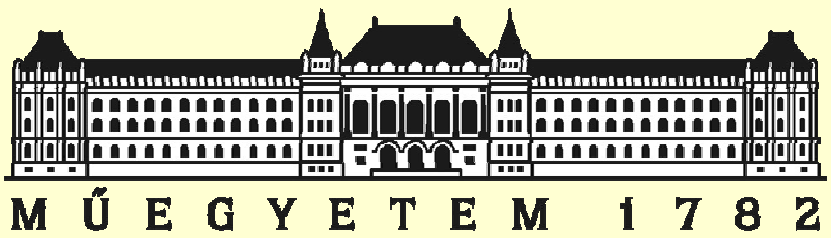
DW @ Adatb haladóknak

Adattárház alapú vezetői információs rendszerek

Gajdos Sándor
gajdos@tmit.bme.hu



Adatbázisok haladóknak 2012.

2012. szeptember 11.

DW @ Adatb haladóknak

Bevezető

- Fő kérdések:
  - Mitől lesz egy adattárház projekt sikeres?
  - Mik a legfontosabb általános tudnivalók egy ilyen rendszerről?
- Szemléletváltás, szemléletformálás kritikus
  - Más módszertan
  - Más eredménytermékek
  - Más technológia
- Nincs kitaposott út

DW @ Adatb haladóknak

Tartalom

- Adattárház fogalmak, elhelyezés
- Egy kis adattárház történelem
- Érettség megítélése
- Projekt előkészítése
- Adattárház architektúrák
- Dimenziós modellezés
- Az ETL egyes fontosabb kérdései

Alapfogalmak I.

Adattárház (Inmon definíciója)

Data Warehouse Definition

A Data Warehouse is a subject-oriented, integrated, time-variant, nonvolatile collection of data in support of management decisions.

- **Subject-oriented:** data that has some commonality from a business perspective, not silos of data based on how they are arranged from a systems perspective.
- **Integrated:** Provide consistent coding and formats.
- **Time-variant:** Data is organized by time and is stored in any number of ways to support historical reporting.
- **Nonvolatile:** No updates are allowed. Only load (append) and retrieval (query) operations is allowed.

Inmon, W. H., Building the Data Warehouse, QED/Wiley, 1991.

Alapfogalmak I.

Adattárház, mint döntéstámogató eszköz helye

- Döntéstámogató rendszerek
  - Kommunikáció-orientált
  - **Adat-orientált**
  - Dokumentáció-orientált
  - Tudás-orientált
  - Modell-orientált

Alapfogalmak II.

Üzleti intelligencia (BI) (EPICOR, 2005):

„Az olyan tudás, amikor már a megtörténés pillanatában tudjuk, hogy mi folyik egy adott területen, birtokunkban vannak a tények a folyamatok megértéséhez, és megvan a képességünk, hogy gyorsan tegyünk is valamit, ha szükséges.”

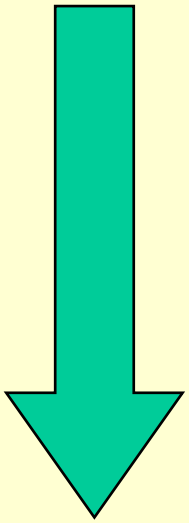
„The art of science of knowing what the heck is going on with your business as it is happening, having the **facts** to **understand** it and **support** it, and having the ability to **quickly do something** about it.”

Alapfogalmak II.

Üzleti intelligencia II.

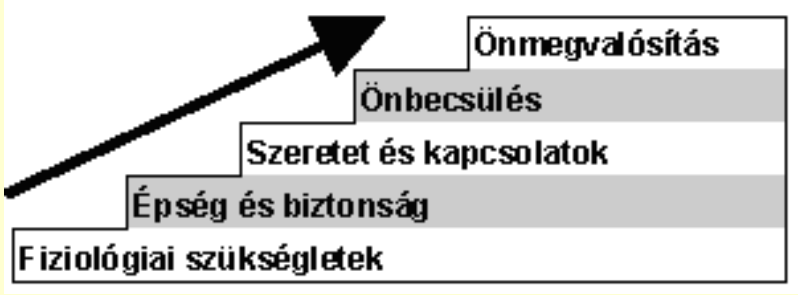
- A gyakorlatban:
  - Konzerv riportok
  - Rendszeres riportok
  - OLAP elemzések
  - Ad-hoc lekérdezések
  - Adatbányászat

