Adatbázisok elmélete

2023.

Dr. Gajdos Sándor BME-TMIT

Információs rendszer tervezési kihívás

Rendszer specifikáció a SkyAlliance helyfoglaló rendszere számára

Feladat: interneten keresztül repülőjegyek foglalása az összes, a szövetséghez tartozó járatra, a foglalások módosítása, a foglalások, a járat foglaltsági adatok és az ügyfelek szokásai elemezhetőségének biztosítása

Információs rendszer tervezési kihívás

(Adjunk rá versenyképes ajánlatot!)

	Követelmények	
1	egyidejűleg átlagosan 5000, maximum 50000 felhasználó kiszolgálása	
2	háromrétegű architektúra (adatbázis szerver, alkalmazás szerver, web kliensek (desktop, mobil))	
3	biztonságos és titkosított kommunikáció a webes kliensekkel	
4	Adat nem veszhet el és nem válhat a cég számára hozzáférhetetlenné	
5	max. válaszidő: 1 sec	
6	rendelkezésreállás: 99,99%	
7	adatkörök: járatok, repülők, ülőhelyek, utasok, keresési adatok, session adatok	
8	nettó adatmennyiség/év: 50 GB, növekmény: 15%	
9	Adatmegőrzési idő: min. 10 év	

Információs rendszer tervezési kihívás

	Követelmények	Kapcsolódás a tárgytematikához
1	egyidejűleg átlagosan 5000, maximum 50000 felhasználó kiszolgálása	párhuzamos feldolgozás
2	háromrétegű architektúra (adatbázis szerver, alkalmazás szerver, web kliensek (desktop, mobil))	DB architektúrák
3	biztonságos és titkosított kommunikáció a webes kliensekkel	Ld. más tantárgyakban
4	Adat nem veszhet el és nem válhat a cég számára hozzáférhetetlenné	elosztott adatbázisok
5	max. válaszidő: 1 sec	párhuzamos feldolgozás, normalizálás, analitikus adatbáziskezelés teljesítménymérés és hangolás
6	rendelkezésreállás: 99,99%	DB architektúrák, elosztott adatbázisok
7	adatkörök: járatok, repülők, ülőhelyek, utasok, keresési adatok, session adatok	Ld. Adatbázisok...
8	nettó adatmennyiség/év: 50 GB, növekmény: 15%	Ld. Adatbázisok...
9	Adatmegőrzési idő: min. 10 év	Ld. Adatbázisok...

Relációs sématervezés OLTP rendszerek számára I.

- Ismétlés, szintrehozás
  - anomáliák,
  - redundancia,
  - implikáció,
  - funkcionális függések,
  - normál formák,
  - helyesség (igazság)
  - teljesség,
  - Armstrong axiómák

Relációs sématervezés OLTP rendszerek számára II.

- Sok mindent már tudunk...
- Mit szeretnénk elérni? Az anomáliák megszüntetésével
  1. DB műveletek hatékonyságának javítása, ill.
  2. Információvesztés elkerülése
     - Beszúrás
     - Törlés
     - Módosítás során

Relációs sématervezés OLTP rendszerek számára III.

- Megoldás: az univerzális/túl nagy/”nem jó” sémákat fel kell bontani. De hogyan?
- Cél: adott normál formákba (+egyéb szempontok…)
- Fel kell tudni ismerni őket (sémaanalízis)
  - Összes funkcionális függés (miért is?, milyen ff?)

Relációs sématervezés OLTP rendszerek számára IV.

- Az összes szemantikailag helyes/igaz érdemi függés kell
- Armstrong axiómák
- Jelentősége: ami helyes, azt le is tudjuk vezetni (teljességi tétel)
- El tudjuk dönteni (hatékonyan), hogy egy tetszőleges X→Y helyes/nem helyes?
    1. Axiómák
    2. függéshalmaz lezárása
    3. attribútumhalmaz lezárása

Relációs sématervezés OLTP rendszerek számára V.

- függéshalmaz lezárása
- attribútumhalmaz lezárása
- Kapcsolatuk és jelentősége
- Tanulság: ha $X \rightarrow Y$ helyessége a kérdés adott F mellett, akkor nézd meg, hogy $Y \in X^{+}(F)$ teljesül-e. Erre hatékony rekurzív algoritmus ismert.

Relációs sématervezés OLTP rendszerek számára VI.

- Függéshalmazok ekvivalenciája
- Minimális függéshalmaz, ha
  - a függőségek jobb oldalán egyetlen attribútum,
  - a függőségek bal oldaláról nem hagyható el attribútum,
  - nincs olyan függőség, amely elhagyható.
- Tétel: Adott függéshalmazzal ekvivalens minimális függéshalmaz mindig előállítható.
- Példa: $F=\{AB\rightarrow C, BC\rightarrow A, A\rightarrow BC\}$ $F_{min}=?$

Relációs sématervezés OLTP rendszerek számára VII.

Sémafelbontások – miért kell?

- Definíció
- Case study

Veszteségmentes sémafelbontások

- Veszteségmentesség definíció
- Project-join mapping és tulajdonságai
- Veszteségmentesség eldöntése...???
- Tétel1 (két részsémára)
- Tétel2 (táblázatos módszer, akárhány részsémára)

Relációs sématervezés OLTP rendszerek számára VIII.

- Veszteségmentes sémafelbontás előállítása két részsémára
- Tétel1 (két részsémára)
  - Bizonyítás
- Tétel2 (táblázatos módszer, akárhány részsémára)
  - Bizonyítás (ha v. mentes, akkor van csupa ‚a' sor)

Relációs sématervezés OLTP rendszerek számára IX.

- Függőségőrző sémafelbontások
- Vetített funkcionális függőségek
- Függőségőrző definíció
- Elvárások:
  - redundancia mentes,
  - veszteségmentes,
  - függőségőrző

Relációs sématervezés OLTP rendszerek számára X.

