Adatbázisok II.

8

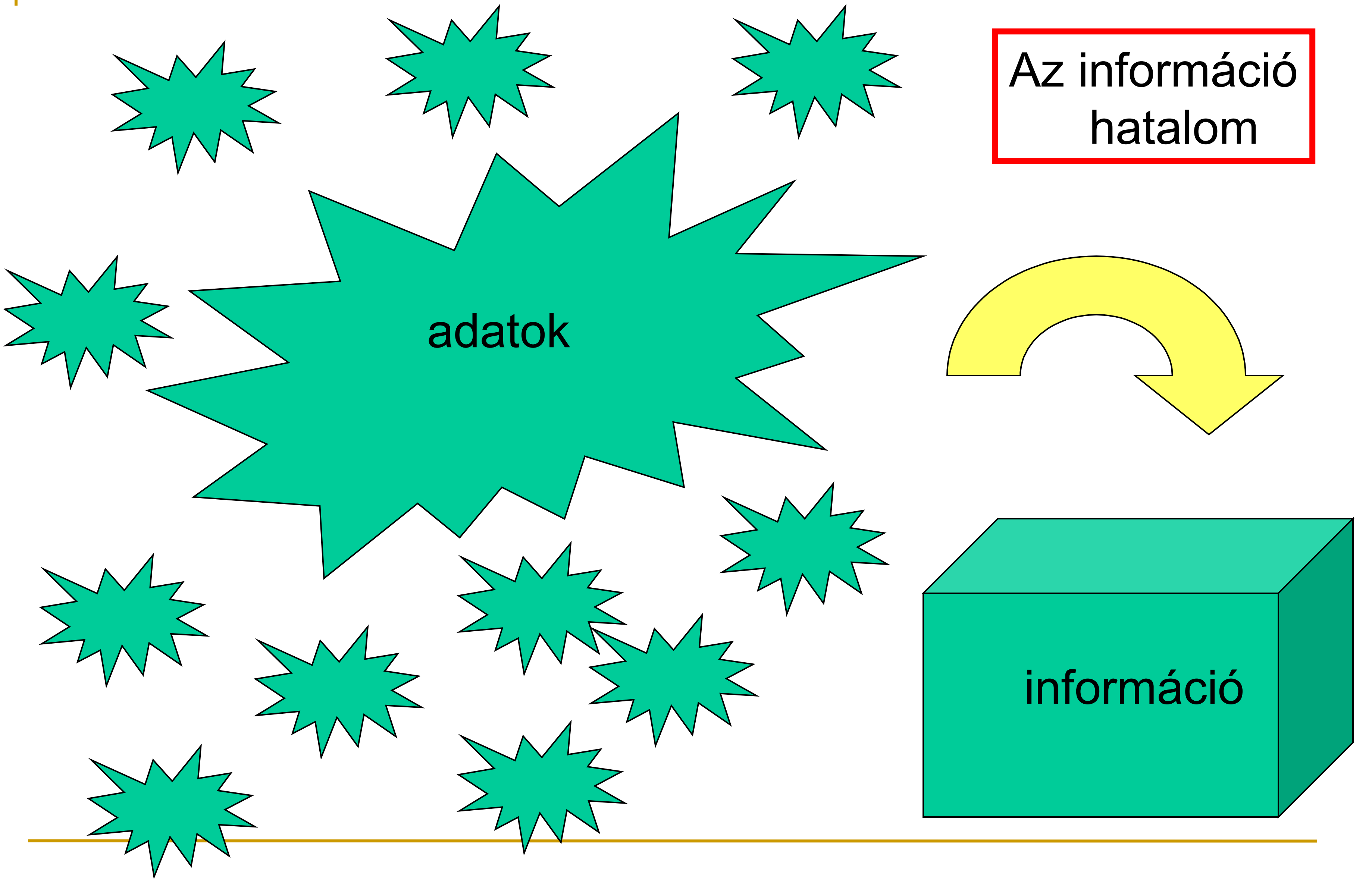
Jánosi-Rancz Katalin Tünde
tsuto@ms.sapientia.ro
327A

Adattárház rendszerek – Data Warehousing

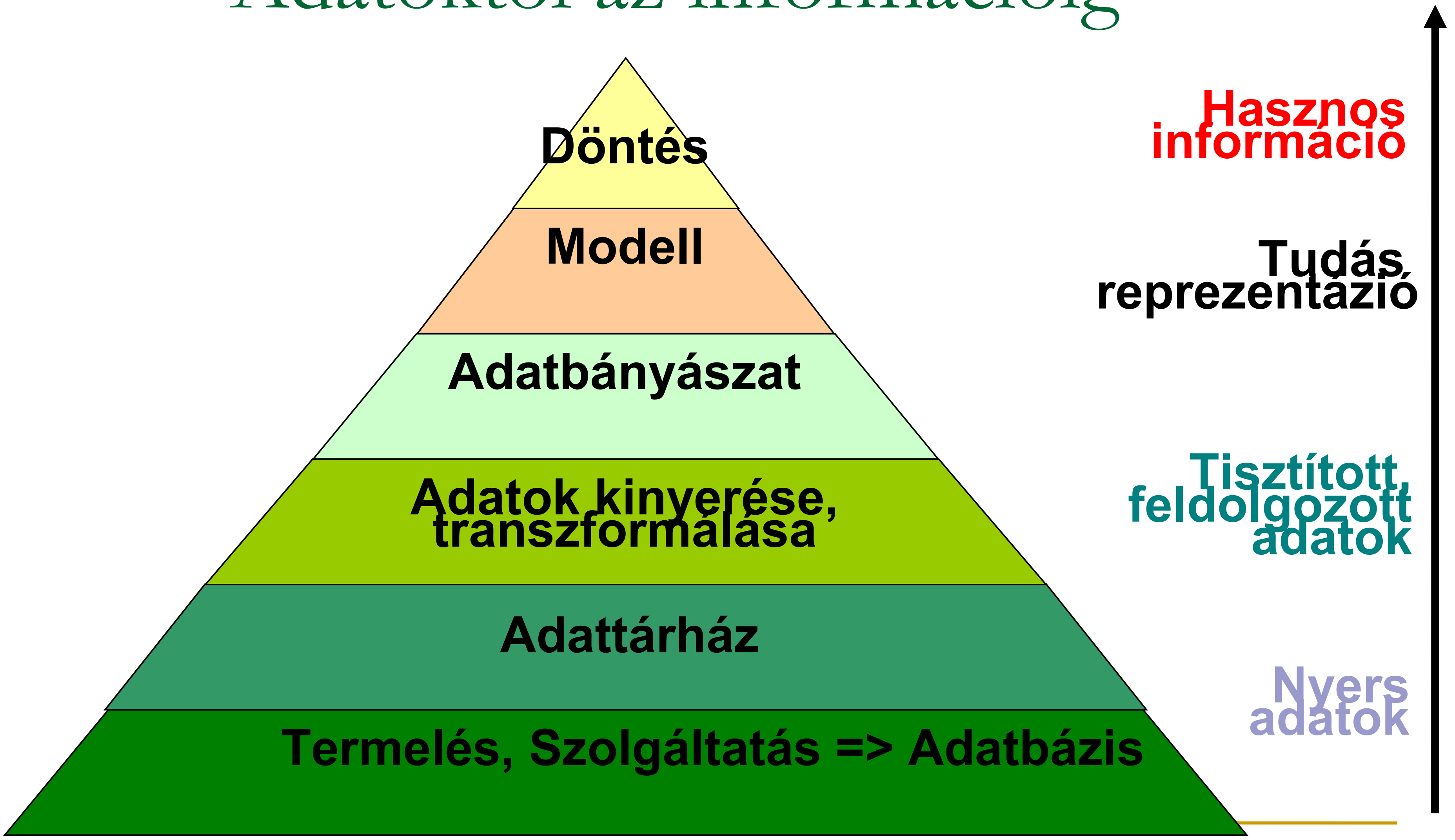
8

Miről lesz szó?

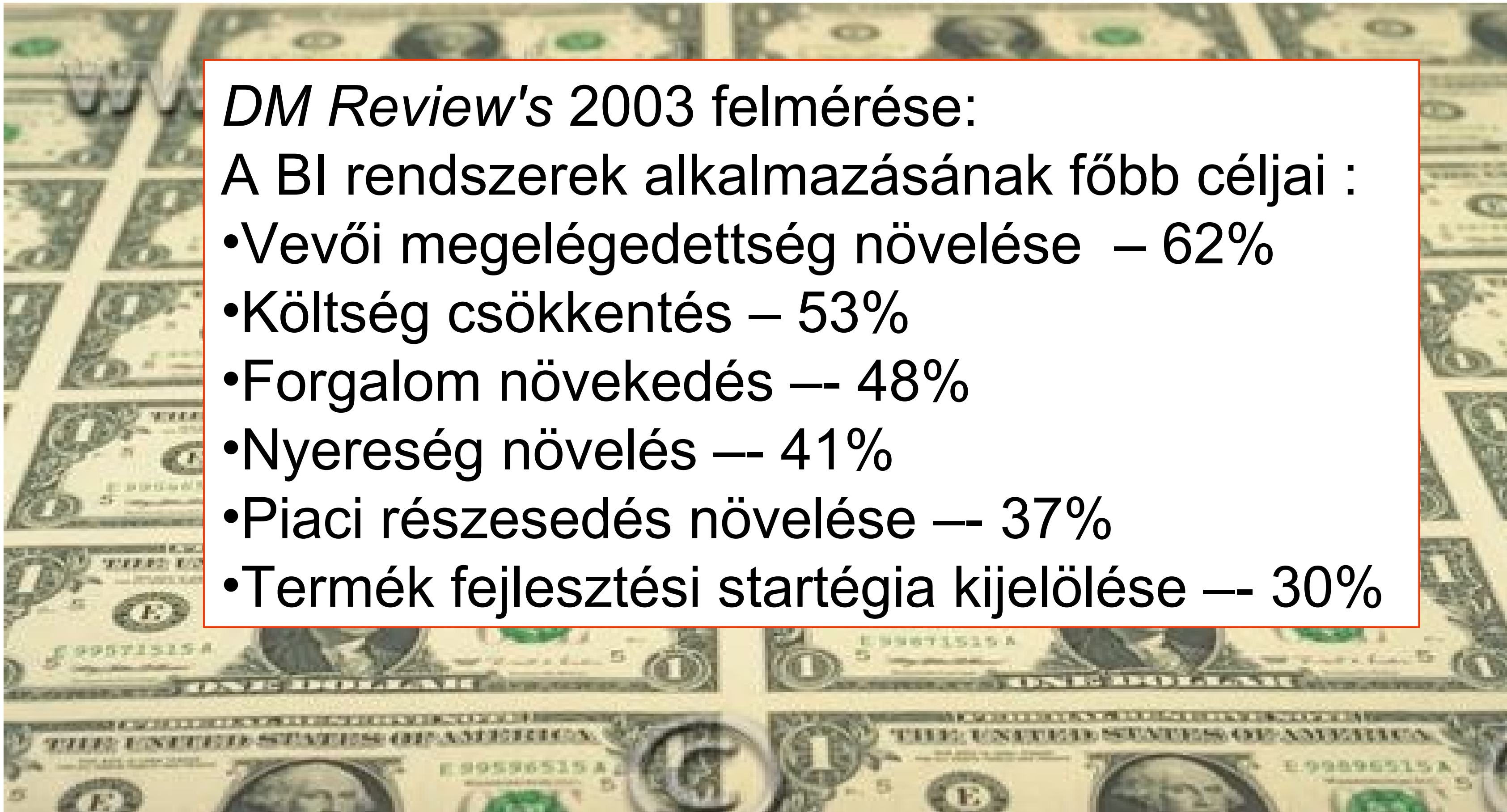
- Motiváció
  - Adattárházak létrehozásának motivációja
  - A tárgy (és tanulásának) motivációja
- Mi az OLAP?
- Mi az adattárház?


Az információ hatalom
adatok
információ

A probléma:
Adatoktól az információig



OLAP rendszerek célja



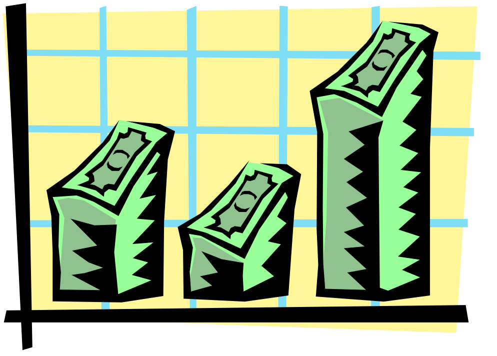
DM Review's 2003 felmérése:
A BI rendszerek alkalmazásának főbb céljai :
- Vevői megelégedettség növelése – 62%
- Költség csökkentés – 53%
- Forgalom növekedés –- 48%
- Nyereség növelés –- 41%
- Piaci részesedés növelése –- 37%
- Termék fejlesztési startégia kijelölése –- 30%

Az OLAP (on-line analitical processing) alkalmazásokhoz információ feldolgozási, elemzési feladatok kapcsolódnak

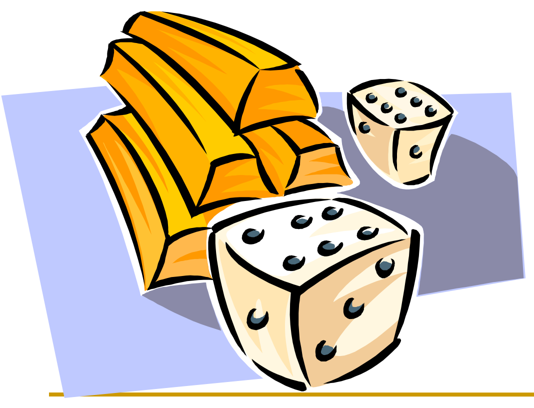
Döntési szintek:

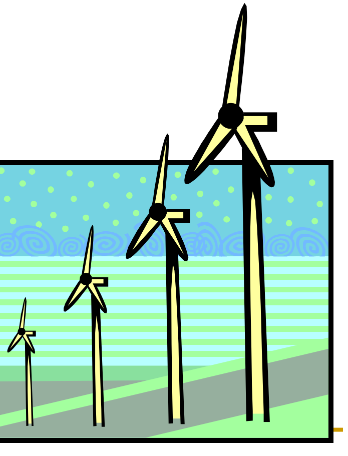
operatív: mely raktárból hozzuk a csokit?
taktikai: mennyi csoki kell a hónapban?
Stratégiai: maradjunk meg a csoki gyártásánál?

