, heterogén adatforrásból. Az adatforrások lehetnek hagyományos adatbázisok vagy különböző formátumú állományok.

Transform

- Az adatforrás formátumában lévő adatot átalakítja adattárház formátumúvá.

Load

- Betölti az átalakított adatot az adattárházba. Az adattárház frissítését is magában foglalja (propagating updates). A frissítési frekvencia változó, havitól a napi többszöriig.

Adattisztítás

A tisztítás folyamatának a lépései:

1. Elemekre bontás (elementizing): Az adatok atomi részekre bontását jelenti. Ez egy cím esetén a város, irányítószám, utca, házszám, emeletszám, ajtószám elkülönítését jelenti.
2. Szabványosítás (standardizing): Egységes jelölés bevezetése, például lakcímeknél a krt, körút, és egyéb rövidítések egységes formára való hozása.
3. Verifikálás (verifying): A szabványosított elemek konzisztenciájának ellenőrzése. Lakcímeknél például az irányítószám és a város egyeztetése.

Adattisztítás

A tisztítás folyamatának a lépései:

4. Illesztés (matching): Az aktuálisan vizsgált rekord (vagy néhány mezője) szerepel-e más helyen a céladatbázisban és tartalmában ugyanazon adatokat tartalmazza-e. A rendszer itt a tárolt adatok alapján bizonyos belső korrelációkat tár fel, és figyeli, hogy a bejövő adatok mennyire felelnek meg a feltárt szabályszerűségeknek. Ha egy adott nevű ügyfél már szerepel az adattárházban és egy rekordban újra találkozunk a nevével, leellenőrizhetjük, hogy a megfelelő lakcím, telefonszám, tartozik-e hozzá.
5. Dokumentálás (documenting): Ha sikerült megtisztítanunk egy adatot, akkor ezt a folyamatot megfelelően dokumentálni kell, általában a metaadatok értelemszerű módosításával.

Végfelhasználói eszközök

OLAP eszközök
- Ad hoc lekérdezések

Riportoló eszközök
- Előre definiált lekérdezések

Statisztikai eszközök
- Statisztikai módszerekkel elemzik az adatokat

Adatbányászati eszközök
- Értékes tudást fedeznek fel

Az adattárházak osztályozása

Általában az adattárházak nagyságrendekkel nagyobbak, mint a forrás adatbázisok.

Az adatok mennyisége meghatározó, amely alapján az adattárházak osztályozhatóak:

- Vállalatszintű adattárház
  Nagy projektek, masszív idő és erőforrás beruházással.
- Virtuális adattárházak
  A forrásadatbázisokon (a hatékony elérés miatt) materializált nézeteket valósít meg.
- Logikai adattárházak
  Adategyesítést, terjesztést és virtualizációt használnak
- Adatpiac (Data mart)
  Általában a cég egy részlegét célozzák meg és kevesebb témára fókuszálnak.

Adattárházak építése

Az adattárház építőjének látnia kell, hogy az adattárházat előre láthatólag mire fogják használni.

- A tervnek támogatnia kell az ad-hoc lekérdezéseket
- A megfelelő sémát kell választani az előre látható használathoz marketing orientált, termék-fogyasztóra fókuszáló cég vagy non-profit jótékonysági, adományokra fókuszáló cég

Az adattárházak tervezése a következő lépéseket foglalja magában

- Az adatok begyűjtése az adattárház számára.
- Annak biztosítása, hogy az adattárolás hatékonyan megfeleljen a lekérdezési követelményeknek.
- Az adattárházat tartalmazó teljes környezet kialakítása.

Adattárházak építése

Az adatok begyűjtése az adattárház számára

1. Az adatokat több, heterogén forrásból kell kinyerni.
2. Az adatokat következetesen kell formázni az adattárházban. A független forrásokból származó adatok neveit, jelentését és tartományait egyeztetni kell.
3. Az adatokat meg kell tisztítani, hogy érvényesek legyenek.
   - A tisztítási folyamatot nehéz automatizálni.
   - Leginkább munkaigényes
   - Back flushing, a tisztított adatokkal való feljavítás.