A felhasználószám csökken



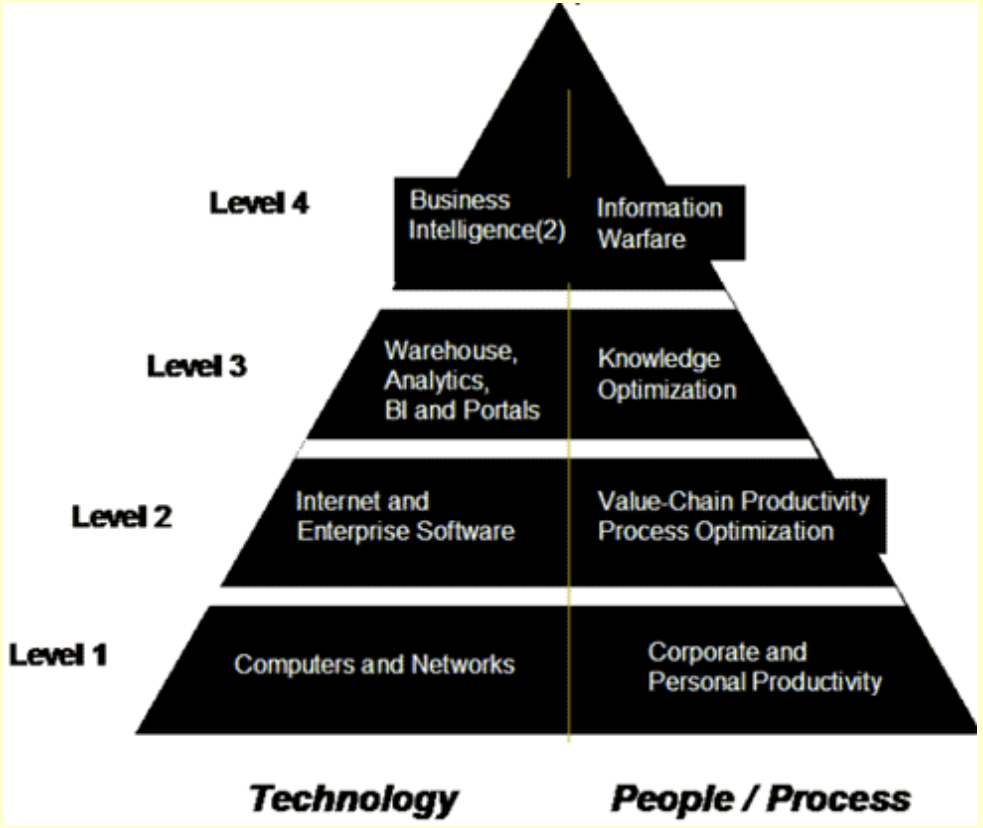
A szükségletek hierarchiája (Maslow)

Avagy: mi „működteti” az embereket



A vállalatok rengeteg energiát ölnek abba, hogy fokozzák alkalmazottaik lelkesedését. Ez igazán szép tőlük, de nézzünk szembe a tényekkel - dolgozni nem jó. Ha az emberek annyira szeretnének dolgozni, ingyen is csinálnák. Azért kell megfizetni az emberek munkáját, mert a munka messze nem tartozik az elképzelhető legkellemesebb időtöltések közé. Az ésszerű vállalat tudja, hogy az alkalmazottak akkor lelkesednek a legjobban a munkájukért, ha segítünk nekik, hogy minél hamarabb abbahagyhassák azt.
Scott Adams: Dilbert elv. SHL Hungary Kft. 2001

És mi „működteti” a vállalatokat?



Alapfogalmak III. – OLTP vs. analitikus rendszerek

- Eddig: az adatbázis a világ egy hű leképezése, amely az aktuális állapotot tükrözi → gyakori update
- Most: mikor mi volt a rendszer állapota? Milyen változások jellemezték? → lekérdezésekre használják, a trendekre akarunk következtetni.
- „Bármit” le lehet kérdezni, de mit érdemes?
- Operatív dolgozók lekérdezései ↔ vezetők lekérdezései

Alapfogalmak III. – OLTP vs. analitikus rendszerek II.

- Jellegzetes a dimenziós modellezés - a lekérdezés orientált megközelítés klasszikus módszere
  - Tudni kell, hogy mit elemzek, minek a függvényében
    - Tények
    - Dimenziók
  - Konform dimenziók fogalma, adattárház busz szerepe és jelentősége
- Vallásháborúk az alkalmazás módjáról
  - A felhasználói réteg szinte mindig dimenziós

Alapfogalmak III. – OLTP vs. analitikus rendszerek III.

- Aggregátumok
  - Előre kiszámított, majd eltárolt lekérdezés eredménye, amely a tényadatokat összegzi a dimenziókban található hierarchiáknak megfelelően.
  - Az adattárház teljesítménye kézben tartásának egyik leghatékonyabb eszköze
  - Lehetséges aggregátumok száma általában igen nagy → aggregátum menedzsment

Alapfogalmak IV. Metaadatok

- „Adat az adatról”
- Alapvetően befolyásolja a használhatóságot mind lekérdezhetőség, mind üzemeltethetőség szempontjából
- Fajtái:
  - Technikai (pl. sémaadatok)
  - Szemantikai (üzleti jelentések)
  - Adminisztrációs (pl. betöltések adatai)
  - Navigációs (pl. aggregátumdefiníciók)

Kis adattárház/döntéstámogatás történelem I.

- `60: batch riportok, papírkötegek
  - nehéz a megfelelő adatot megtalálni egy nagy halom papírban, lassú a kérés kiszolgálása, nehéz az adatok további feldolgozása
- `70: terminál alapú rendszerek
  - nehézkes felhasználói lekérdezések, gyenge felhasználói interfész, gyenge integráció az irodai szoftverekkel, adatforrások gyenge integrációja
- `80: PC alapú hozzáférés és EIS (Executive Inf. Systems)
  - grafikus interfészek, de zavaros adatok, inkonzisztens kódok, bonyolult adatbázis sémák, kevés historikus adat
- `90: adattárházak
  - megoldja a korábban nyitva maradt problémák legtöbbjét. Kezdetben kliens szerver megoldások, desktop OLAP, később web technológiák.
- `00: valós idejű rendszerek
  - Minimális késleltetés, stratégiai mellett operatív döntéstámogatás is.
- `10: mobil hozzáférés
  - Hozzáférés bárhonnan, bármikor, akár a legfrissebb adatokhoz/elemzésekhez

Kis adattárház történelem II.