OLAP Alkalmazási területek



Bankok:
- tranzakció figyelés
- ügyfél minősítés
- ügyfél menedzsment
- beruházások
- tőzsde





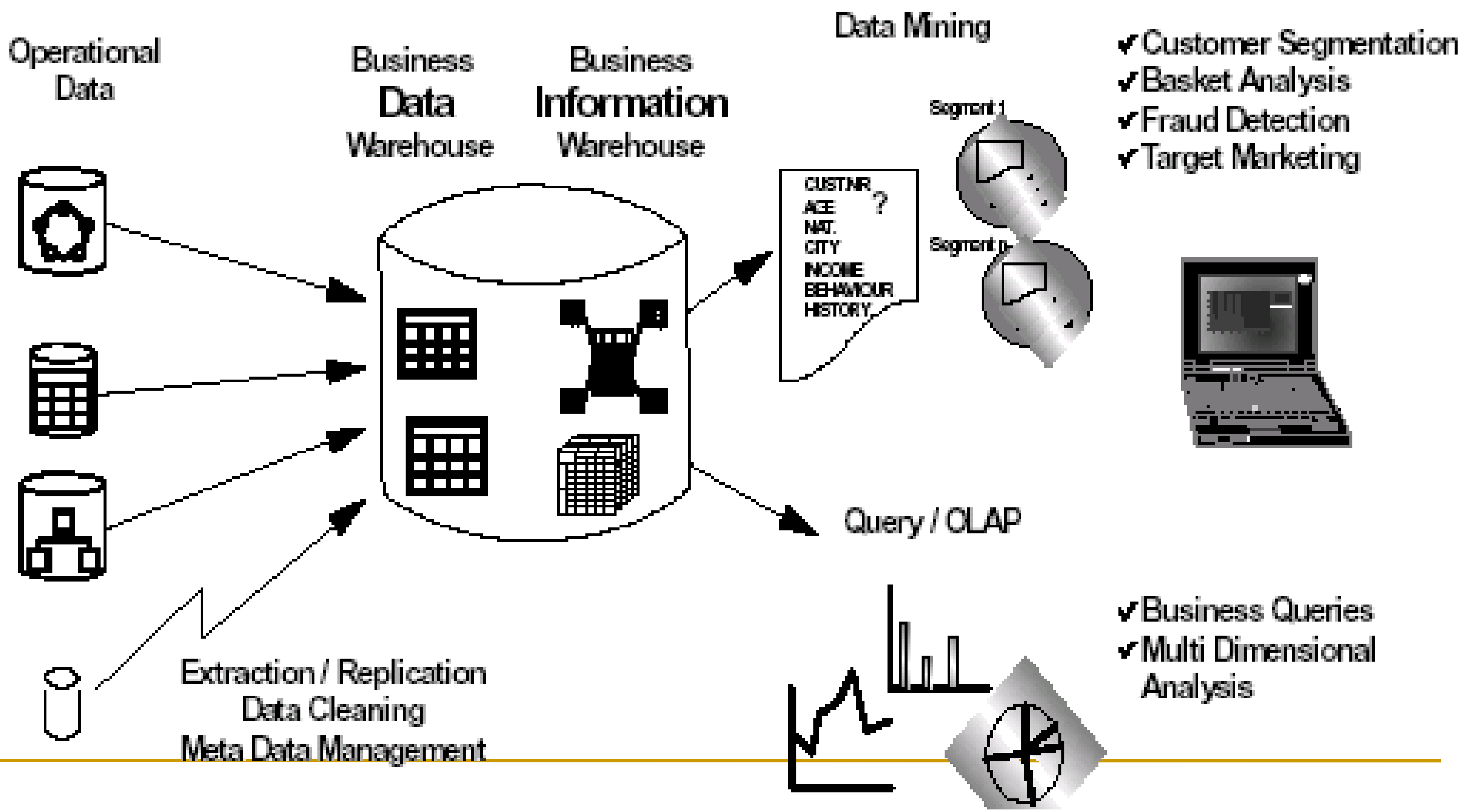
Cégek:
- fogyasztás alakulás
- piac elemzés
- döntés előkészítés
- termelés optimalizálás

… ?……?. …

Motiváció

- A vállalatok fulladoznak az adatokban, de éheznek az információért
- Vállalati környezet – a táptalaj
- Vállalat vezetése: döntések sorozata, gyors, jó minőségű döntések -> eredményesség
- A döntések minősége nagyban függ a döntéshozók informáltságától, a rendelkezésre álló adatok, információk minőségétől (több forrásból konszolidálja, illetve integrálja az információt, és egy értelmes formára hozza)
- A döntések megfelelő támogatására jelenthet megoldást az adattárház technológia
  - *Hogyan építsünk adattárházat?*
  - *Hogyan rendezzük adatainkat?*
  - *Hogyan nyerhetünk ki információt?*

Üzleti intelligencia környezet



Adattárházak

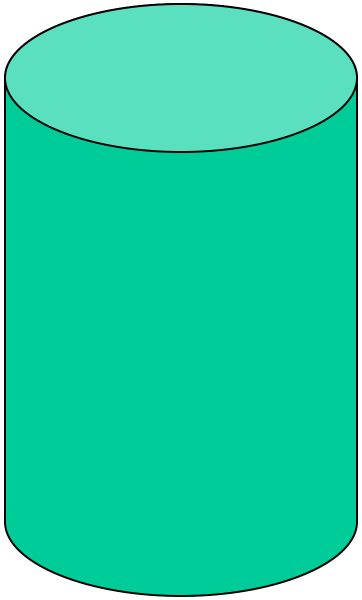
OLAP igényeket kielégítő adattárolás

Inmon: Témaorientált, integrált, az adatokat történetiségében tároló adatrendszer (1992)

Kell hatékony QUERY modul

Kell nem normalizált nézet

Kell adatintegrátor, betöltő



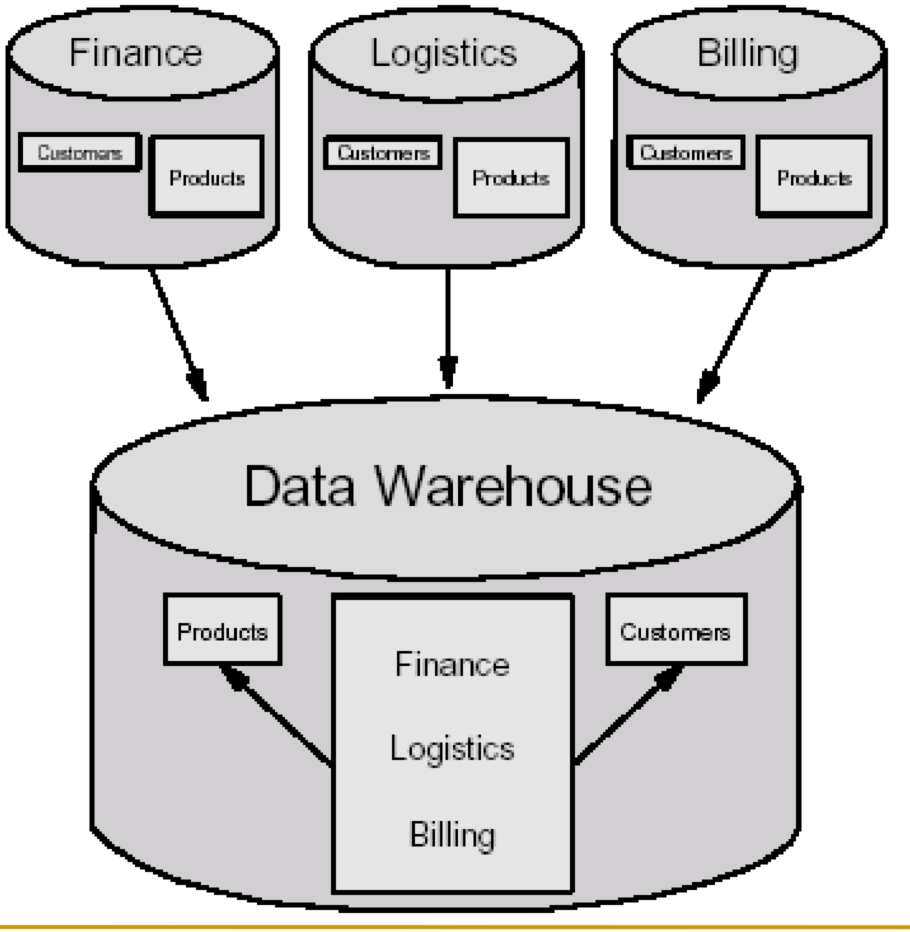
Nem kell bonyolult tranzakció kezelő

Nem kell bonyolult aktív DBMS modul

Miért is?

- Strukturált módon tárolt adatokhoz egyszerű hozzáférés
  - Különböző formátumok, platformok
- Heterogén adatforrások,
  - adattisztítás
  - szűrés
  - átalakítás
  - tárolás könnyen hozzáférhető és áttekinthető formában
- El kell különíteni a tranzakciós rendszert az információs rendszertől, hogy növeljük a teljesítményt

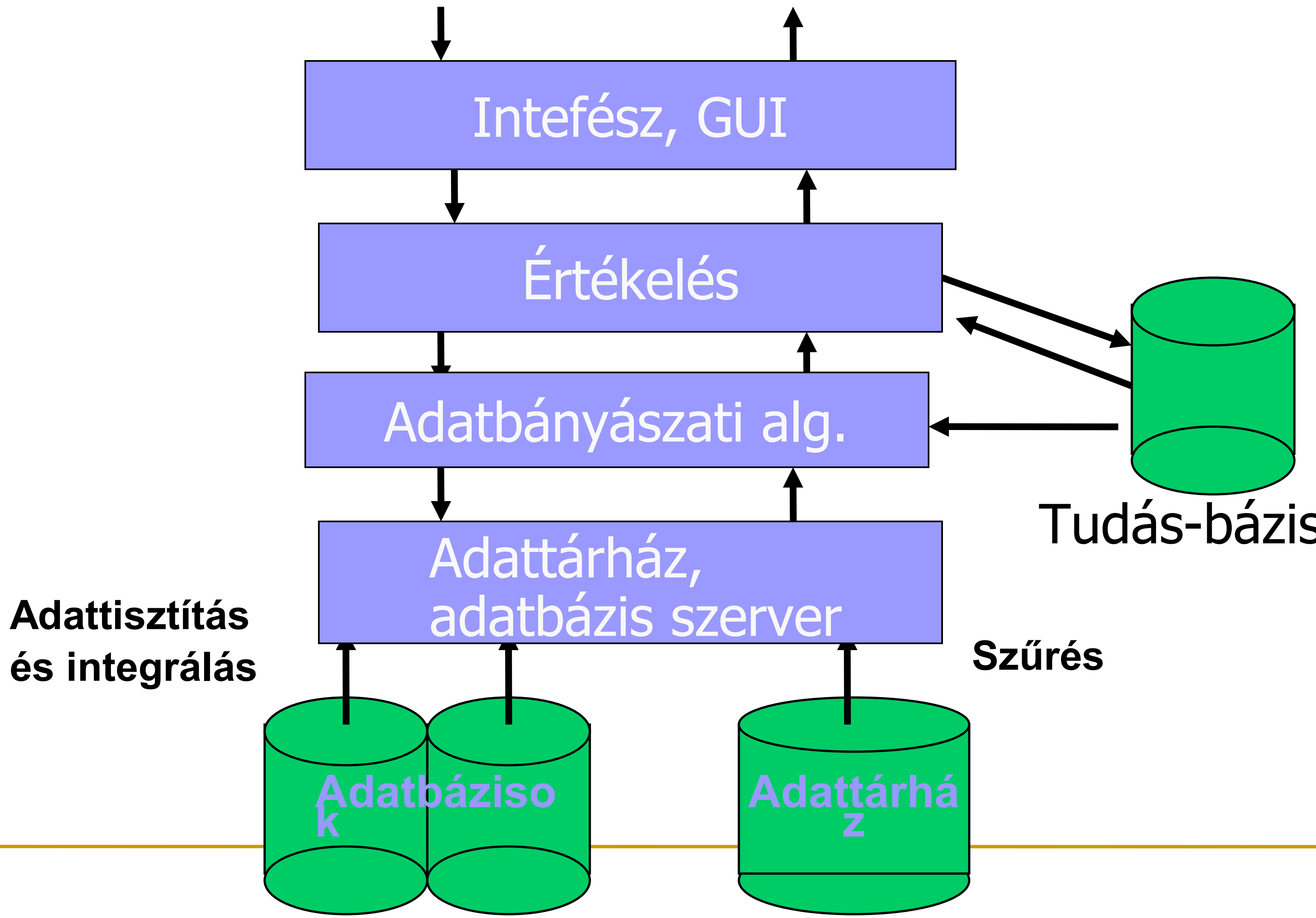
Példa adattárházra



Adattárház funkciója

- *OLTP: Hogyan vigyünk be és tároljunk adatokat ???*
- *DSS: Decision Support System, Hogyan nyerjünk ki információt ??*
- *EIS: Executive Information System, Hogyan használjuk az információt ??*
- Összefüggés és téma orientált
- Trend-adatok (időbeliség)
- gyakran nem normált
- több forrású

Tipikus architektúra



Adattárház definiálása

- Döntéstámogató adatbázis melyet külön üzemeltetnek a szervezet működéséhez kapcsolódó adatbázistól
- Támogató információ feldolgozó egység mely egy megbízható, feldolgozott hisztorikus, elemzések céljából összegyűjtött adatokat tartalmaz.