- Függőségőrző sémafelbotás 3NF részsémákba
  - Példa: R(TEIHDO), F={T->E, IH->T, IE->H, TD->O, ID->H}
  - A sémafelbontás: {TE, IHT, IEH, TDO, IDH}
- Függőségőrző és veszteségmentes sémafelbontás 3NF részsémákba
  - Példa, bizonyítás
- Veszteségmentes sémafelbontás BCNF részsémákba
- A veszteségmentes és függőségőrző sémafelbotás BCNF részsémákba nem garantálható.
  - Biz: konstruktív
  - BCNF tulajdonság tesztelése...

Relációs sématervezés OLTP rendszerek számára XI.

- Függőségőrző sémafelbotás 3NF részsémákba
  - Példa: R(TEIHDO), F={T->E, IH->T, IE->H, TD->O, ID->H}
  - A sémafelbontás: {TE, IHT, IEH, TDO, IDH}

T: tantárgy

E: előadó

I: időpont

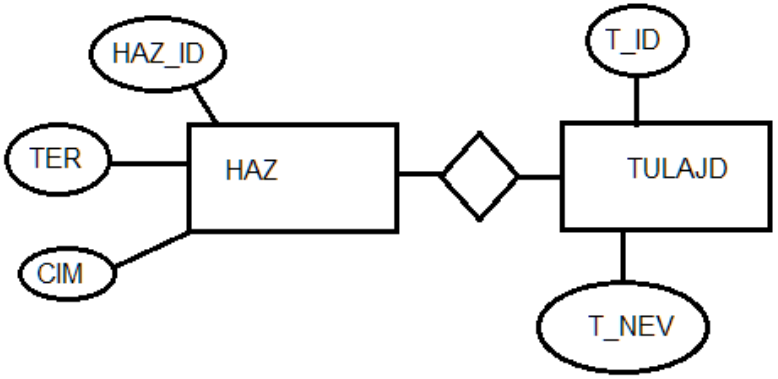
H: hely/terem

D: diák

O: osztályzat

Relációs sématervezés OLTP rendszerek számára XII.

Esettanulmány:



F={HAZ_ID->TER, CIM;
T_ID->T_NEV}

F_{MIN}={HAZ_ID->TER;
HAZ_ID-> CIM;
T_ID->T_NEV}

R_{univ}(HAZ_ID, T_ID, TER, CIM, T_NEV)
Ennek veszteségmentes, függőségőrző felbontása 3NF részsémákba:
{R_1(HAZ_ID, TER), R_2(HAZ_ID, CIM), R_3(T_ID, T_NEV), R_k(HAZ_ID, T_ID)}
---------- F_{min}-ből ---------- R_{univ} kulcsából

Relációs sématervezés OLTP rendszerek számára XIII.

Sémaösszevonás után:

{R_H(HAZ_ID, TER, CIM),
Ez megfelel a
HAZ táblának

R_T(T_ID, T_NEV),
Ez megfelel a TULAJ
táblának

R_k(HAZ_ID, T_ID)}
Ez megfelel a
„kapcsoló”táblának

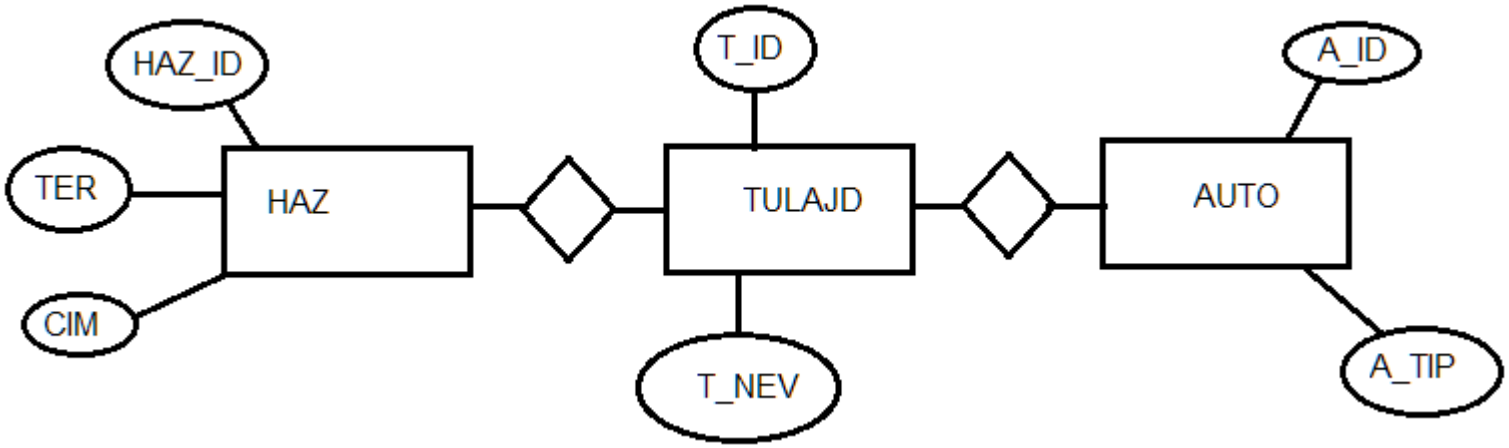
Relációs sématervezés OLTP rendszerek számára XIV.