Megjelenés tipikus sorrendje:

1. Kereskedelem
2. Távközés
3. Pénzügyi szektor
4. Államigazgatás, költségvetési intézmények

Érettség megítélése

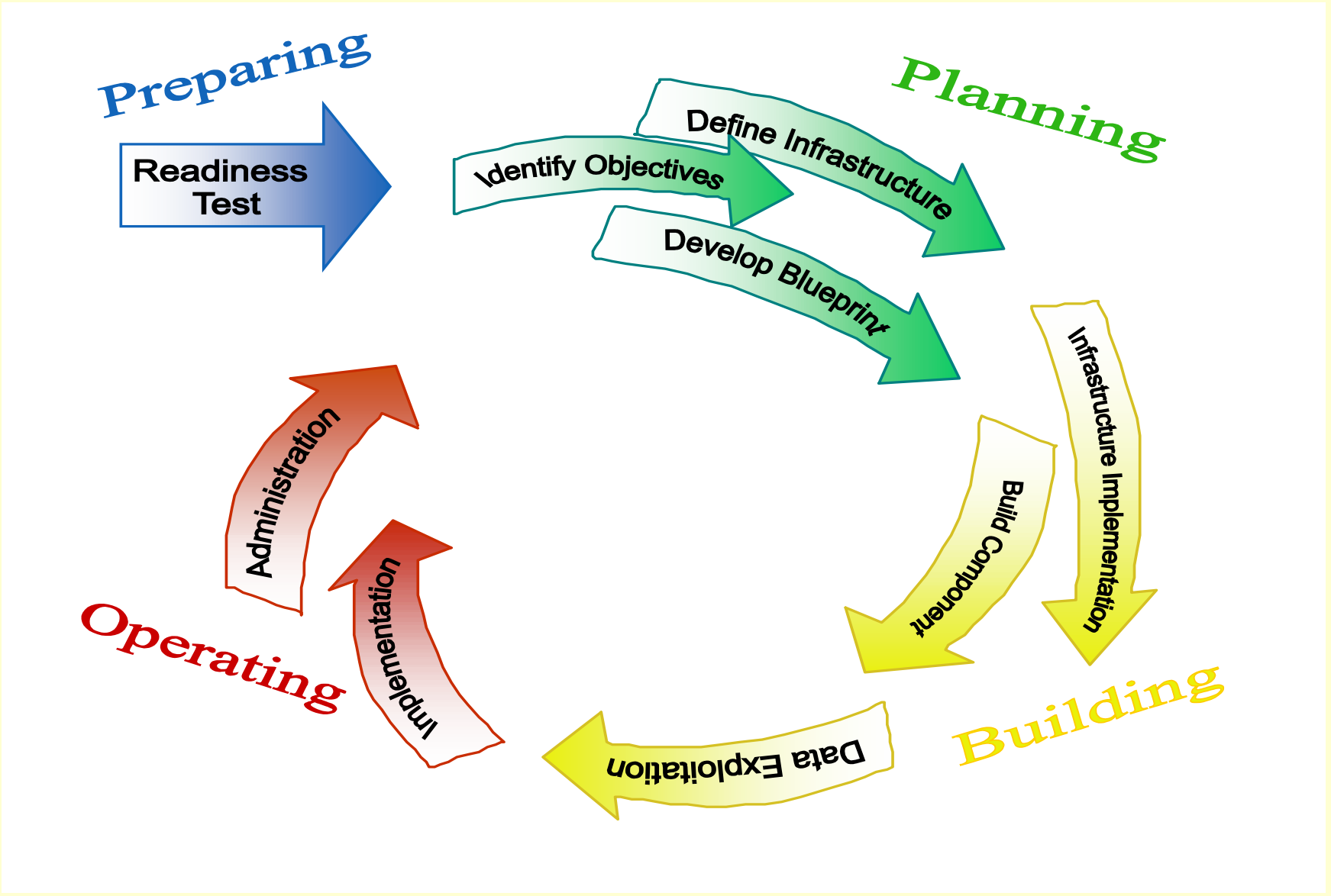
Viszonylag magas a sikertelen projektek aránya

- Érettség objektív megítélése fontos
- Lakmusz teszt (több száz projekt tanulsága alapján, forrás: Kimball)
  - Erős üzleti szponzor: 60 %
  - Valós üzleti motiváció: 15%
  - Technológia - üzleti vezetés partnersége: 5%
  - Meglévő analitikus kultúra: 5%
  - Különböző megvalósíthatósági szempontok: 15%
- Nem mindenható, de...