- “A data warehouse is a <u>subject-oriented</u>, <u>integrated</u>, <u>time-variant</u>, and <u>nonvolatile</u> collection of data in support of management’s decision-making process.”—W. H. Inmon

Subject oriented (tárgy- v. témaorientált)

- Témakörök köré szervezett, pl. vásárlók, termékek, eladások.
- A döntéshozók számára szükséges adatok modellezéséhez és elemzéséhez kötődik, nem a napi működéshez, illetve adatátvitelhez.
- Egyszerű és tömör nézetet nyújt a fontos témakörökben, de nem tartalmazza azokat az adatokat, melyek nem fontosak a döntéshozatal szempontjából.

Integrated (integrált)

- Több, különböző jellegű adatforrás integrálásával épül fel
  - Relációs adatbázisok, különálló fájlok, on-line adatátviteli források
- Adat tisztítási és adatintegrációs eszközöket alkalmaznak
  - Amikor az adat bekerük az adattárházba konvertálódik
  - A konzisztenciát az elnevezési konvenciók, a struktúrák, stb. biztosítja a különböző adatforrások között

Time variant (időfüggő)

- Az időhorizont sokkal nagyob mint egy operációs adatbázisban.
  - Operációs adatbázis: aktuális adatok (pl. az elmúlt nap).
  - Adattárház: hisztorikus adatok elemzésére (pl., az előző 5-10 év)
- Minden fontosabb (kulcs) struktúra tartalmaz
  - Idő elemet (explicit vagy implicit módon)

Nonvolatile (tartós)

- Változatlan adatok
- Alapvetően nem törlődő adatok

- Fizikailag külön tárolt, a működési környezetből transzformált adatok.
- Az üzemvitelből adódó adatfissítés nem fordul elő az adattárházban.
  - Nincs szükség on-line adatátvitelre, adat mentésre és vissza, és konzisztenciát biztosítű eljárásokra
  - Csak két fő adatkezelési mód:
    - *adattárház feltöltése* and *adatok lekérdezése* .

Data Warehousing

"Data Warehousing is the process, whereby organizations extract value from their informational assets through the use of special stores called data warehouses."

Három kulcsmozzanat:

- Adatkinyerés a tranzakciós (vagy más vállalat-működtetési) forrásrendszerekből
- A kinyert adatok átformálása riport (beszámoló) készítés számára
- A riportok, beszámolók elérhetővé tétele a döntéshozók számára.

Business Intelligence (BI, üzleti intelligencia) fogalma:

„Olyan módszerek, fogalmak halmaza, melyek a döntéshozás folyamatát javítják ún. *tényalapú rendszerek* használatával."

(Howard Dresdner, 1989)

Tényalapú rendszerek:

- Vezetői információs rendszerek (EIS, *Executive Information System)*
- Döntéstámogató rendszerek (DSS, *Decision Support System*)
- Vállalati információs rendszerek (*Enterprise Information System)*
- *On Line Analytical Processing* (OLAP)
- Adat- és szövegbányászat
- Adatvizualizáció
- Geográfiai Információs rendszerek (GIS)

Ezek egy szeletét fedik le az adattárház megoldások.

Business Intelligence Platform

Olyan platform, amely támogatja a következő technológiákat:

- Adattárház jellegű adattárolás
- OLAP
- Adatbányászat
- Nyílt interface-ek (OLAP, adatbányász, stb.)
- Ezeket támogató, megvalósító komponensek, eszközök

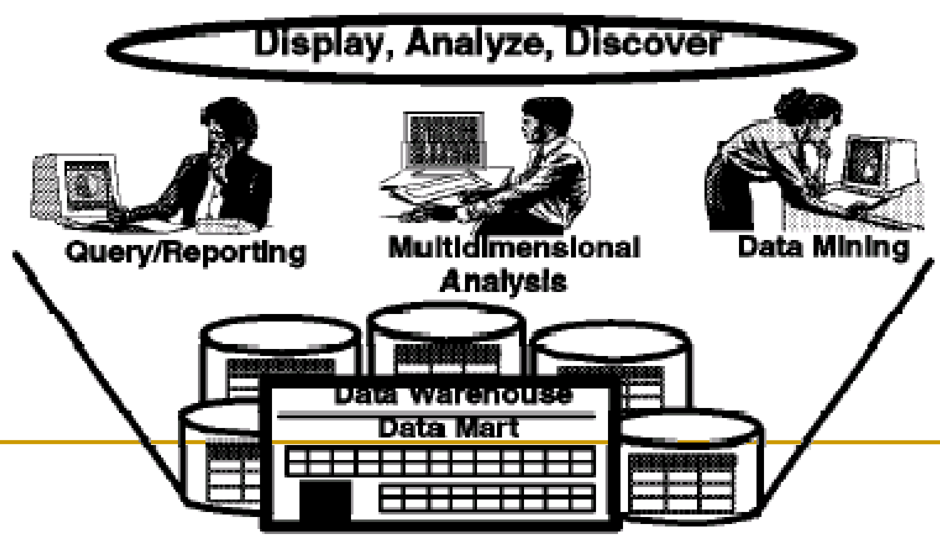
Pl.: Oracle9i, IBM DB2, MSSQL

Adattárház vs. Heterogén Adatbázisok

- Hagyományos heterogén adatbázis integráció:
  - Lekérdezés alapú megközelítésmód
    - Amikor kliens oldalról lekérdezés érkezik, egy meta-könyvtár segítségével a lekérdezés a heterogén adatbázis egy eleméhez kapcsolódó lekérdezésre fordítódik, és az egyes lekérdezések eredményei egy globális válaszá integrálódnak
- Adattárház: feltöltés-alapú integritás biztosítás, nagy teljesítmény
  - A heterogén adatforrások információi a lekérdezés előtt kerülnek integrálásra és tárolódnak
  - Direkt lekérdezésekhez és elemzések

Adattárház alkalmazásai

- Jelentések
  - a szervezeten belüli információ megosztás hatékony eszköze
  - Automatikus (web, e-mail, intranet)
  - Saját jelentések (infohoz való hozzáférés, munkamegosztás, teljes áttekintés)
- Statisztika
  - Interpretáció
  - Valószínűség
  - Minta (szignifikáns)
- Adatbányászat



Végfelhasználók igényei

- Tipikus felhasználók
  - „non-frequent user”
    - nem érdekli őket az adattárház, csak időről időre információra van szükségük
  - Előre definiált, friss jelentéseket igénylő felhasználó
    - Speciális érdeklődés, rendszeres időközönként
  - Dinamikus, ad hoc lekérdezéseket igénylő
    - Üzleti elemző
  - Profi felhasználó
    - Számára minden adat fontos
    - Specializált adatpiacok

→ *Különböző felhasználók különböző igények*

OLAP (On-line Analytic Processing)

- OLAP ötletét E.F. Codd, a relációs adatbázisok atyja 1993 -ban egy Computerworld cikkben vetette fel.
- Codd rájött, hogy az OLTP elérte alkalmazásainak határát, rendkívül nagy számítási igény szükséges amikor relációs adatbázisokból végzünk lekérdezéseket. Rájött amit már a döntéstámogatással foglalkozó szakértők már régóta hangoztattak: *pusztán az operciós adatok nem alkalmasak a menedzserek kérdéseire választ adni*.
  - Idáig a **relációs adatbázis** képes válaszolni tipikus kérdésekre mint *„ Mi?, Mit?”*
  - Az **adattárházak** a múltbeli adatok összesítésével képesek válaszolni olyan kérdésekre mint *„Mi volt a teljes forgalom a keleti régióban a második negyedévben ?”*
  - Az **OLAP** célja az adatok elemzése és megértése alapján a *„Miért?, Mi lenne ha?”* kérdések megválaszolása

OLAP II.

- OLAP és az adattárház komplementer fogalmak
  - Az adattárház tárolja és menedzseli,
  - az OLAP stratégiai információvá alakítja az adatokat
- Az OLAP alapötlete, hogy a menedzserek képesek legyenek az adatok több dimenziót figyelembe vevő kezelése, és annak megértése, hogy azok miként fordulnak elő, illetve hogyan változnak.
- Felhasználási területei:
  - Piac szegmentálás, marketing kutatás, termelés tervezés, ...
- A megoldás a „multi-dimensional” azaz több dimenziós adatbázis.

Codd 12 szabálya

- 1. Többdimenziós áttekintés
- 2. Felhasználó számára áttekinthető támogatás
- 3. Elérhetőség
- 4. Konzisztens naplók készítése
- 5. Kliens -szerver architektúra
- 6. Általános dimenzió aggregálás
- 7. Dinamikus ritka mátrixok
- 8. Multi-user támogatás
- 9. „Cross-dimensional operations”
- 10. Intuitív adatkezelés
- 11. Rugalmas jelentések
- 12. Korláttalan dimenziók

OLTP – OLAP tulajdonságok

	OLTP	OLAP
Felhasználó	adatrögzítő, informatikus	'knowledge worker'
Funkció	napról napra történő	döntés támogatás
Tervezés	alkalmazás-orientált	témakör-orientált
Adat	aktuális, naprakész, részletes, relációkba foglalt izolált	történeti, összesített, többdimenziós integrált, konszolidált
Használat	Ismétlődő	ad-hoc
Elérés	írás/olvasás	Sok lekérdezés
Munka egysége	rövid, egyszerű tranzakciók	Komplex lekérdezés
Elért rekordok száma	tizes nagyságrend	milliós nagyságrend
Felhasználók száma	ezres nagyságrend	százas nagyságrend
Méret	100MB-GB	100GB-TB
Mérték	Tranzakciós idő	Lekédzezési idő

Mikor használjunk OLAP-ot ?

- Az adatok iránti igény nem tranzakciós hanem elemző jellegű
- Az elemzett információ nem elérhető közvetlen módon
- Jelentős számítás és összesítés igény
- Főként numerikus adatok
- Az elemek, melyek az adatpontokat definiálják nem változnak időben

Miért külön adattárház?