Tehát:

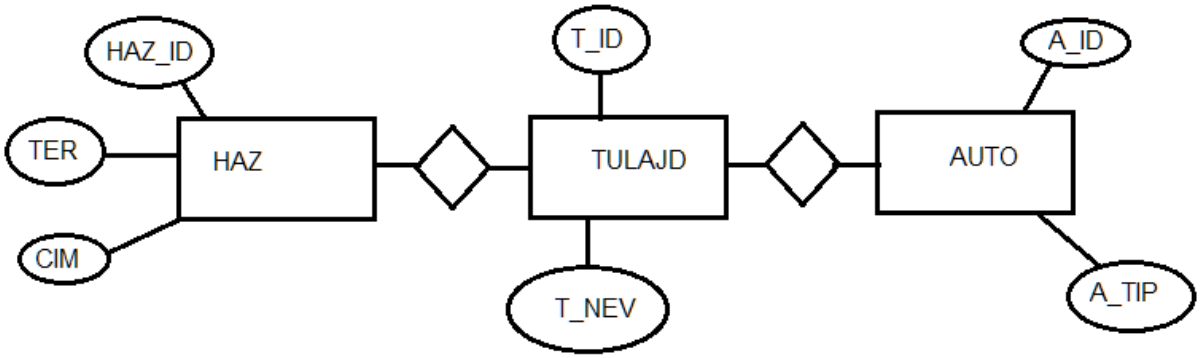
ER-ből: (HAZ_ID, TER, CIM),(T_ID, T_NEV),(HAZ_ID, T_ID)

FF-ből: (HAZ_ID, TER, CIM),(T_ID, T_NEV),(HAZ_ID, T_ID), ami (most) ugyanaz, de tudható, hogy legalább 3NF részsémákat tartalmazó v. mentes, függőségőrző felbontás.

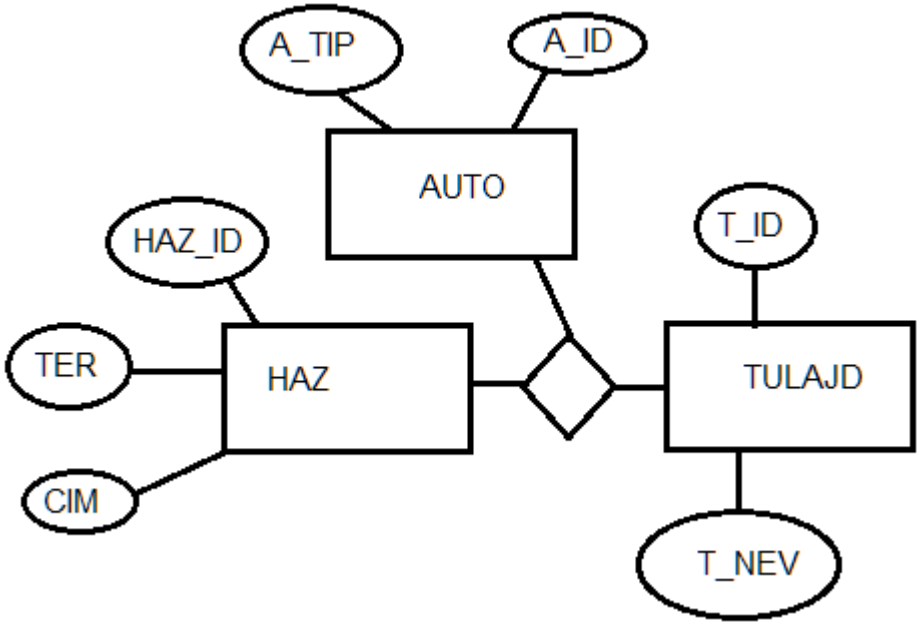
Viszont: (esettanulmány2):



Relációs sématervezés OLTP rendszerek számára XV.



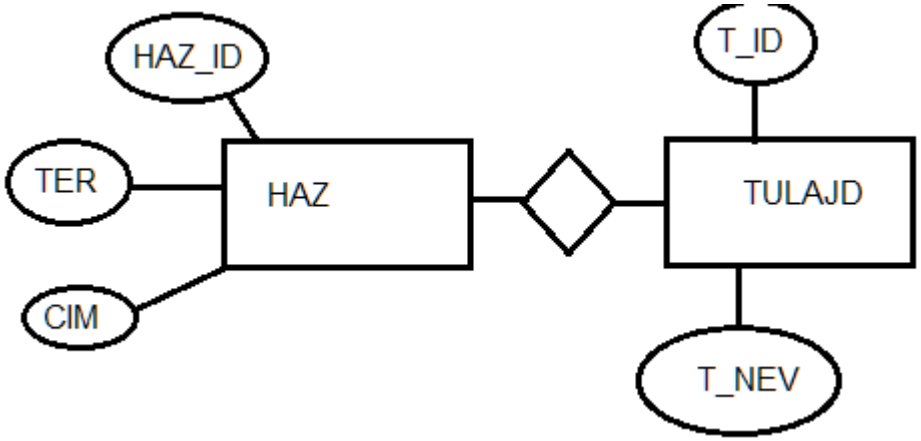
A funkcionális függésekből kapható sémák valójában ennek fognak megfelelni:



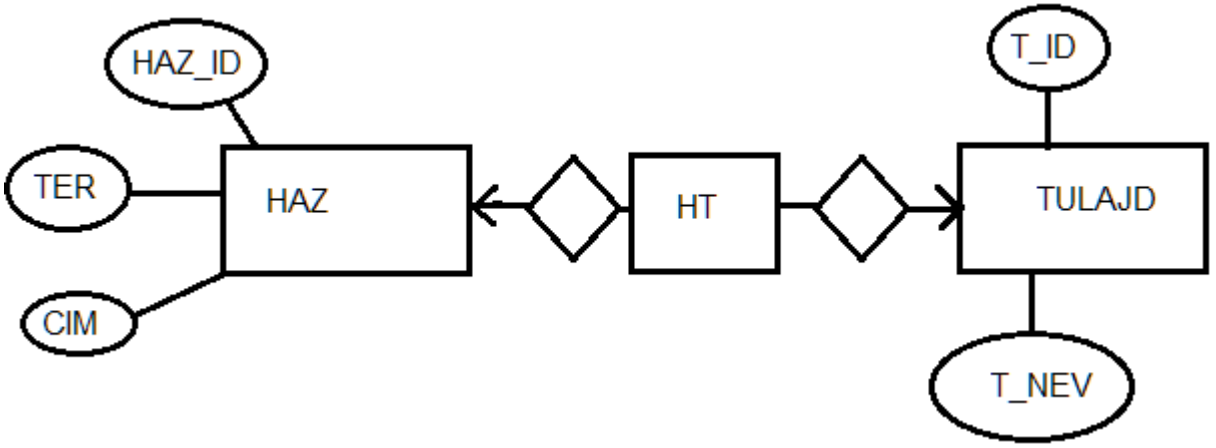
Tanulság: A funkcionális (“függvényszerű”) függőségek csak függvényszerű viszonyokat tudnak leírni. ☺

Relációs sématervezés OLTP rendszerek számára XVI.