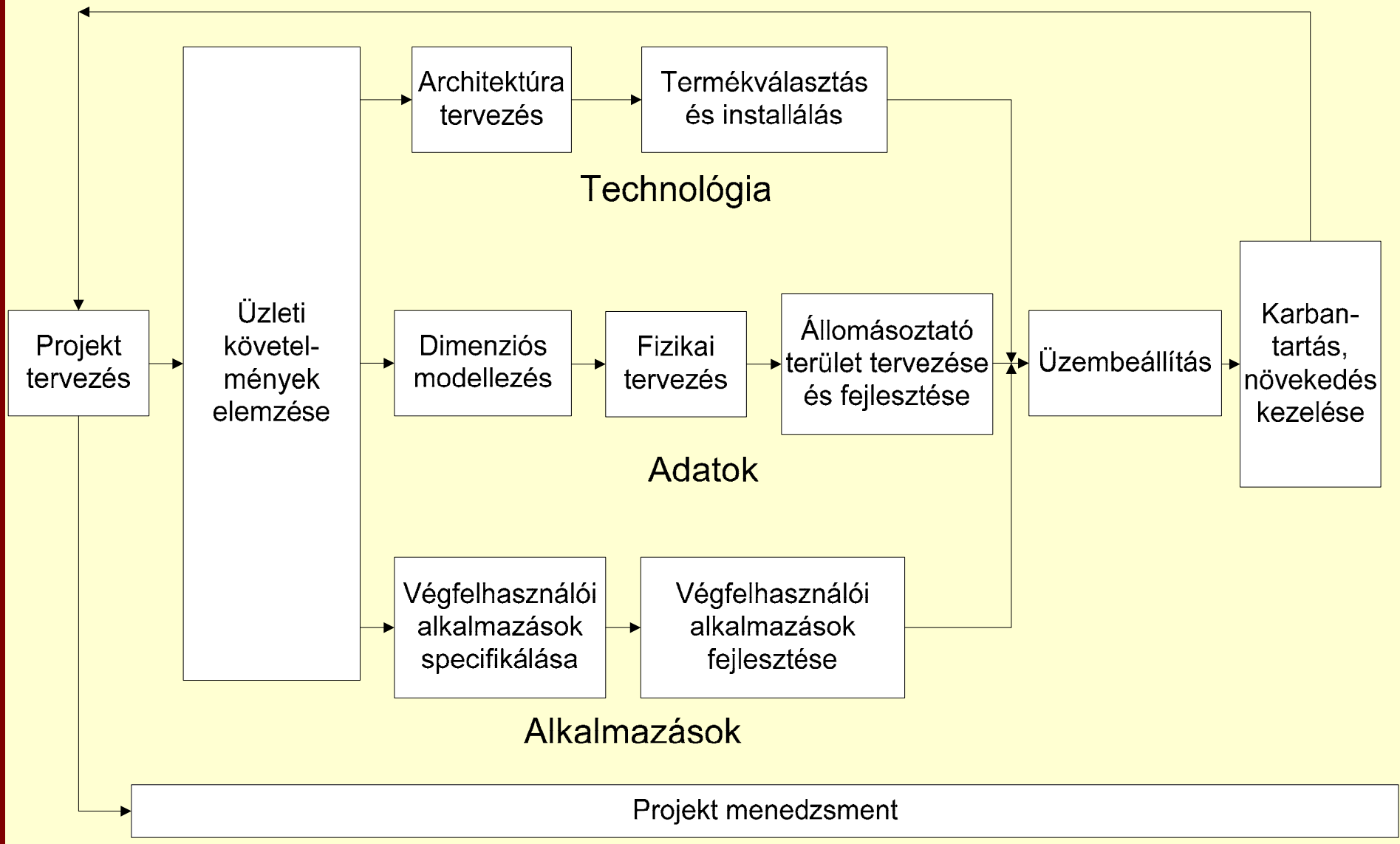
1. Adattárház megvalósítási módszertanok I.

- Hadden-Kelly
- HP Open Warehouse
- Oracle Warehouse Methodology
- Ralph Kimball-é
- SAS
- …

Hadden-Kelly



Ralph Kimball módszertana



1. Adattárház megvalósítási módszertanok II.

Konklúzió: **általános jellemző a fázisok definiálása és az iteratív szemlélet**

Előnyei:

1. Megvalósítási kockázat csökkentése
2. Megvalósult részek korábbi kiaknázása
3. Humán erőforrások egyenletesebb terhelése
4. Hasznosulás mértékének növekedése a fokozatos rendszerbeállítással
5. Lehetőség a megrendelői oldal fokozatos bevonására a megvalósításba

Hátránya:

Valamivel nagyobb átfutási idő, ill. ráfordításigény

2. Kezdeti munkamegosztás a megrendelő és szállító szervezet között

Ha a megrendelői oldal vállalja az...

- üzleti igények részletes specifikálását:
  - – élőmunka-igényes, spec. erőforrásokat igényel, de
  - \+ csökkenti mindkét oldal kockázatát, így a vállalási árat.

- adatforrások felmérését, adatok minőségbiztosítását:
  - – spec. erőforrásokat igényel, jelentős ráfordításigényű
  - \+ csökkenti a megrendelői oldal kockázatát

Projekt előkészítése II.

Konklúzió: minél pontosabban érdemes tudni, hogy

- **Mit** (üzleti igények meghatározása)
- **Miből** (forrásadatok felmérése, minőségbiztosítása)
- (**Hogyan** (technológiai elvárások)).

Ezek eldöntése is rábízható a szállítóra, de ez

- növeli a projekt kockázatát (idő, szkóp bizonytalanság)
- a szállítói kockázattal együtt növeli a költségeket.

Projekt előkészítése III.