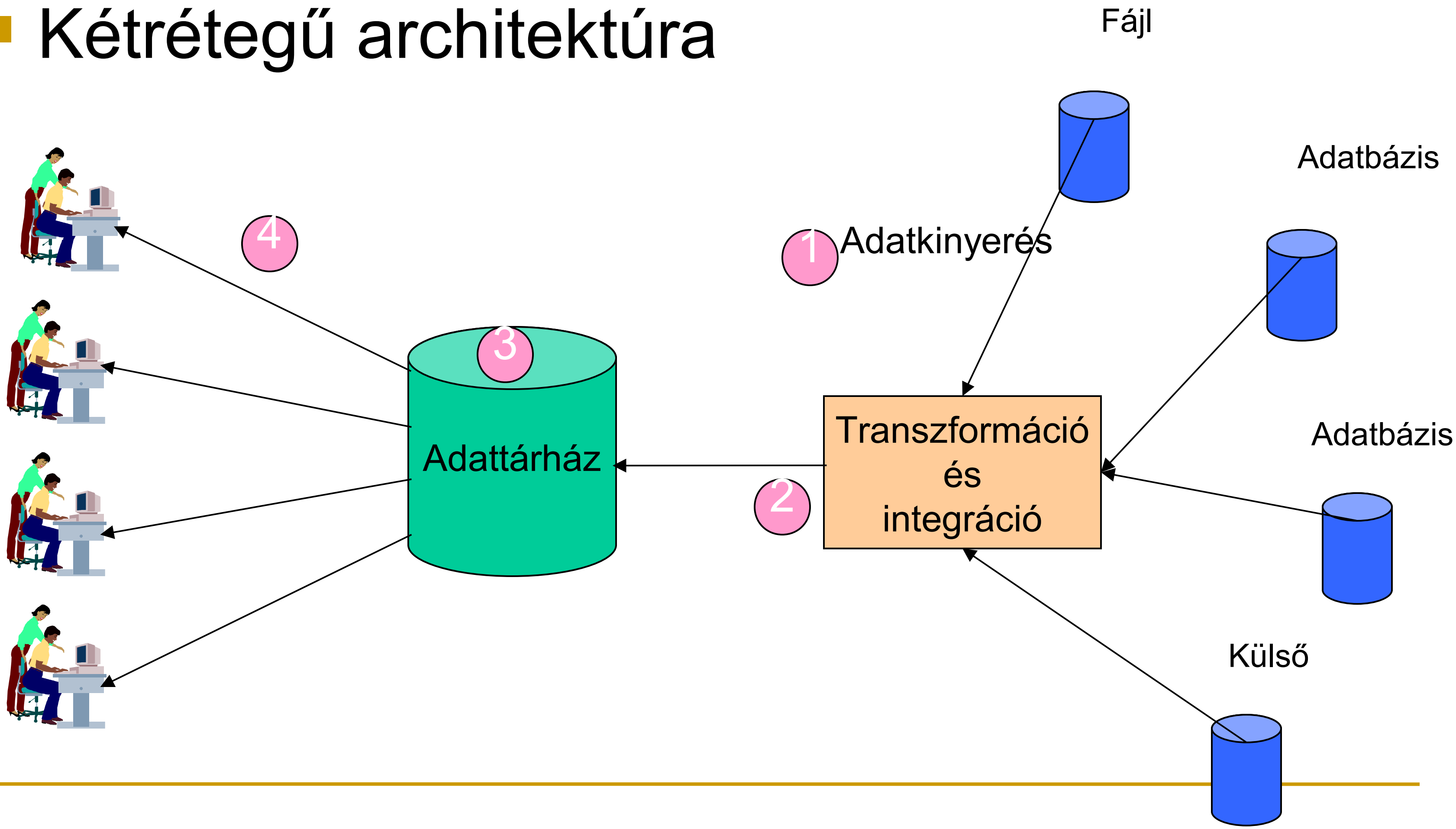
- Mindkét rendszer jó teljesítményt nyújt
  - Relációs adatbázis—OLTP-re hangolva: elérési módok, indexelés
  - Adattárház—OLAP-ra hangolva: összetett OLAP lekérdezések, többdimenziós nézet, konszolidáció.
- Különböző funkciók és különböző adatok:
  - Hiányzó adatok: Döntéstámogató rendszer olyan **hisztorikus** adatokat kíván melyeket egy tipikus relációs adatbázisban nem tárolnak
  - Adat konszolidáció: Pl. heterogén forrásból származó adatok aggregálása, összegzése
  - Adat minőség: Különböző adatforrások általában inkonzisztens reprezentációt alkalmaznak, pl. idő formátumok

Adattárház vs. Heterogén Adatbázisok

- OLTP (on-line transaction processing)
  - A hagyományos relációs adatbázisok alapfeladata
    - Napról napra történő működés: vásárlás, bank, gyártás, regisztráció, számlázás, stb.
- OLAP (on-line analytical processing)
  - Az adattárházak alapfeladata
    - Adatelemzés és döntéshozatal
- OLTP vs. OLAP:
  - Felhasználó és rendszer orientáltság: vásárló vs. piac
  - Adat tartalom: aktuális, részletes vs. történeti, konszolidált
  - Tervezési módszer:
    ER (entity-relationship) + alkalmazás vs. csillag + témakör
  - Nézet: aktuális, lokális vs. evolóciós, integrált
  - Hozzáférés: frissítés vs. csak olvasható de komplex lekérdezések

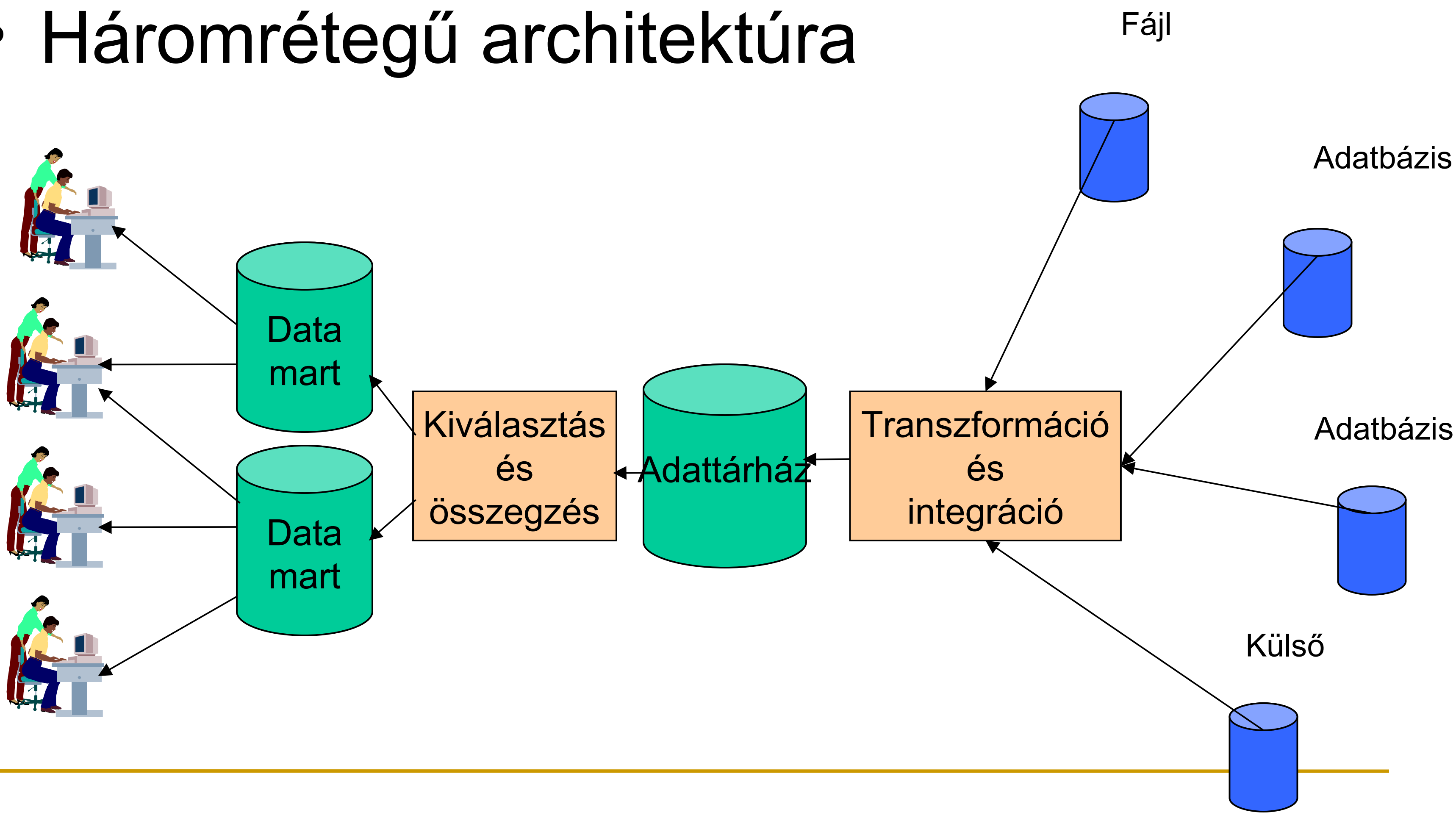
Az adattárházak architektúrája

- Kétrétegű architektúra



Az adattárházak architektúrája

- Háromrétegű architektúra



Speciális adattárház típusok

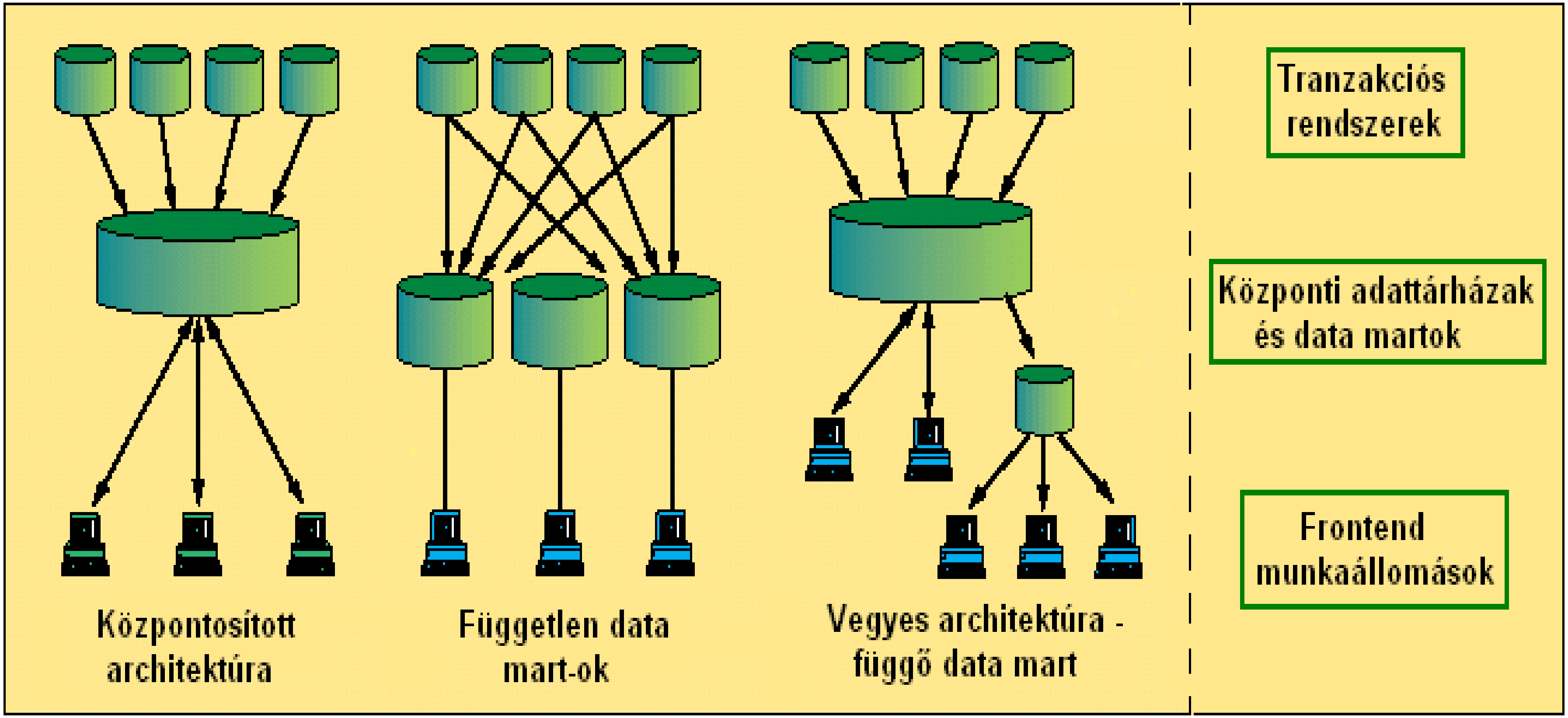
Jól skálázható technológia:

- *Data Mart* (adatpiac)
  lokális, szűk felhasználói kör, konkrét feladatok, kis adatfeldolgozó és analizáló egység adattárház funkciókkal
- *Operational Data Store (ODS)*
  Adatok tisztítására, gyűjtésére használt egység, teljes részletezettségű operációs adatokkal
- *Extraprise Data Warehouse*
  Helyi megkötés nélkül összefutnak benne B2B és B2C adatok, elemzési céllal
- Virtuális adattárház
  Nem épül külön rendszer az adattárház adatainak számára, azt az OLTP rendszer keretein belül valósítják meg

Az adat útjának fő állomásai

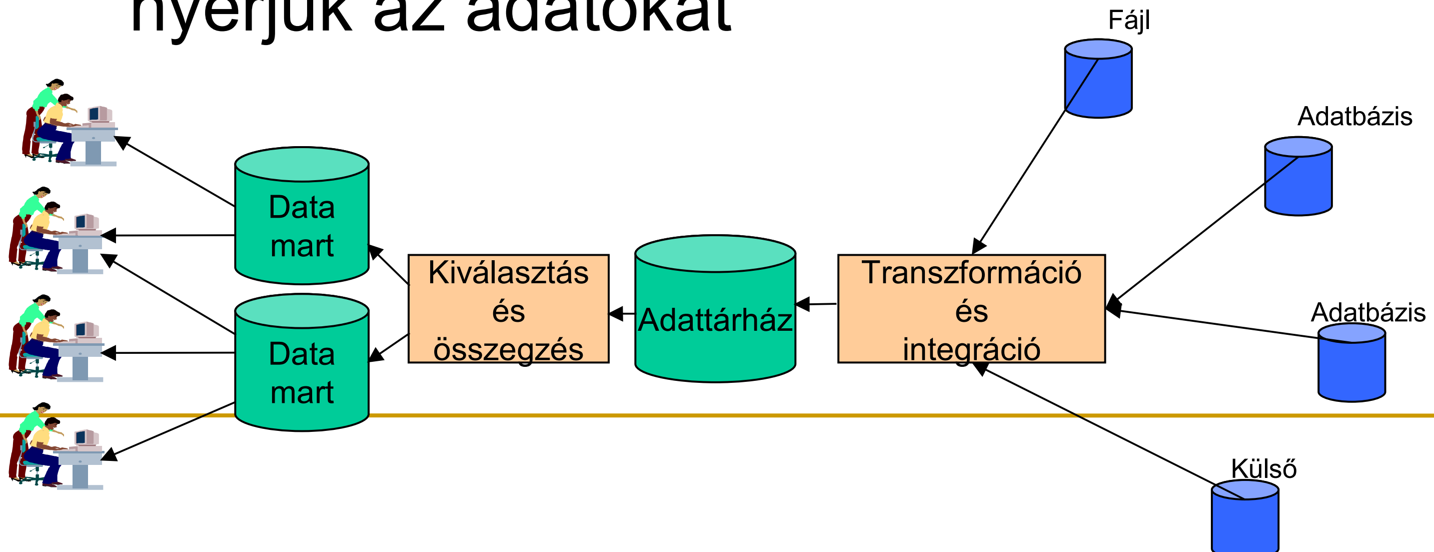
- Forrásrendszerek
- Adattárház
- Elemző frontend alkalmazások

Architektúra változatok
(kliens-szerver modellek)



Független és kapcsolódó DataMart-architektúra

- Kapcsolódó DataMart
  - csak a DataWarehouse-ból nyeri az adatait
- Független DataMart
  - nincs DataWarehouse, közvetlenül a forrásból nyerjük az adatokat



Ha több DataMartra van szükségünk, akkor mindenképpen alkalmazzunk egy közbülső DataWarehouse-t.