Tehát



helyett így



gondolkozva f. függésekkel is „helyes” modellt kaphatunk.

Adatbáziskezelés analitikus környezetben

Szóhasználat és kontextus

- Analitikus, lekérdezésorientált, OLAP, adattárház, DSS...
- Kontextus: döntéstámogatás

Tartalom

- Döntéstámogatás általában
- OSS vs. DSS
- Multidimenziós modellezés
- Hozzáférési módok, BI eszközök
- Lekérdezés optimalizálás dim. struktúrákon
- Adattárház architektúrák
- Megvalósítási módszertanok
- Tervezési kérdések
- Implementációs kérdések
- Dimenziós modellezési gyakorlat

Döntéstámogatás

- Jelentősége...
- Hol van/honnan szerezhető meg a releváns info?
  - Kommunikáció-orientált
  - Adat-orientált
  - Dokumentáció-orientált
  - Tudás-orientált
  - Modell-orientált

Döntéstámogatás II.

- Kommunikáció-orientált
  - Kommunikáció, együttműködés, megosztott döntéstámogatás
  - Hirdetőtábla, lev. lista
  - Telefon(konferencia), doku megosztás
- Adat-orientált
  - (sok, idősoros) adathoz való hozzáférés
  - EIS/VIR, GIS, DW, OLAP,

Döntéstámogatás III.

- Dokumentáció-orientált
  - Strukturálatlan dokuk garmadája (audio, video is)
  - „Information retrieval”
  - AI/MI
  - Fuzzy módszerek,...
- Tudás-orientált („szakértő rendszerek”, intelligens DSS)
  - Szűk szakterület tudásanyaga
  - Spec. probléma megoldásának képessége

Döntéstámogatás IV.

- Modell-orientált („computation-oriented DSS”)
  - matematikai/formális modellezés alkalmazása
  - Tip: statisztikai, pénzügyi, optimalizálási, szimulációs
  - What if?
  - Általában nem adat-intenzív
- Döntéstámogatás a gyakorlatban...

Adat-orientált DSS története

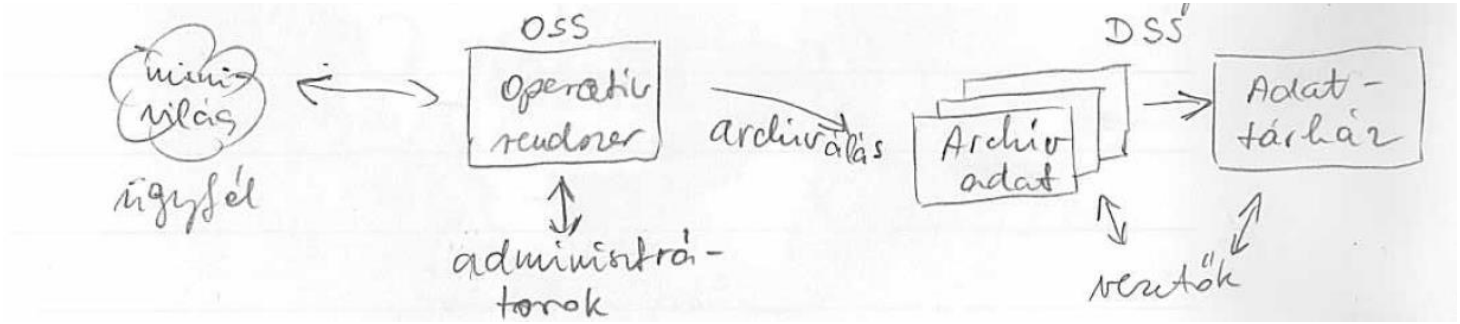
- 60-as évek: batch riportok, nyomtatva,
- 70-es évek: terminál alapú (nehézkes lekérdezések, gyenge UI, gyenge forrásintegráció)
- 80-as évek: PC alapú hozzáférés, GUI, inkonzisztens adatok, kevés adat,
- 90-es évek: adattárházak (korábbi problémák megoldása, desktop OLAP, trendanalízis)
- 95-től: webes elérhetőség
- 2005- valós idejű
- 2010- mobil

Lekérdezések támogatása I.

- Támogass „mindent”
  - Hardver támogatással (érdemben nem foglalkozunk vele)
    - Brute force, MPP,...
- Támogass kiválasztott lekérdezéseket
  - NoSQL/Big Data technológiák (ld. később)
  - Hagyományos technológia, dimenziós adatstruktúrák (most)

Lekérdezések támogatása II.

Hogyan????



A DSS rendszerek fizikailag elkülönülnek, mert...