Milyen követelményeket fogalmazzunk meg az adattárházzal kapcsolatban?

- Teljesítmény
- Meghatározott elemző eszközök támogatása
- Skálázhatóság
- Adatbiztonság
- Működés biztonság
- Egyszerű karbantartás (ld. fenntartási költségek)

DW architektúrák

„Rendszertervezési döntés, amely általában nem könnyen változtatható meg”

„Fontosabb szempontok, amiket figyelembe kell venni.”

- Mire jók a különböző architektúrák?
  - Kommunikáció
  - Tervezés
  - Tanulás
  - Hatékonyságnövelés és újrahasznosítás

Architektúrák

- Koncepcionális architektúra
- Adat(konzisztencia) architektúra
- Front-end architektúra és back-end architektúra
- Eszközarchitektúra (HW, SW)
- Üzemeltetési architektúra
- Biztonsági architektúra
- ...

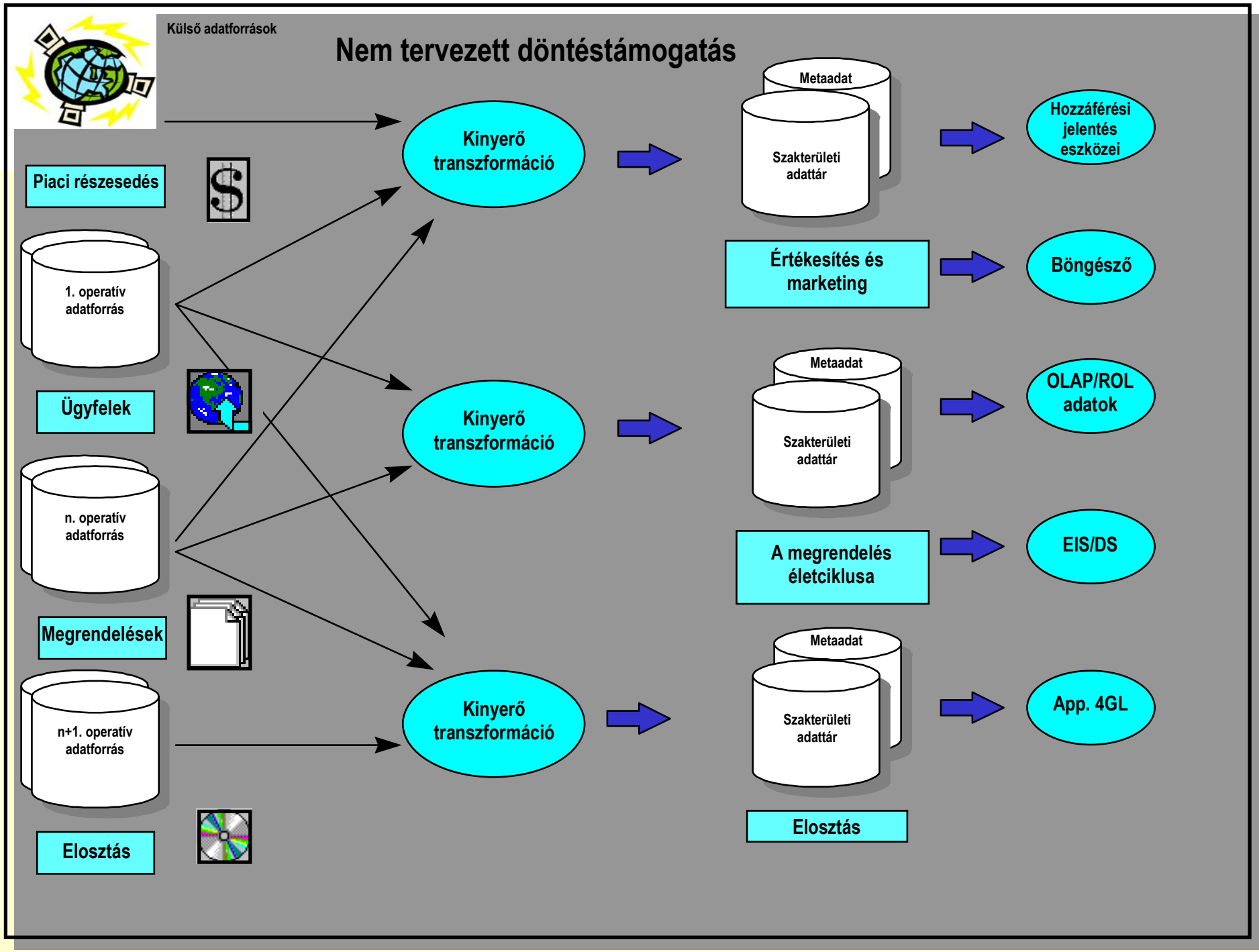
Koncepcionális architektúra főbb elemei

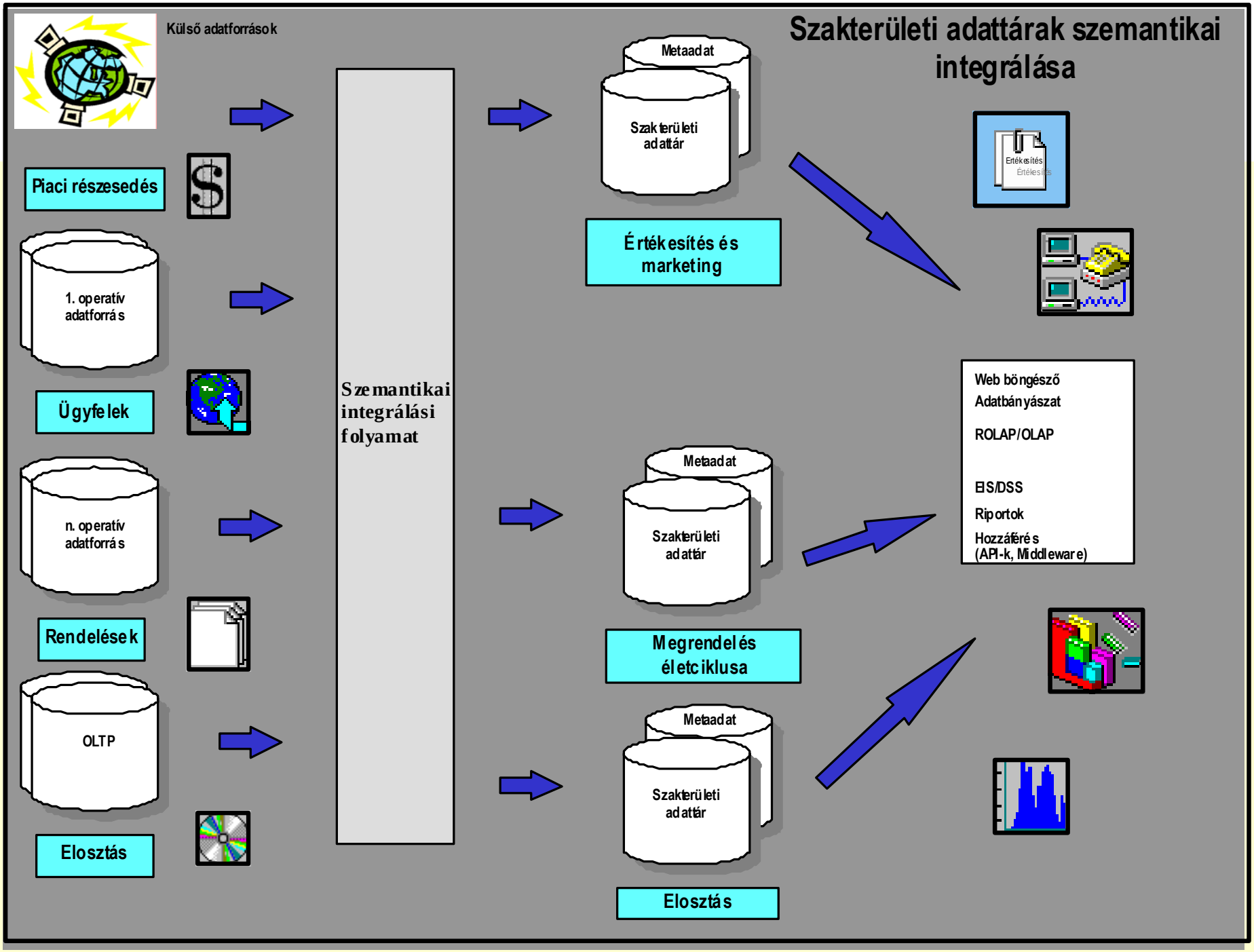
- forrásrendszerek
- adatkinyerés-integrálás
- állomásoztató terület (staging area, SA)
- elemi adattár (detailed storage, DS)
- szakterületi adattár (data mart)
- metaadattár
- (üzemi adattár (operational data store, ODS))
- megjelenítés támogatás