- Ha nincs középső réteg
  - inkonzisztens adatok a különböző forrásokból
  - redundáns adattárolás
  - nincsenek integrálva az adataink
  - több-platformon átívelő kapcsolatok (JOIN)
  - a különböző felhasználóknak különböző frissességű adatok kellenek

OLAP elemzések

OLAP elemzések

- Multidimenzionális adatnézet
- Intuitív kezelőfelület, rugalmas lekérdezések
- On-line, válaszidő orientált szolgáltatás
- Közép-felsővezetők
- Lehetőség összetett elemzésekre, látványos, jól használható vizualizációra

Adattárházak - adatbányászat

- Adatbányászat- Data Mining -DM: „Érdekes (nem triviális, implicit, eddig ismeretlen és valószínű hasznos) információk vagy mintázatok nagy adatbázisok tartalmából való kinyerése.”
- OLAP korlátok: adatmennyiség, lekérdező nyelv

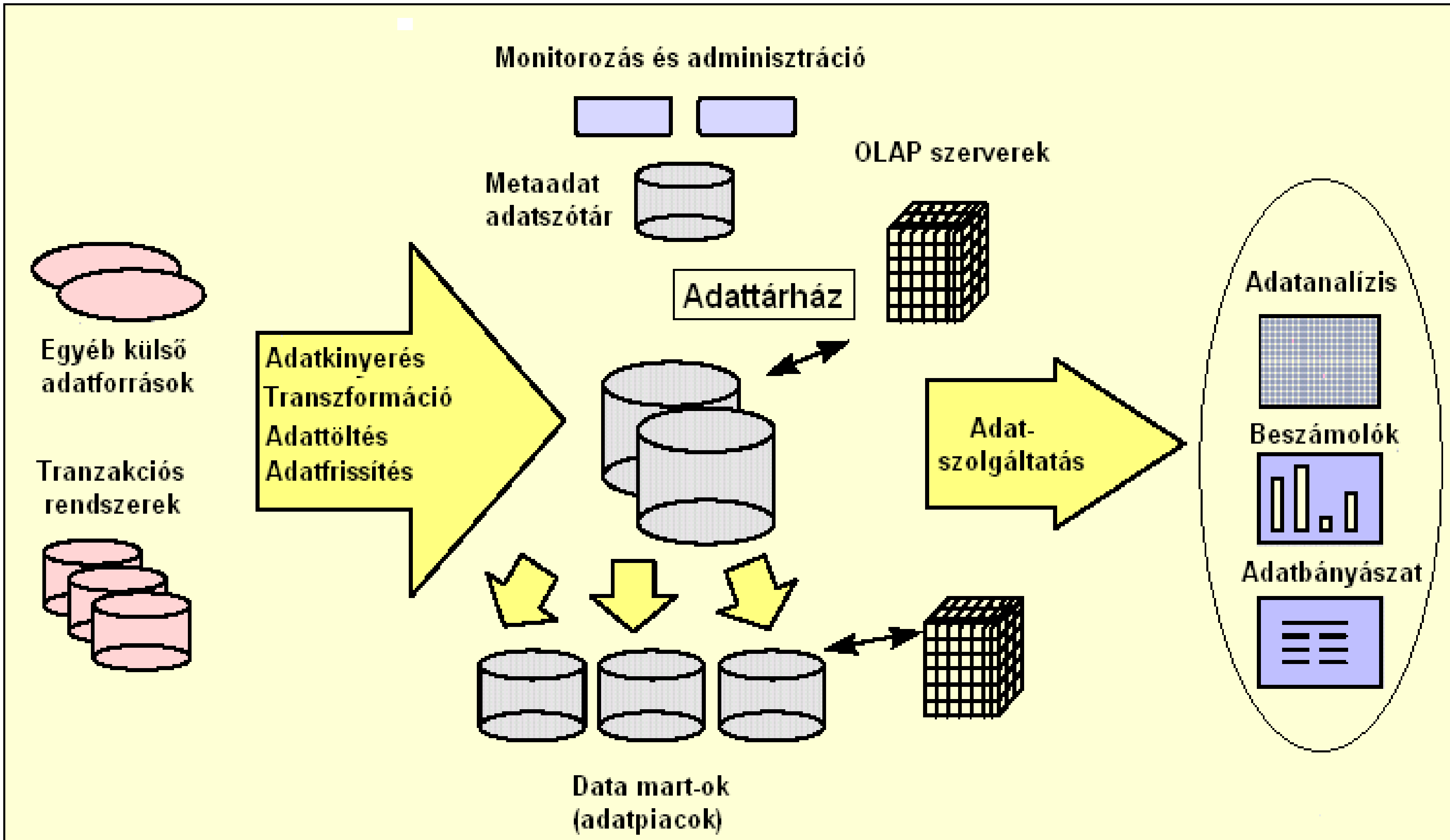
Tudáskinyerés folyamata

1. Alkalmazási terület felmérése, előzetes ismeretek rendszerezése
2. Céladatbázis kiválasztása, létrehozása
3. Adattisztítás, előfeldolgozás
4. Adatintegráció
5. Adattér csökkentés: cél szempontjából fontos attribútumok kiemelése
6. Adatbányászati algoritmusok kiválasztása (klaszterezés, mintakeresés, osztályozás)
7. Adatbányászati algoritmus, paraméterek előállítása
8. Algoritmus alkalmazása
9. Kinyert információ értelmezése, finomítások
10. A megszerzett tudás megerősítése, összevetése az elvárásokkal, dokumentálás

Adattárházak - adatbányászat

- Az adattárházak megfelelő alapot biztosíthatnak adatbányász módszerek alkalmazásához
- Részben hasonló célok
- OLAP elemzések – adatbányász elemzések: jól kiegészíthetik egymást
- Probléma: OLAP jellegű és adatbányász rendszerek hatékony, rugalmas illesztése
- Megoldást jelentheti:
  - Következtetési szabályok a DW-ben (induktív adatbázisok)
  - Megfelelő adatbányász interface alkalmazása (még nincs elfogadott szabvány)

Adattárház komponensek



Komponens csoportok

ETL: *Extraction Transformation and Load*

- Adatkinyerés az operatív rendszerekből *(extraction)*
- Adattranszformáció (különböző adatformátumok, mértékegységek, nyelvek stb.)
- Adatminőség ellenőrzése, adattisztítás (cleaning)
- Adatbetöltés az adattárház struktúráiba (loading)

Komponens csoportok 2.

- OLAP Tools:
  OLAP lekérdezéseket lehetővé tévő komponensek (OLAP szerver, interface-ek)
- Felügyelet, adminisztráció
  adattárház működtetése, felügyelete

Metaadat kezelés

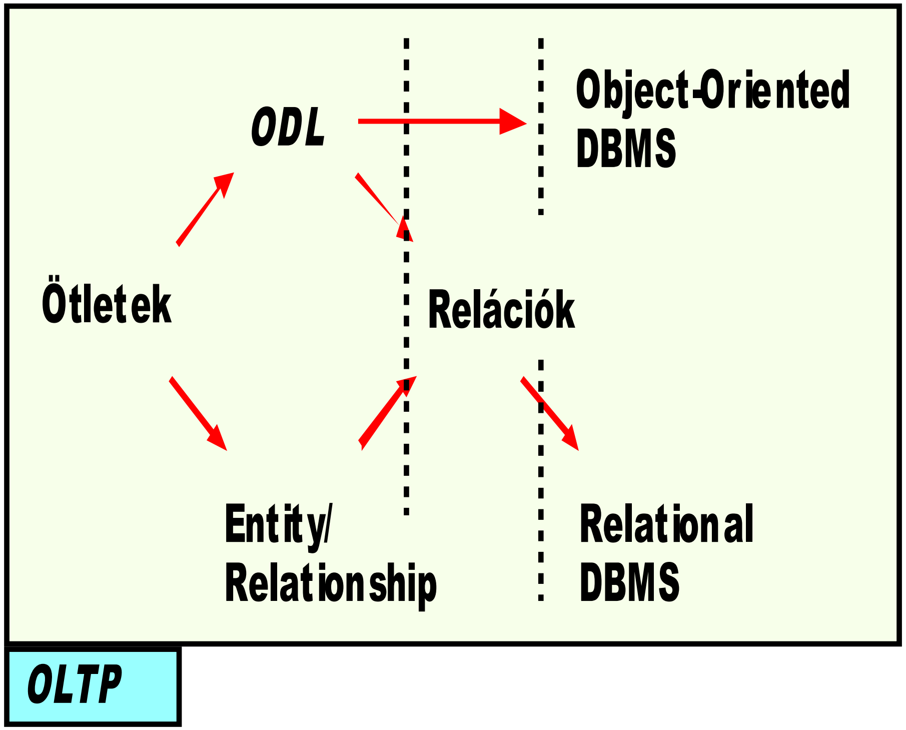
Metaadat: „adat az adatokról”

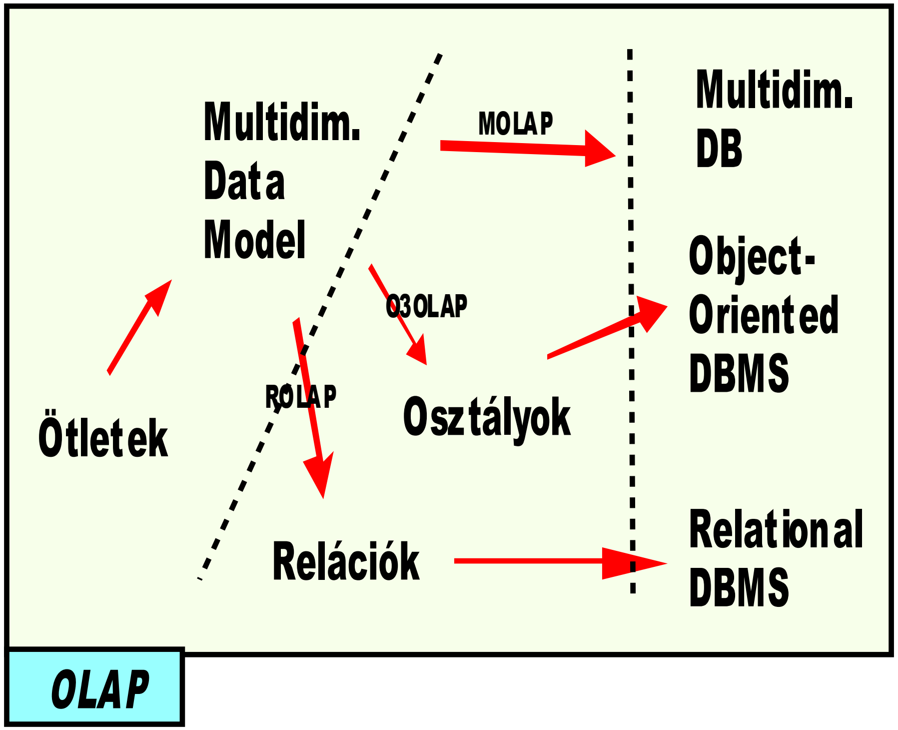
- Az adattárház szerkezetét, a bent lévő adatok jellemzőit tároló szerkezet
- Fontos: adatintegrációhoz szabványos adatkezelés
- A megfelelő metaadat kezelési stratégiát gyakran említik mint az adattárház projekt kulcskérdését
- Példa: adatkockáink leírása, az adattöltéseink eredményei, az adatforrások mezőinek jelentése, stb.

Komponens csoportok 3.