Adatok módosítása helyett elemzés

Cél: a trendek felderítése

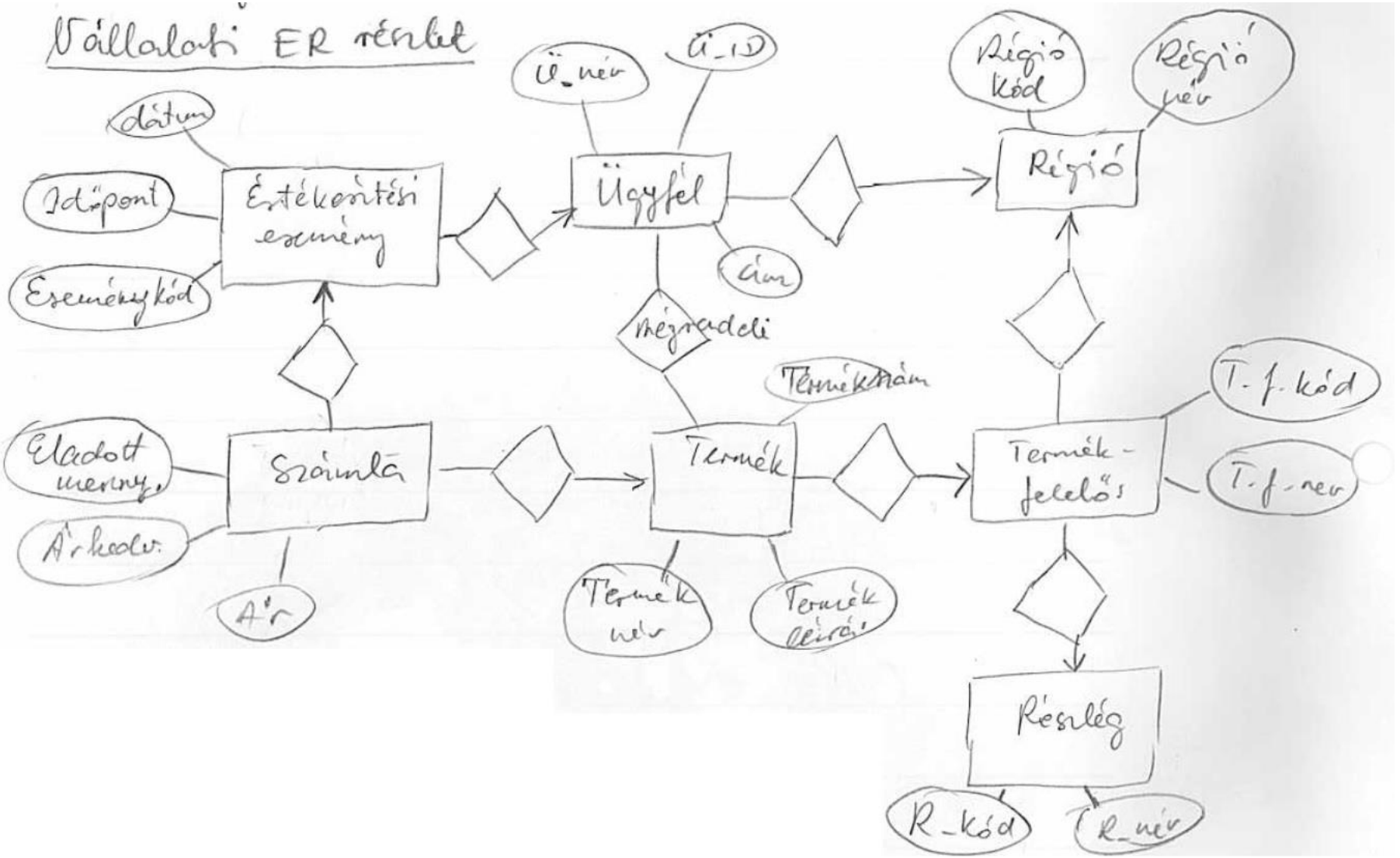
A pontos előrejelzésekhez sok adat kell

Időbeli fókuszú tárolás

Összehasonlítás

	OLTP/OSS	OLAP/DSS
alapfunkció	adatfeldolgozás	döntéstámogatás
eredménytermék	funkcionalitás	információ
válaszidő	mp	akár órák, napok
érintett rekordszám műveletenként	néhány	igen sok
struktúra	sok tábla, kevés oszlop, magasan normalizált	kevés tábla, sok oszlop, redundáns
adatok jellege	dinamikus	statikus
frissítés	folyamatos	periodikus változások, bulk loading
felhasználószám	nagy	kevesebb
gép terhelése	stabil, jól méretezhető	dinamikus, rosszul méretezhető
rendelkezésreállás	kritikus	átlagos

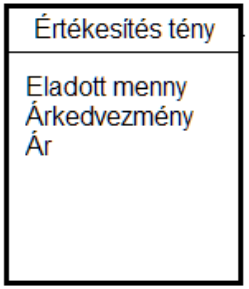
Lekérdezések támogatása - példa



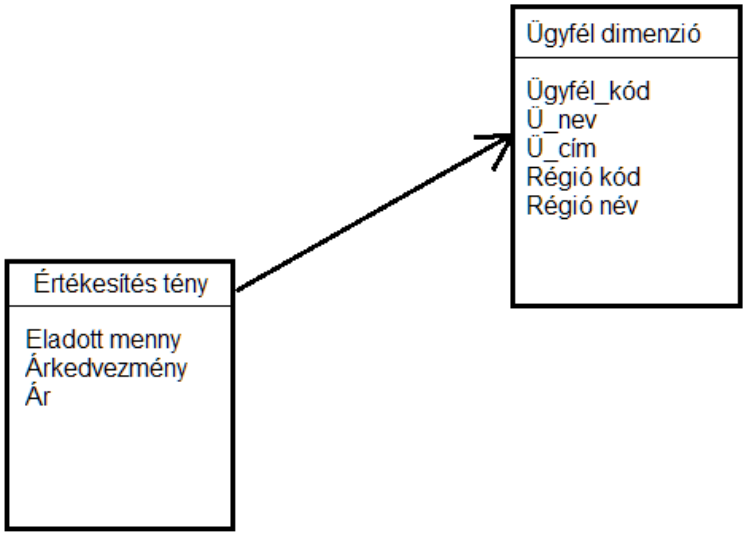
Lekérdezések támogatása III.