Koncepcionális architektúra főbb fejlődési fázisai

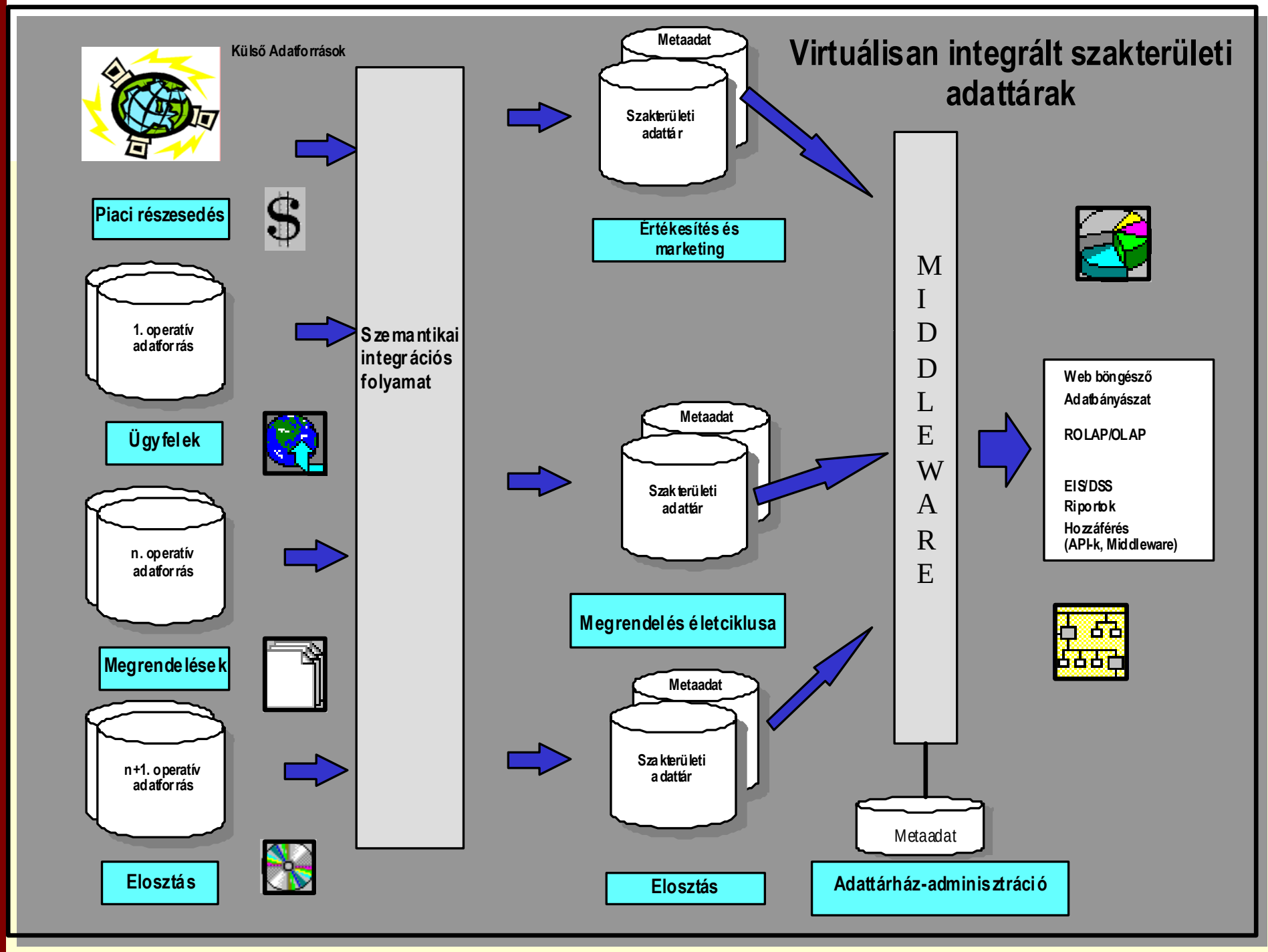
- Nem tervezett döntéstámogatás
- Szemantikailag integrált szakterületi adattárak
- Virtuálisan integált szakterületi adattárak
- Függő szakterületi adattárak (hub and spoke architektúra)