- Frontend adatelemző alkalmazások
  OLAP elemzők, adatbányász eszközök, vizualizáció, egyéb kliens alkalmazások
- Adatbázis komponensek
  - ROLAP: relációs OLAP – relációs adatbáziskezelő
  - MOLAP: multidimenzionális OLAP, közvetelen multidimenzionális adattárolás
  - HOLAP: hibrid OLAP - keverék

Adatmodellezés
(koncepcionális, logikai, fizikai)





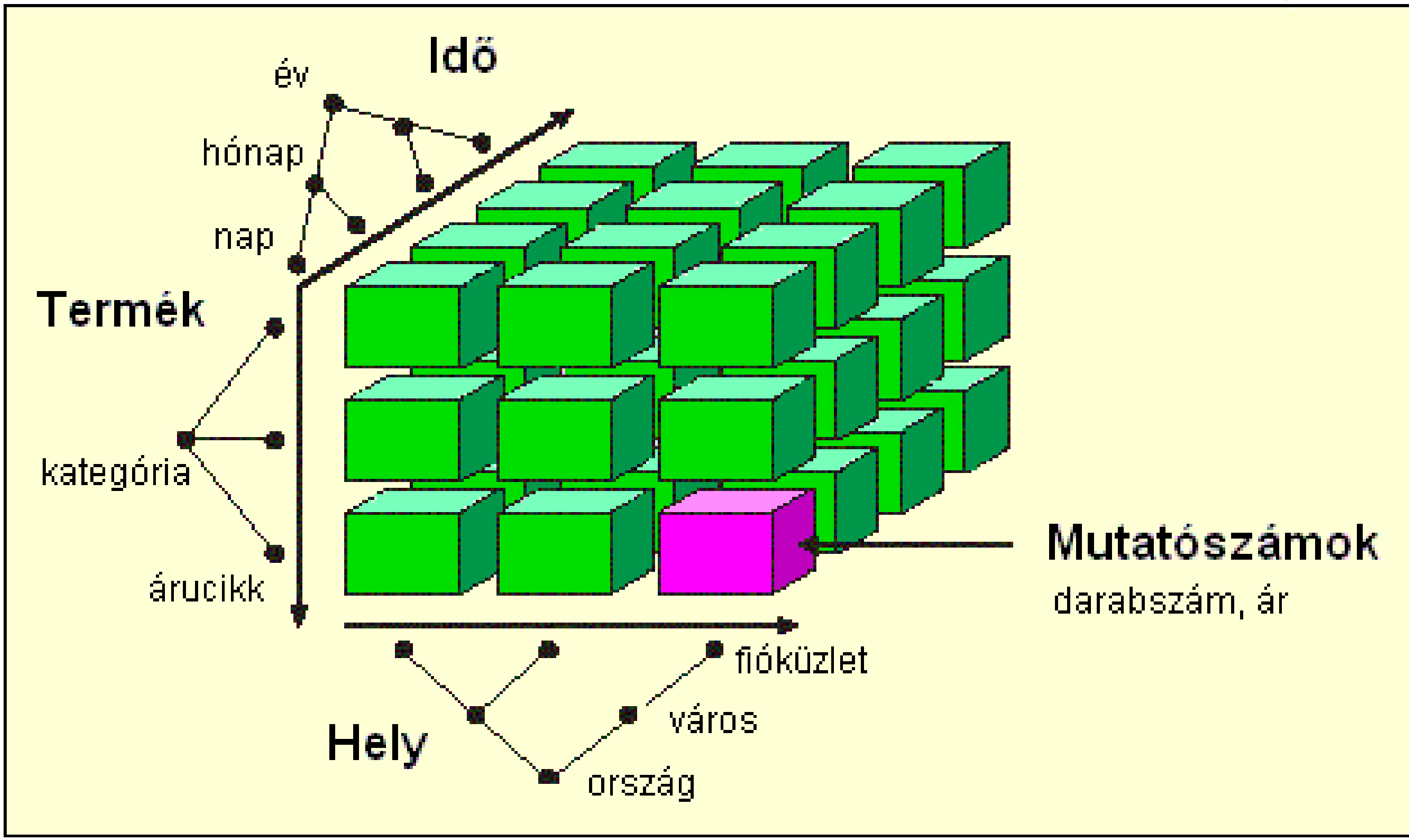
OLTP adatmodellek

- Hagyományos, kiforrott módszerek
- Relációs adatmodell
- Relációs algebra alapú lekérdezőnyelvek, SQL
- Egyed/Kapcsolat Modell (E/R M), UML

A multidimenzionális adatmodell (MD, multi dimensional data model

- A multidimenzionális adatmodellben a multidimenzionalitás arra utal, hogy itt az elemi adatokat nemcsak egy kulcs függvényében lehet elérni, hanem több kulcstól való függése is nyilvántartott az adatbázisban.
- Az egyes kulcsok mint dimenziók szerepelnek az adatelemek elérésekor.
- Az adatelemek ábrázolására ekkor egy kockát szoktak alkalmazni, amit adatkockának neveznek.
- Az egyes dimenziók itt egyes kulcsokhoz tartoznak
- Az alkalmazások döntő többségében a modellezett problémakör nem annyira egyszerű, hogy egyetlen egy adatkockával leírható lenne. (mindegyik kockát külön szerepeltetik a sémában (vs EK modell))

Adatkocka példa: nemzetközi kereskedelmi cég értékesítési adatainak multidimenzionális nézete



OLAP multidimenzionális adatfogalma

Fogalmak:

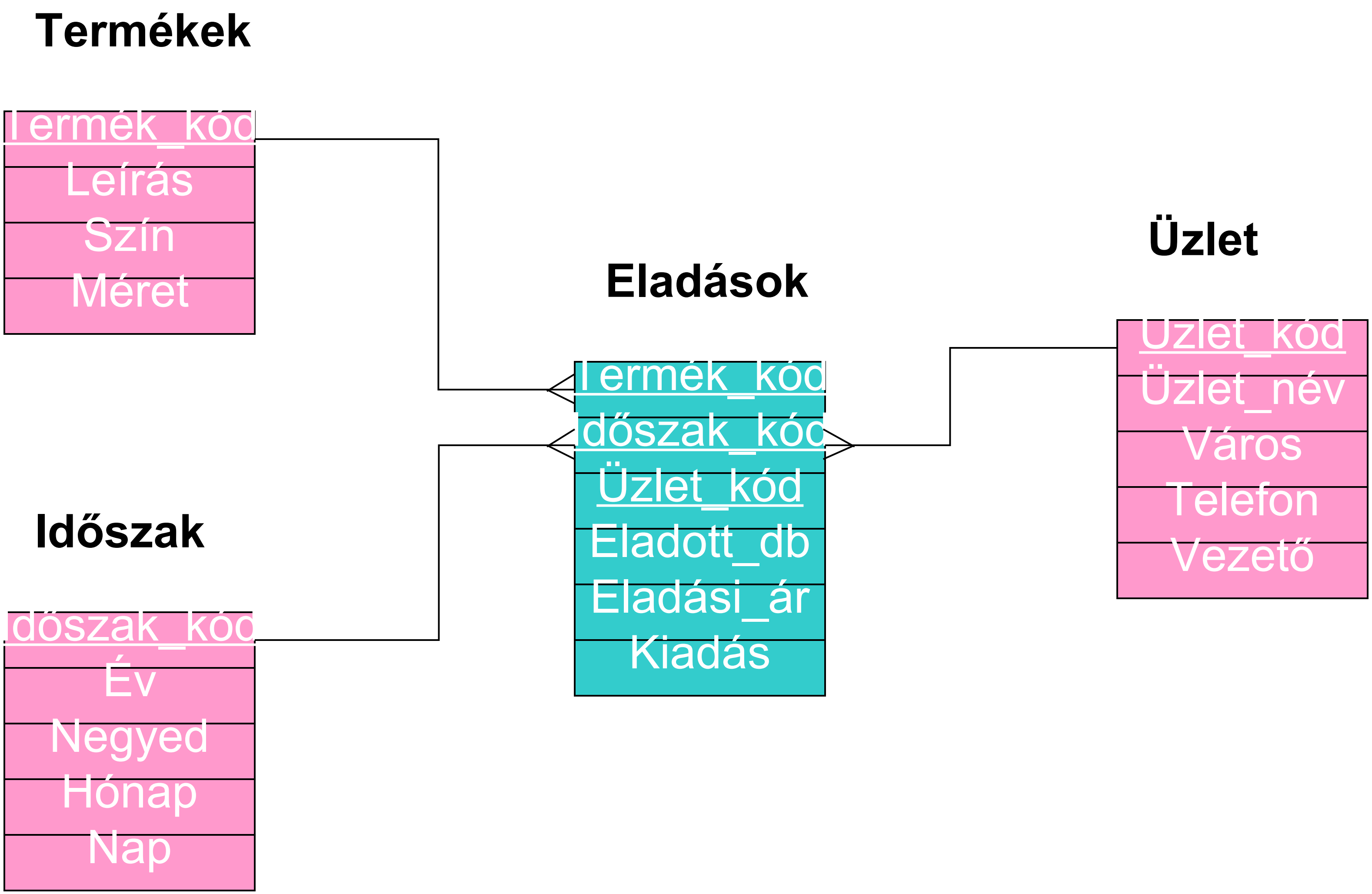
- Ténytábla (mutatószámok)
  - Amit tárolunk az adatkockában, aminek az értékeit vizsgáljuk.
  - A ténytábla tulajdonképpen egy összekapcsoló entitás
  - üzleti egységet, tranzakciót, eseményt jelöl
  - Kulcs tábla, melyben numerikus adatok szerepelnek

- Dimenziók (jellemzők)
  - A tények hátterét definiálják, leíró jellegű adatok (pl. idő, hely, üzletkötő …)
  - Gyakran nem numerikus egységek
    - pl. termék márka, alkalmazott
    - Diagrammokban tengelyként ábrázolva
  - Paraméterek, melyekre OLAP elemzést szeretnénk végezni
    - pl. Idő, Hely, Vásárló ...

- Dimenzió-hierarchiák
- N-dimenziós adatkocka

Csillag-séma

A csillag modell célja az adatkocka szerkezetének megadása.



Csillag-séma

Termék_kód	Leírás	Szín	Méret
100	Pulóver	Kék	40
110	Cipő	Zöld	39
125	Kesztyű	Barna	M

Időszak_kód	Év	Negyed	Hónap
100	1999	1	1
110	1999	1	2
125	1999	1	3

Termék_kód	Időszak_kód	Üzlet_kód	Eladott_db	Eladási_ár	Kiadás
110	002	S1	30	1500	1200
125	003	S2	50	1000	600
100	001	S1	40	1600	1000
110	002	S3	40	2000	1200
100	003	S2	30	1200	750

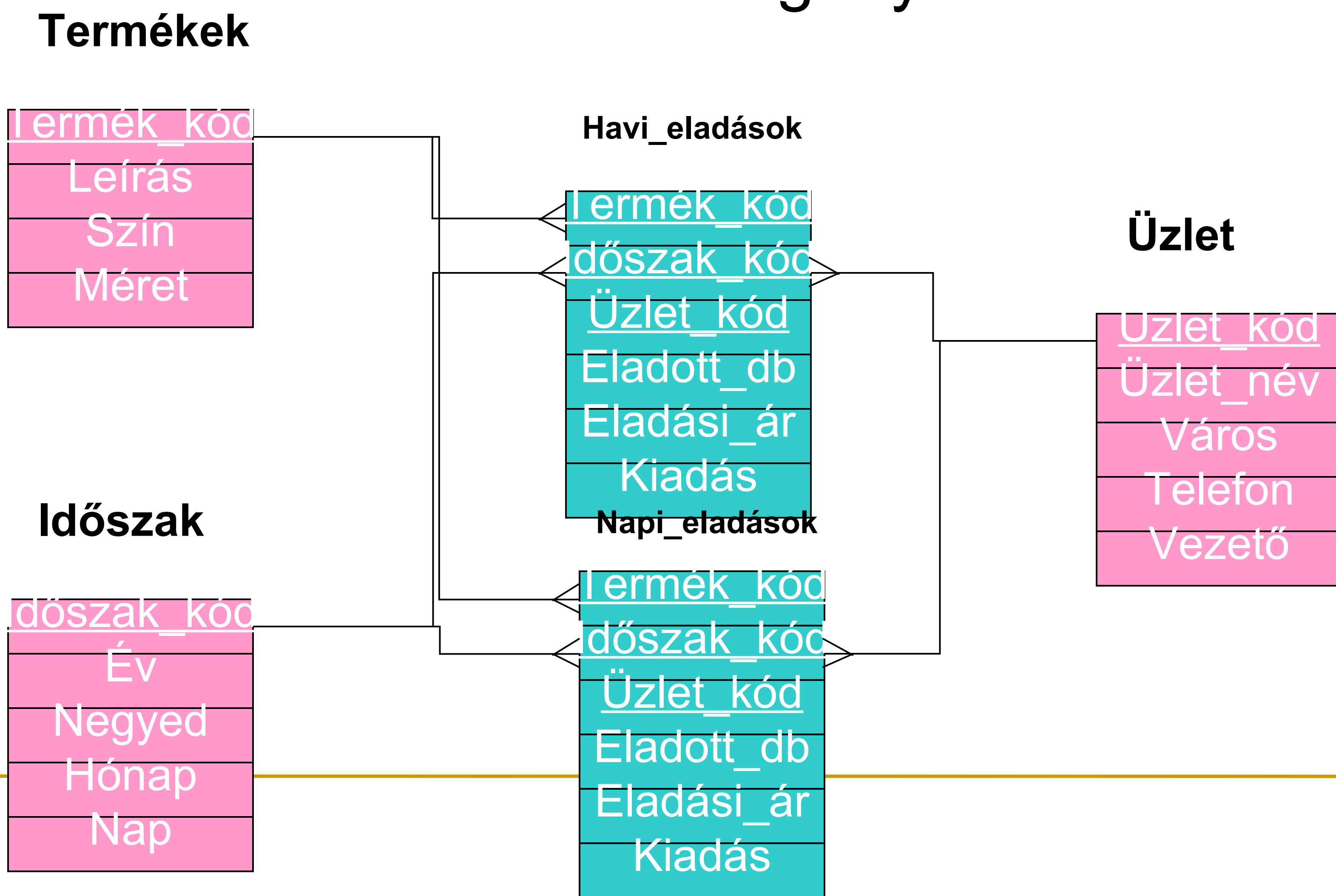
Üzlet_kód	Üzlet_név	Város	Telefon	Vezető
S1	Újpest	Bp	432-3243	Kovács
S2	Pólus	Bp	654-5464	Lajtai
S3	Mammut	Bp	234-4353	Csurka