Adattárházak építése

Az adatok begyűjtése az adattárház számára

4. Az adatoknak illeszkedniük kell az adattárház adatmodelljébe.
5. Az adatokat be kell tölteni az adattárházba.
   A frissítési elvekhez megfelelő terv készítése.
   - Milyen frissnek kell lennie az adatnak?
   - Az adattárház lehet offline? Milyen hosszú ideig?
   - Az adatok között milyen kölcsönös függés van?
   - A tár milyen elérhetőséggel rendelkezik?
   - Mik az elosztási követelmények?
   - Mennyi idő alatt lehet az adatot betölteni? (beleértve a tisztítást, a másolást, az átadást, és az indexek újraépítését)

Adattárházak építése

Az adattárházban történő adattárolás a következő folyamatokat foglalja magában:

- Az adatok tárolása az adattárház adatmodelljének megfelelően
- A szükséges adatstruktúra létrehozása és fenntartása
- Megfelelő elérési utak létrehozása és karbantartása
- Időfüggő adatok biztosítása, amikor új adat érkezik
- Az adattárházbeli adatok módosításának támogatása
- Az adatok frissítése
- Az adatok tisztítása

Adattárházak építése

Tervezéskor figyelembe vesszük a leendő környezetet:

- A használat tervezése (ki fogja használni az adattárházat és hogyan fogja használni)
- Az adatmodell illeszkedése
- Az elérhető források jellemzői
- A metaadat komponensek tervezése
- Moduláris komponensek tervezése
- A kezelhetőség és a változások tervezése
- Az elosztott és a párhuzamos architektúrák megfontolása
- Elosztott és szövetséges (autonóm) adattárházak

Az adattárházak implementálásának a nehézségei

Sok időbe telik adattárházat építeni

- Évekbe mérik.

A minőség és a konzisztencia fontos kérdés.

A használati igények felülvizsgálata, hogy megfeleljen az aktuális igényeknek.

- Az adattárházakat úgy kell tervezni, hogy forrás beillesztésekor vagy eltávolításakor ne kelljen nagyon átszervezni

Az adattárházak adminisztrációja szélesebb körű ismereteket igényel, mint a hagyományos adatbázisoké.

Üzleti Intelligencia

Az **üzleti intelligenciát (Business intelligence -BI)** általában úgy írják le, mint
„azon technikák és eszközök halmaza, amelyek segítségével a nyers adatokat üzleti elemzési célokra alkalmas értelmes és hasznos információvá transzformálhatjuk.”

Az adattárház az üzleti informatika alapja.

Üzleti intelligencia

Alkalmazások

- Közvetlen lekérdezés és riportoló eszköz (a felhasználók közvetlenül kérhetik le az adatokat)
- Adatbányászat
- Standard riportok (előredefiniáltak, formázottak)
- Elemző alkalmazások
- Dashboardok és scoreboardok (riportok és diagramok)
- Működési (Operational) BI alkalmazások (a történetiség a fontos, az egyszerű felhasználók használják, akik az alapfeladataikat végzi, pl. ügyfélszolgálatos visszakeresi a régi címet)

Dashboard


Finance
Healthcare
Legal
Manufacturing
Marketing
Projects
Sales
Service Management
Add a dashboard item:
Hospital Aging Summary
Hospital Bill Lag Category Summary
Bill Lag Category
(Current Period)
Hospital Payer Summary
McKesson ATB Aging Summary
(current period)
McKesson ATB Aging (Over 360 Days)
McKesson Occupancy Percentage
Occupancy % by Location
Occupancy (%)
McKesson Occupancy Trend
Occupied
Total Beds
Meditech Monthly Overtime by Dept
Showing 1 to 88 of 88
Dept Number
Dept Name
Overtime Hours
Meditech MTD Revenue by Dept (Top 20)
Full size
Export Data
Settings

Tableau dashboard (click on it)