- Multidimenziós logikai adatstruktúra
  - Tényadatok: a dim/csillagstruktúra közepe. Numerikus, folyamatos értékkészlet, kevés attribútum, sok rekord
  - Dimenziós adatok: a dim/csillagstruktúra „ágai”. Amik mentén a tényadatokat jellemezzük vagy változásait figyelemmel kísérjük. Sok, leíró jellegű attribútum.
- Többváltozós függvény analógia

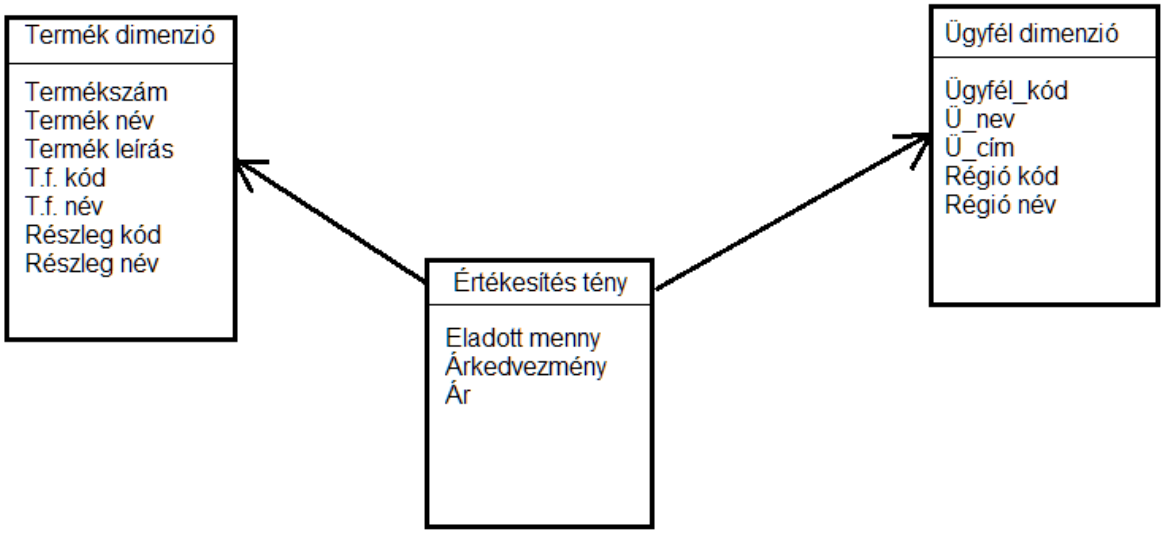
Lekérdezések támogatása - példa



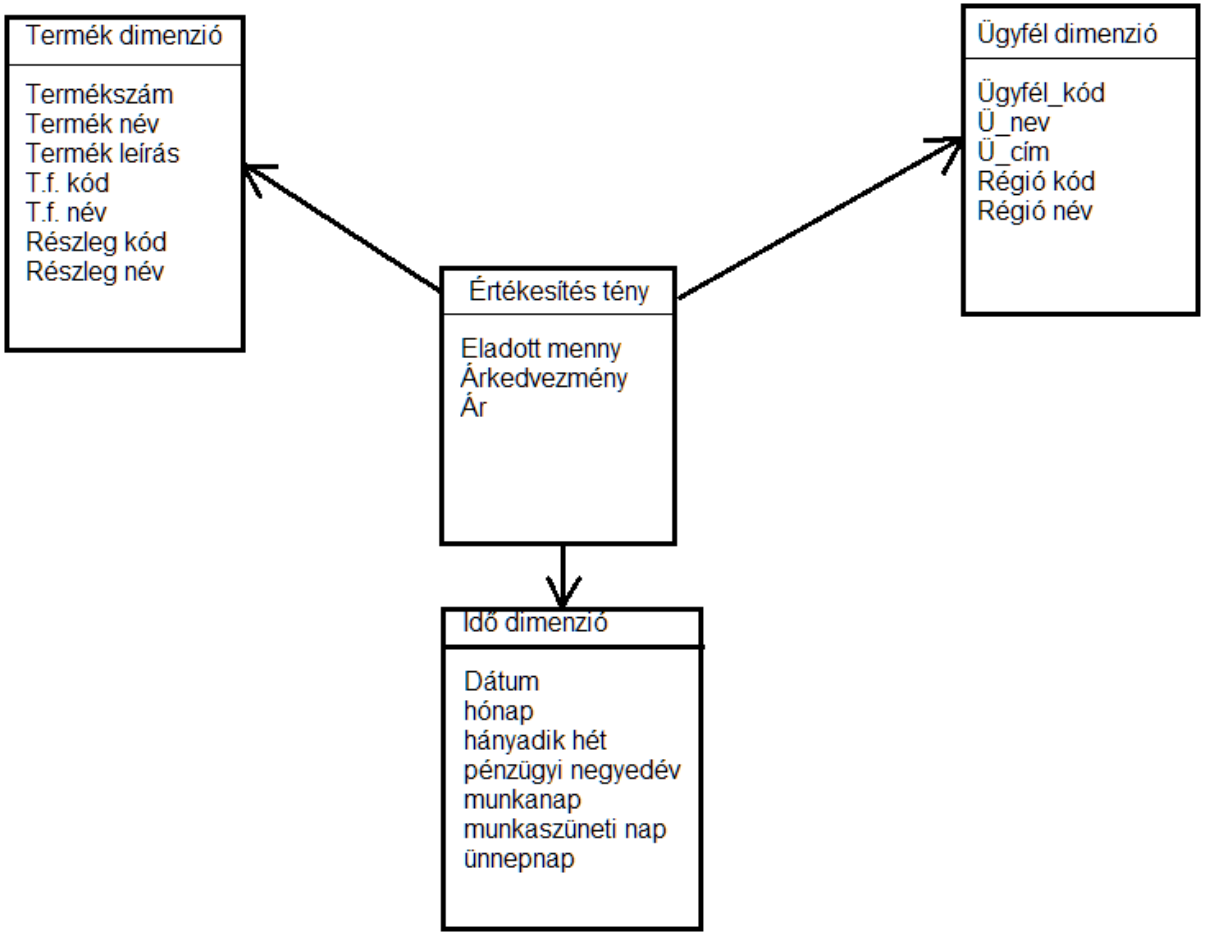
Lekérdezések támogatása - példa