A ténytábla mérete

- Az adatok granularitása
  - év, negyedév, hónap, nap
- Üzletek száma: 1000
- Termékek száma: 10,000
- Időszakok száma: 24 (két év)
- sorok száma = 1000 * 5000 (aktív) * 24 = 120,000,000 (24byte/rekord: 2.88GB)
- Napi adatok esetén 34.56GB

Több ténytábla

- Különböző felhasználók különböző igényekkel



Csillagséma tulajdonságai

Előnyök:

- Egyszerű, intuitív adatmodell
- Kevés *join* művelet lekérdezésekhez
- Kevés tábla olvasása
- Könnyű megvalósíthatóság, a modell leíró adatai egyszerűek

Hátrányok:

- Nehézkes aggregátum (összeg) képzés
- Nagy dimenziótáblák esetén a hierarchiák kezelése nagyban lassítja a lekérdezéseket
- Dimenzióelemek tárolása redundáns, denormalizált (vagyis tárhhely-pazarló)

Egyéb csillagséma variánsok

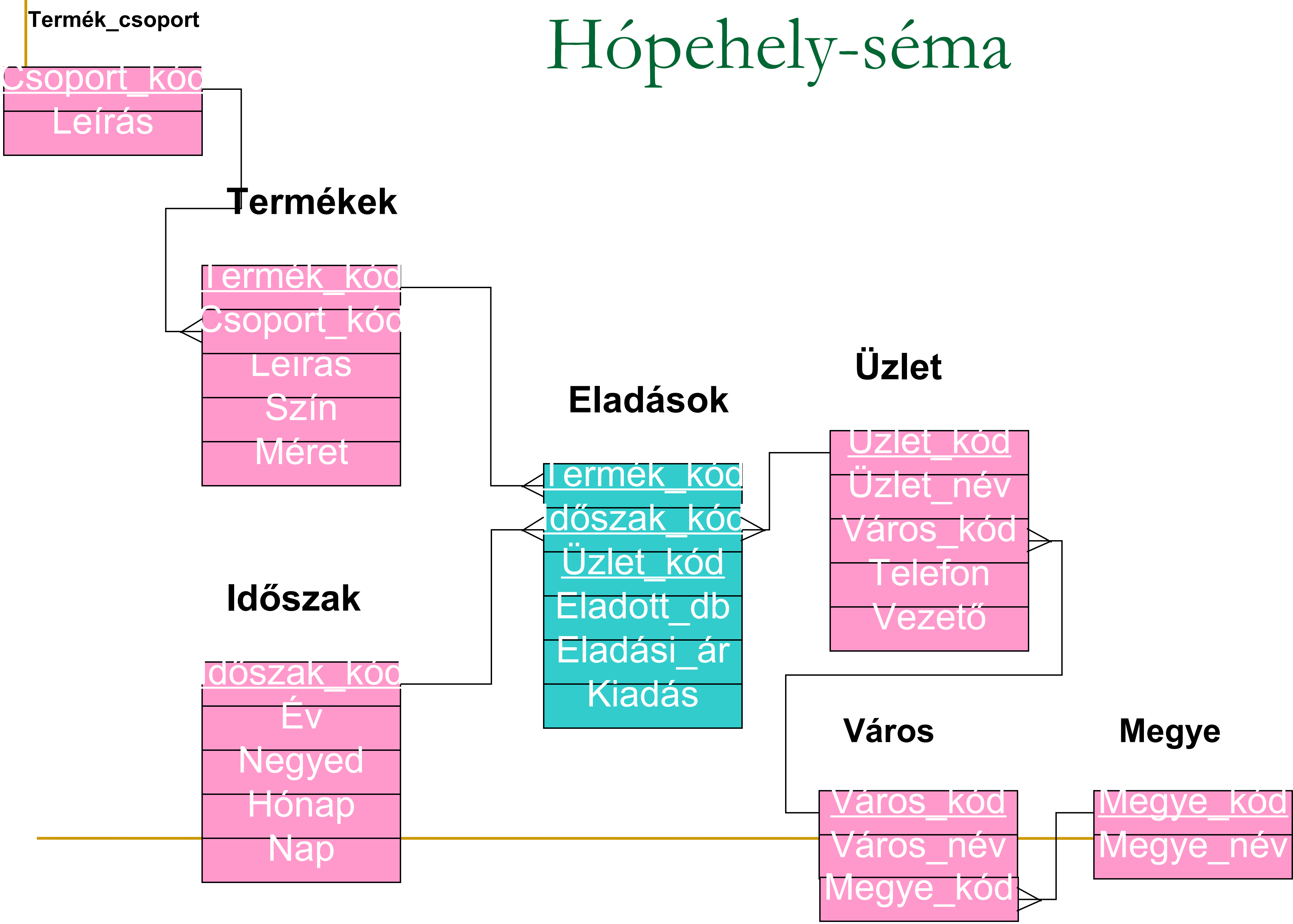
- Hópehely séma
  normalizált dimenziótáblák (pl. hierarchiaszerkezetek kialakítása, stb. – hagyományos normalizálás folyamata)
- Konszolidált csillagséma
  aggregált adatok tárolása a ténytáblában
- „Terraced” séma – a szélsőséges eset
  egyetlen, elfajult ténytáblából álló séma
- Galaxis séma
  több adatkocka megvalósítása külön ténytáblákkal, de közösen használt dimenziótáblákkal
- „Fact consellation schema”
  hierarchikus kapcsolatban álló ténytáblák

Hópehely-séma

- A dimenziók egy természetes hierarchiába rendezhetők
  - Üzletek városban
  - Városok megyékben

  - Termékek csoportosítása
- Denormalizált alak: egy tábla
- Normalizált alak: több tábla

Hópehely-séma



Analízisoperátorok

Műveletek: adatkocka → adatkocka

- Aggregáció (roll up)
  dimenzió elhagyása v. lépés hierarchiában felfelé
- Lefúrás (drill down)
  áttérés nagyobb részletezettségre
- Pivoting
  adatkocka elforgatása
- Szelekció (selection, filtering)
  konkrét jellemzők kiválasztása
- Szeletelés (slicing and dicing)
  adatkocka szeletének kiválasztása, részkocka kiválasztása

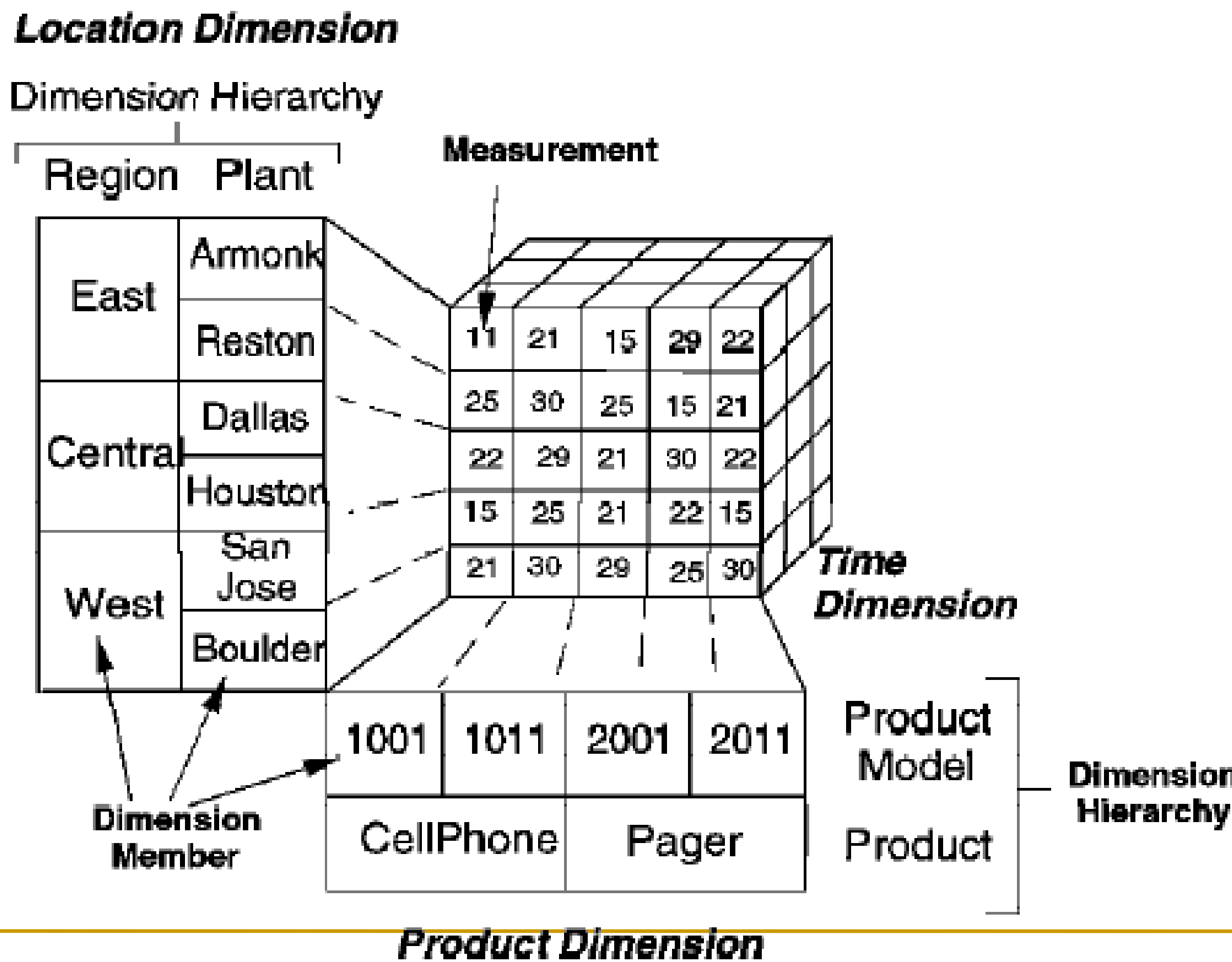
Hatékony adatkocka kezelés

- Az adatkocka cuboidok hálójaként értelmezhető
  - A legalsó cuboid az alap cuboid
  - A legfelső cuboid (apex) csak egy cella
  - Hány cuboid fordul elő egy n-dimenziós L szintből felépülő adatkockában?

$$T = \prod_{i=1}^{n} (L_i + 1)$$

- Adatkocka materializációja
  - full materialization, Minden cuboid kiszámítása és tárolása
  - no materialization,
  - partial materialization, Csak néhány cuboid materializációja, a lekérdezések gyakorisága, a méret, stb. alapján

Tipikus Olap Műveletek 0. Példa



Olap Műveletek

Aggregáció (roll up)

Egy adott dimenziót kihagyunk a felbontásból, azaz a dimenzió elemein végighaladva az adatokat felösszegezzük. Előfordulhat az is, hogy a dimenzió felbontását nem teljesen hagyjuk ki, hanem áttérünk egy kisebb elemszámú hierarchia alkalmazására az adott dimenzióra. (Pl. városok helyett országok szerint nézzük adatainkat)

Lefúrás (drill down, roll down)

Ennek ellentéte, mikor egyre részletezettebben nézzük az adatokat. Pl. felbontjuk az összesített eladási adatokat termékekre, vagy a havi összesített adatokat lebontjuk napi adatokra.

Pivoting

Az adatkocka elforgatását értjük alatta. A kocka felbontása marad, csak a dimenziókat cseréljük fel, ezáltal más nézetét kapva az adatoknak.

Szelekció (selection, filtering)

Ebben az esetben egy adott dimenzió egy adott elemét kiválasztjuk, és a hozzá tartozó adatokat nézzük, a többi adatot pedig figyelmen kívül hagyjuk. Ilyen pl., ha kíváncsiak vagyunk egy konkrét fióküzlet bevételeinek alakulására.

Szeletelés (slicing and dicing)

Slicing alatt a szelekcióhoz hasonlóan azt értjük, mikor adott dimenziót fix értékkel lekötünk, és így nézzük a kocka nézetét, szeletét. Dicing alatt a kocka egy részkockájának kivágását értjük.

Tipikus OLAP Műveletek I.