DW @ Adatb haladóknak



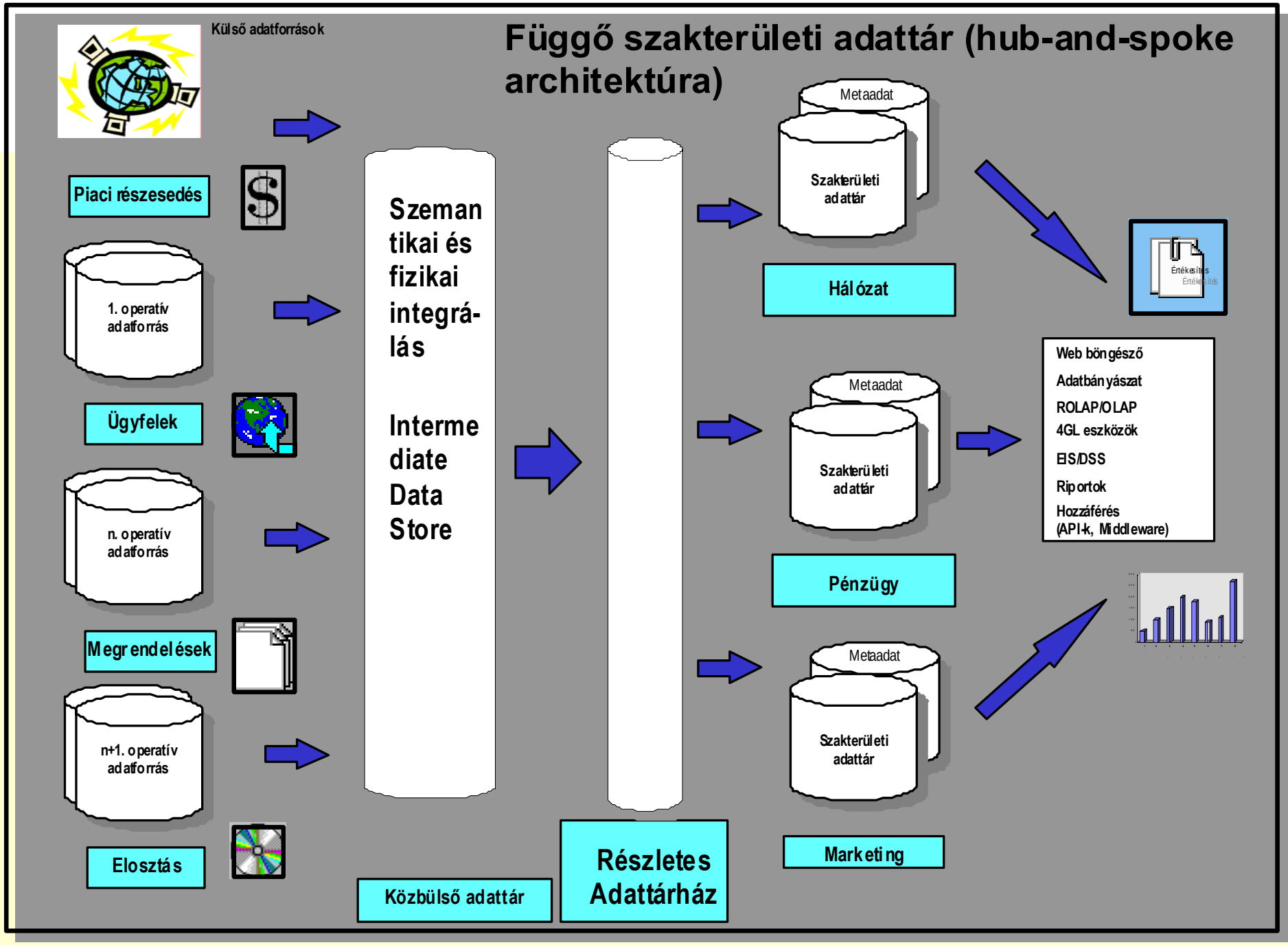

Szakterületi adattárak szemantikai integrálása
Külső adatforrások
Piaci részesedés
1. operatív adatforrás
Ügyfelek
n. operatív adatforrás
Rendelések
OLTP
Elosztás
Szemantikai integrálási folyamat
Metaadat
Szakterületi adattár
Értékesítés és marketing
Metaadat
Szakterületi adattár
Megrendelés életciklusa
Metaadat
Szakterületi adattár
Elosztás
Értékesítés
Web böngésző
Adatbányászat
ROLAP/OLAP
EIS/DSS
Riportok
Hozzáférés (API-k, Middleware)

DW @ Adatb haladóknak



DW @ Adatb haladóknak

Függő szakterületi adattár (hub-and-spoke architektúra)



Dimenziós modellezés

- **Dimenziós modellezés előnyei:**
  - lekérdezése könnyen optimalizálható
  - a modell bővítése viszonylag egyszerű
  - laikusok által is könnyen lekérdezhető

Négylépéses dimenziós modellezés

1. Üzleti folyamat azonosítása
2. Tényadat granularitásának megválasztása
3. Dimenziók azonosítása
4. Tények azonosítása

1. Üzleti folyamat izolálása

Példák:

- szolgáltatás használata,
- hitelek igénylése és felvétele,
- bevételek alakulása,
- kinnlevőségek,
- rendelések
- személyzeti ügyek
- számlázás
- javítások és reklamációk, stb.

2. Tényadat granularitásának megválasztása

- milyen részletes adatok tárolását támogatjuk
- túl részletes: sok adat, nagy diszkigény, nagy CPU igény (különösen a betöltések során)
- nem elég részletes: elemzéseket akadályozhat meg

3. Dimenziók azonosítása

- Mi alapján akarjuk rendezni, lekérdezni, csoportosítani a tényadatokat?
- Sok és részletes dimenzió - változatosabb analízisek
- Dimenziók azonosítása szigorúan az adatok használata (ld. üzleti igények) alapján
- Dimenzió lesz minden, ami...
- Inkább szöveges attribútumok, de lehet numerikus is

4. Tények azonosítása

- Az elemzendő mennyiségek konkrét meghatározása (pl. eladási ár Ft-ban, darabszám, átlagos kisker. ár, …)
- Általában folytonos értékkészletűek és numerikusak.

Dimenziós tervezési elvek

- Pontosan ismerni és érteni az adatokat
- Dimenziós táblák: leíró attribútumok, akár 50 is, a rekordok hossza kevéssé kritikus.
- Ténytáblák: a rekordok legyenek rövidek
- Konform dimenziókban gondolkodunk
- Minden dimenziónak legyen **surrogate** (anonym, kiegészítő, jelentés nélküli, mesterséges) kulcsa.

Surrogate kulcs

Előnyei:

- méretcsökkentés a ténytáblában
- forrásrendszeri kulcs változásaitól függetlenek leszünk
- az entitások időbeli változásait is le tudjuk így írni

Hátránya:

- újra kell kulcsolni a tény és dimenziós rekordokat (jelentős betöltési többletteher)

Dimenziós tábla tervezés

- A felesleges dimenziók teljesítményveszteséget eredményeznek.
- A dimenziós adatok nem feltétlenül nyerhetők ki valamely forrásrendszerből.
- Az idő, termék, hely, ügyfél a leggyakoribb dimenziók

Állapot- és esemény-tények

- Esemény-tény: egyetlen időponthoz kapcsolódik
- Állapot-tény: két időpont (kezdete, vége)
  - Új tényrekord beszúrása egy másik lezárásával jár → alacsonyabb hatékonyság
  - valószínűbb információvesztés (ld. később)
- Általában egymásba átalakíthatók
  - Kik, mikor, hol, mit, stb. vásároltak
  - Kik azok a vásárlók, akiknek van …
  - Melyek azok a termékek, amelyeket eladtak…
  - …
- A lekérdezések hatékonysága erősen különböző!

Ha a dimenzió is változik idővel… (“slowly changing dimensions”, SCD)

Pl. az valaki költözik, címe megváltozik

1. régi rekord felülírása
2. “old” mező képzése a dim. táblában
3. új rekord a dim. táblában a surrogate kulcs új értékével