- Roll up (drill-up):
  adatok összegzése
  - *A hierarhikus dimenziók összesítése (nap vs. év) vagy dimenzió redukció (pl. nem érdekel minket a hely)*

Volume of Prod (numbers in 1000)		1996			
		Qtr 1	Qtr 2	Qtr 3	Qtr 4
West	San Jose	78	45	34	56
	Boulder	90	67	87	91

Roll Up

Dimension: Time

Volume of Prod (numbers in 1000)		Quarter 1		
		Jan	Feb	Mar
West	San Jose	30	26	22
	Boulder	28	30	32

- Drill down (roll down):
  a roll-up ellentettje
  - *Nagyobb szintű összesítésből részletekre bontás, illetve új dimenziók bevezetése*

Volume of Prod (numbers in 1000)		CellPhone		Pager	
		1001	1011	2001	2011
West	San Jose	33	12	8	12
	Boulder	45	34	20	23

Drill-Down

Dimension: Location

Member: San Jose

Volume of Prod (numbers in 1000)		CellPhone		Pager	
		1001	1011	2001	2011
San Jose	Team1	20	8	6	7
	Team2	13	4	2	5

Tipikus OLAP Műveletek II.

- **Slice and dice:**
  - *Projekció és szelekció*
- **Pivot (rotate):**
  - *A kocka átszervezése, megjelenítés, 3D mint 2D síkok halmaza.*
- Más műveletek
  - *drill across:* *ténytábla használata*
  - *drill through:* *alsó szintjének és annak relációs táblájánal (SQL)*

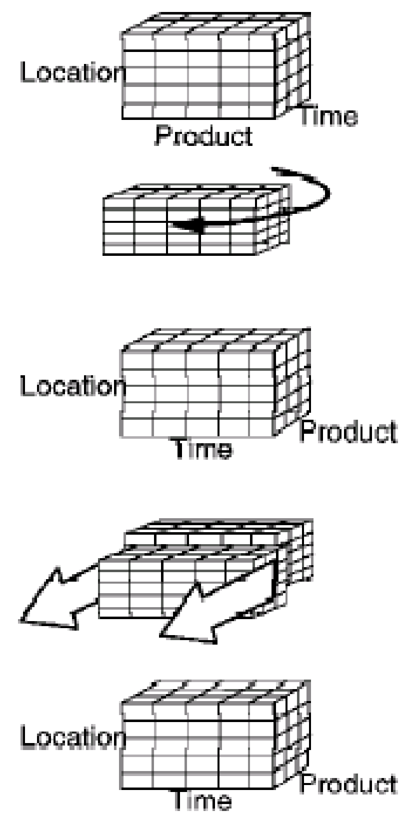
Volume of Prod. (numbers in 1000)		CellPhone		Pager	
		1001	1011	2001	2011
West	San Jose	33	12	8	12
	Boulder	45	34	20	23

Dice

Volume of Prod. (numbers in 1000)		1996 (CellPhone & Pager)			
		Qtr 1	Qtr 2	Qtr 3	Qtr 4
West	San Jose	78	45	34	56
	Boulder	90	67	87	91

Slice

Volume of Prod. (numbers in 1000)		1996 (CellPhone Only)			
		Qtr 1	Qtr 2	Qtr 3	Qtr 4
West	San Jose	53	35	20	48
	Boulder	76	57	40	80



Oracle Warehouse Builder

• Integrált eszköz vállalati adatok kezelésére üzleti intelligencia alkalmazásokhoz

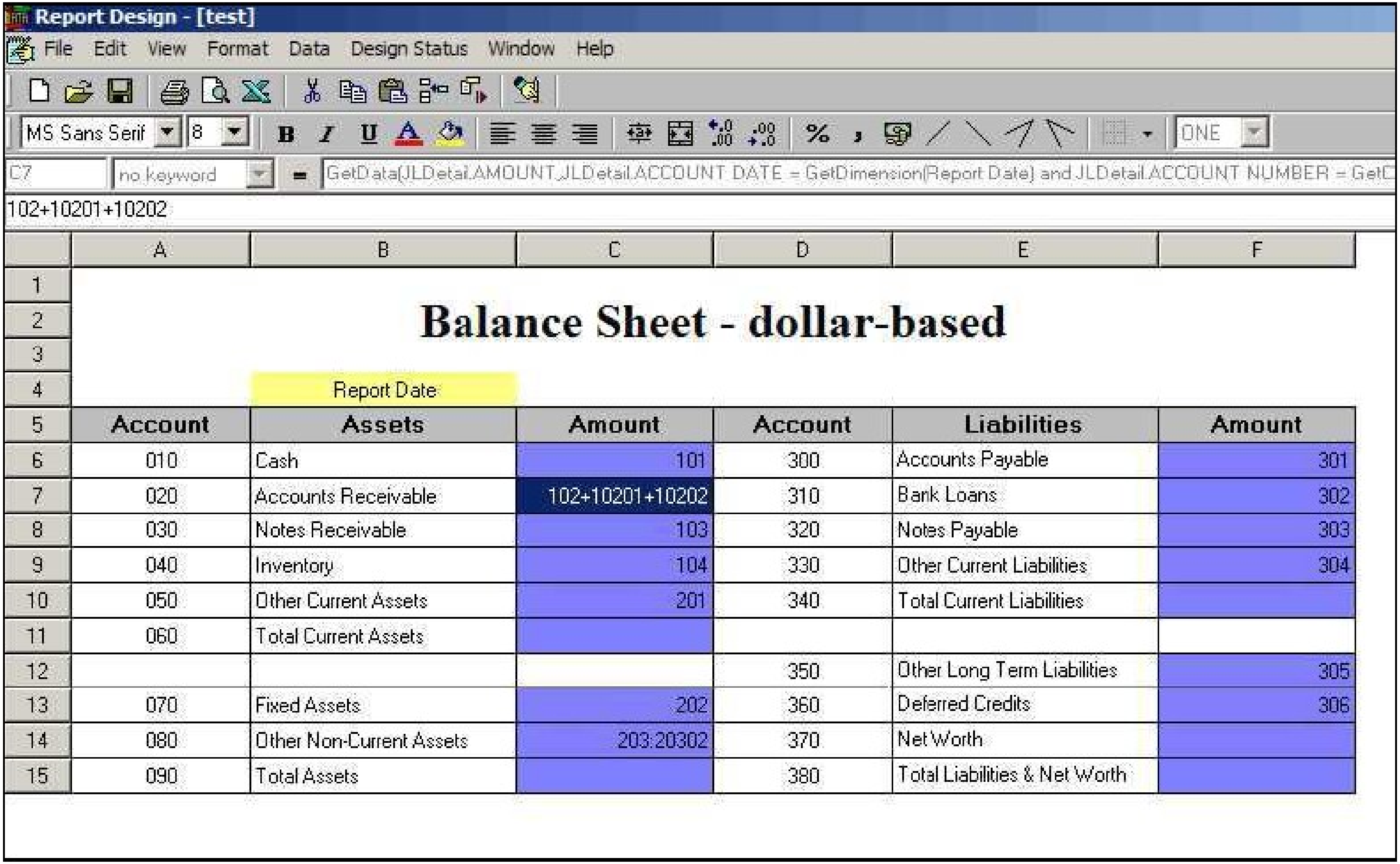
• Segítségével nagyban csökkenthető az ETL folyamatokra (adatkinyerés, transzformálás és adattöltés) fordított idő és költsége

• Vizualizált, grafikus felületet biztosít az adatok útjának menedzselésére

•ETL folyamat támogatása mellett adattárház vagy adatpiac tervezésére is alkalmas

• Képes az Oracle9i adatbázis speciális adattárház-célzatú funkcióinak használatára valamint a multidimenzionális OLAP engine használatára

Oracle Warehouse Builder



SQL bővítések: Group by kiegészítői: ROLLUP, CUBE operátorok

- ROLLUP(o_kif1, ... , o_kifN) elvégzi a csoportosításokat az első N, N-1, N-2, … 0 darab o_kif szerint. Ezek közül az első N darab o_kif szerinti adja az eddigi normál gyűjtősorokat, a többi pedig a szupergyűjtő-sorokat. Összesen N+1 darab különböző csoportosítás van végrehajtva.

  - SELECT t_kod, beosztas, AVG(fizetes), COUNT(*)
    FROM alkalmazott
    GROUP BY ROLLUP(t_kod, beosztas);

SQL bővítések: Group by kiegészítői: ROLLUP, CUBE operátorok

- CUBE(o_kif1, ... , o_kifN) elvégzi a csoportosításokat az o_kif -ek összes lehetséges kombinációja szerint. Ezek közül az első N darab o_kif szerinti adja az eddigi normál gyűjtősorokat, a többi pedig a szupergyűjtő-sorokat. Összesen 2N darab különböző csoportosítás van végrehajtva.
  - SELECT t_kod, beosztas, AVG(fizetes), COUNT(*)
    FROM alkalmazott
    GROUP BY CUBE(t_kod, beosztas);

SQL bővítések: Group by kiegészítői: ROLLUP, CUBE operátorok

- Mindkét esetben a szupergyűjtő-sorokban az "összesen oszlopérték" egy speciális NULL értékkel van reprezentálva, melyet egy speciális függvénnyel tudunk kezelni:
- GROUPING(o_kif) = 1 ha o_kif "összesen oszlopértéket" reprezentáló speciális NULL érték,
    - = 0 egyéb érték és a normál NULL esetén.

```
SELECT premium, GROUPING(premium),
              AVG(premium), COUNT(*)
FROM alkalmazott
GROUP BY ROLLUP(premium);
```

- /*Kiegészíthető: */ HAVING GROUPING(premium)=1; -- csak a szupersorokat adja vissza
- /*vagy: */ HAVING premium IS NULL; -- szupersorok és a premium IS NULL sorok

SQL bővítések: Group by kiegészítői: ROLLUP, CUBE operátorok

- SELECT DECODE(GROUPING(t_kod), 1, 'Össz t_kód', t_kod) AS t_kód,
  DECODE(GROUPING(beosztas), 1, 'Össz beosztás', beosztas) AS beosztás,
  COUNT(*) "Alk. szám",
  AVG(fizetes) * 12 "Átlag fiz"
  FROM alkalmazott
  GROUP BY CUBE (t_kod, beosztas);

/*Kiegészíthető: */ HAVING GROUPING(t_kod)=1 OR GROUPING(beosztas)=1;

- T_KÓD BEOSZTÁS Alk. szám Átlag fiz
- ---------- --------------- ---------- ----------
- 10 IGAZGATO 1 648000
- **10 Össz beosztás 1 648000**
- 20 ELADO 1 156000
- 20 TELEPHELYVEZETO 1 426000
- 20 VIZSGABIZTOS 1 240000
- **20 Össz beosztás 3 274000**
- 30 ELADO 1 159000
- 30 SZERELO 1 216000
- 30 TELEPHELYVEZETO 1 348000
- **30 Össz beosztás 3 241000**
- 40 ELADO 2 141000
- 40 TELEPHELYVEZETO 1 450000
- 40 VIZSGABIZTOS 1 252000
- **40 Össz beosztás 4 246000**
- 50 ELADO 1 156000
- 50 SZERELO 1 264000
- 50 TELEPHELYVEZETO 1 390000
- **50 Össz beosztás 3 270000**
- 60 SZERELO 1 252000
- 60 TELEPHELYVEZETO 1 300000
- **60 Össz beosztás 2 276000**
- Össz t_kód ELADO 5 150600
- Össz t_kód IGAZGATO 1 648000
- Össz t_kód SZERELO 3 244000
- Össz t_kód TELEPHELYVEZETO 5 382800
- Össz t_kód VIZSGABIZTOS 2 246000
- **Össz t_kód Össz beosztás 16 283687.5**

OLAP támogatás 2

- SQL bővítések
  - Group by kiegészítői: ROLLUP, CUBE operátorok

```
select channel_desc,calendar_month_desc, country_id,
        to_char(sum(amount_sold), '9,999,999,999') SALES$
from sales, customers, times, channels
where
    sales.time_id=times.time_id and
  sales.cust_id=customers.cust_i and
  sales.channel_id= channels.channel_id and
   channels.channel_desc IN ('Direct Sales', 'Internet') and
    times.calendar_month_desc IN ('2002-09', '2002-10') and
   country_id IN ('CA', 'US')
group by cube(channel_desc,calendar_month_desc,country_id);
```

CHANNEL_DESC	CALENDAR	CO	SALES$	
Direct Sales	2002-09	CA	1,378,126	
Direct Sales	2002-09	US	2,835,557	
Direct Sales	**2002-09**		**4,213,683**	**BY Channel and Month**
Direct Sales	2002-10	CA	1,388,051	
Direct Sales	2002-10	US	2,908,706	
Direct Sales	**2002-10**		**4,296,757**	**BY Channel and Month**
Direct Sales		CA	2,766,177	BY Channel and Country
Direct Sales		US	5,744,263	
Direct Sales			**8,510,440**	**BY Channel**
Internet	2002-09	CA	911,739	
Internet	2002-09	US	1,732,240	
Internet	**2002-09**		**2,643,979**	**BY Channel and Month**
Internet	2002-10	CA	876,571	
Internet	2002-10	US	1,893,753	
Internet	**2002-10**		**2,770,324**	**BY Channel and Month**
Internet		CA	1,788,310	BY Channel and Country
Internet		US	3,625,993	
Internet			**5,414,303**	**BY Channel**
	2002-09	CA	2,289,865	BY Month and Country
	2002-09	US	4,567,797	
	2002-09		**6,857,662**	**BY Month**
	2002-10	CA	2,264,622	
	2002-10	US	4,802,459	
	2002-10		**7,067,081**	
		CA	4,554,487	
		US	9,370,256	
			13,924,743	**Everything**