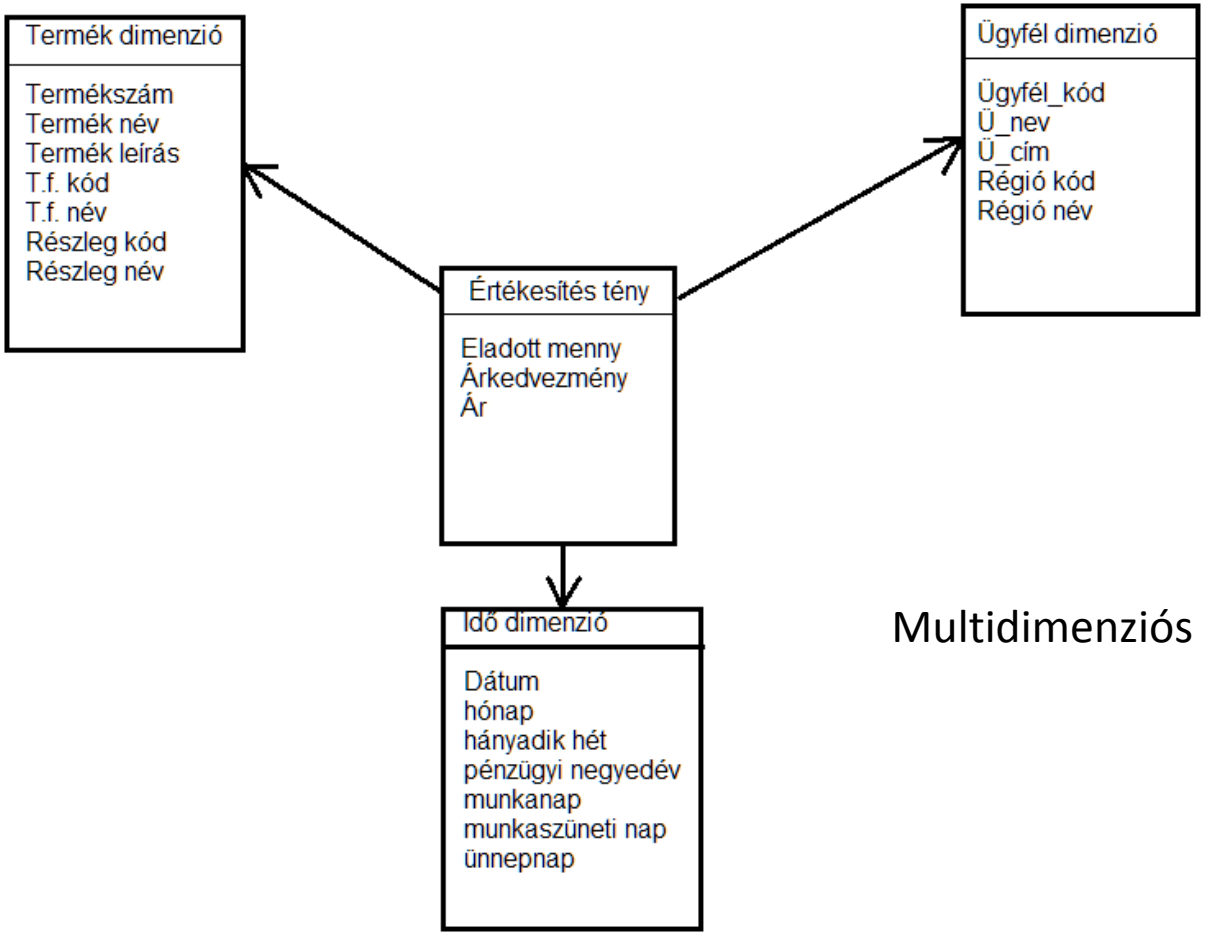
Lekérdezések támogatása - példa



Lekérdezések támogatása - példa



Lekérdezések támogatása - példa

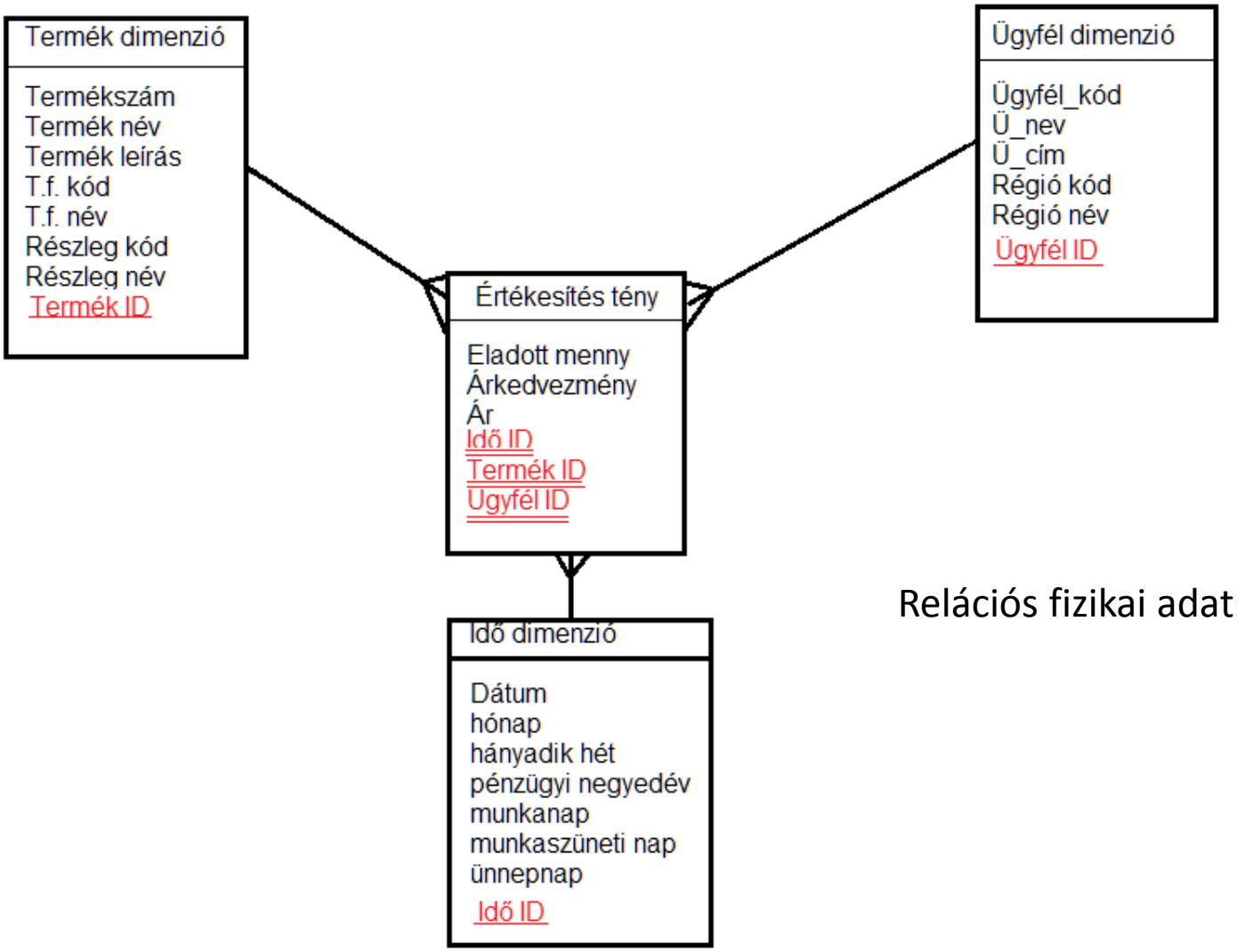


Multidimenziós logikai modell

Lekérdezések támogatása IV.

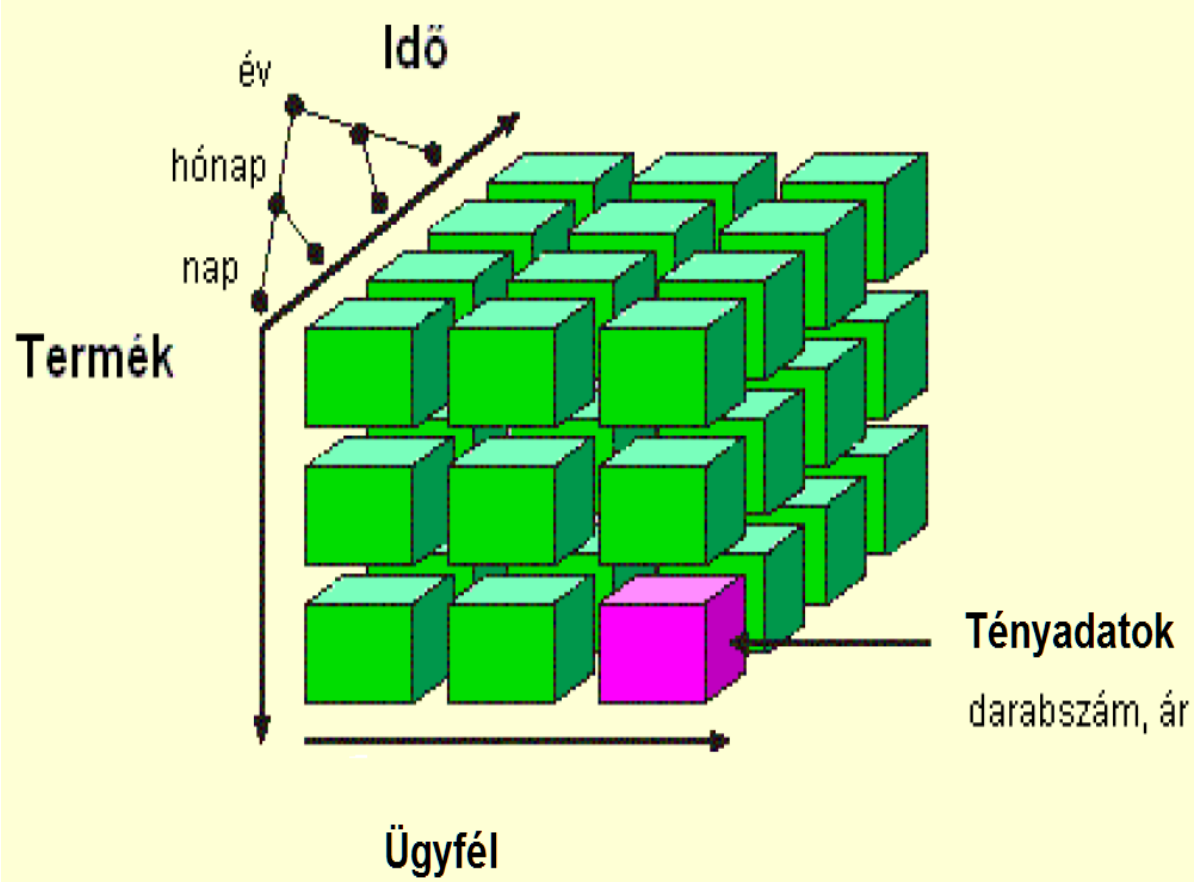
- Implementációs lehetőségek
  - Relációs
  - Natív multidimenziós
  - OO
  - …

Lekérdezések támogatása – relációs implementáció



Relációs fizikai adatmodell

Lekérdezések támogatása – natív multidimenziós implementáció



Lekérdezések támogatása IV.

Teljes modell

- A ténytáblák csak dimenziókat, a dimenziók csak tényeket kapcsolnak össze
- Adattárház busz
- Konform dimenziók
  - Definíciója
  - Jelentősége

Lekérdezések támogatása V.

- Aggregátumok
  - Előre kiszámított, majd eltárolt lekérdezés eredmény
  - Tip: tényadatok összegzése a dimenziók hierarchiái mentén
  - Használati jellegzetességek
  - “Teljesítmény” kézben tartásának fontos eszköze
  - Aggregátumok lehetséges száma

Lekérdezések támogatása VI.

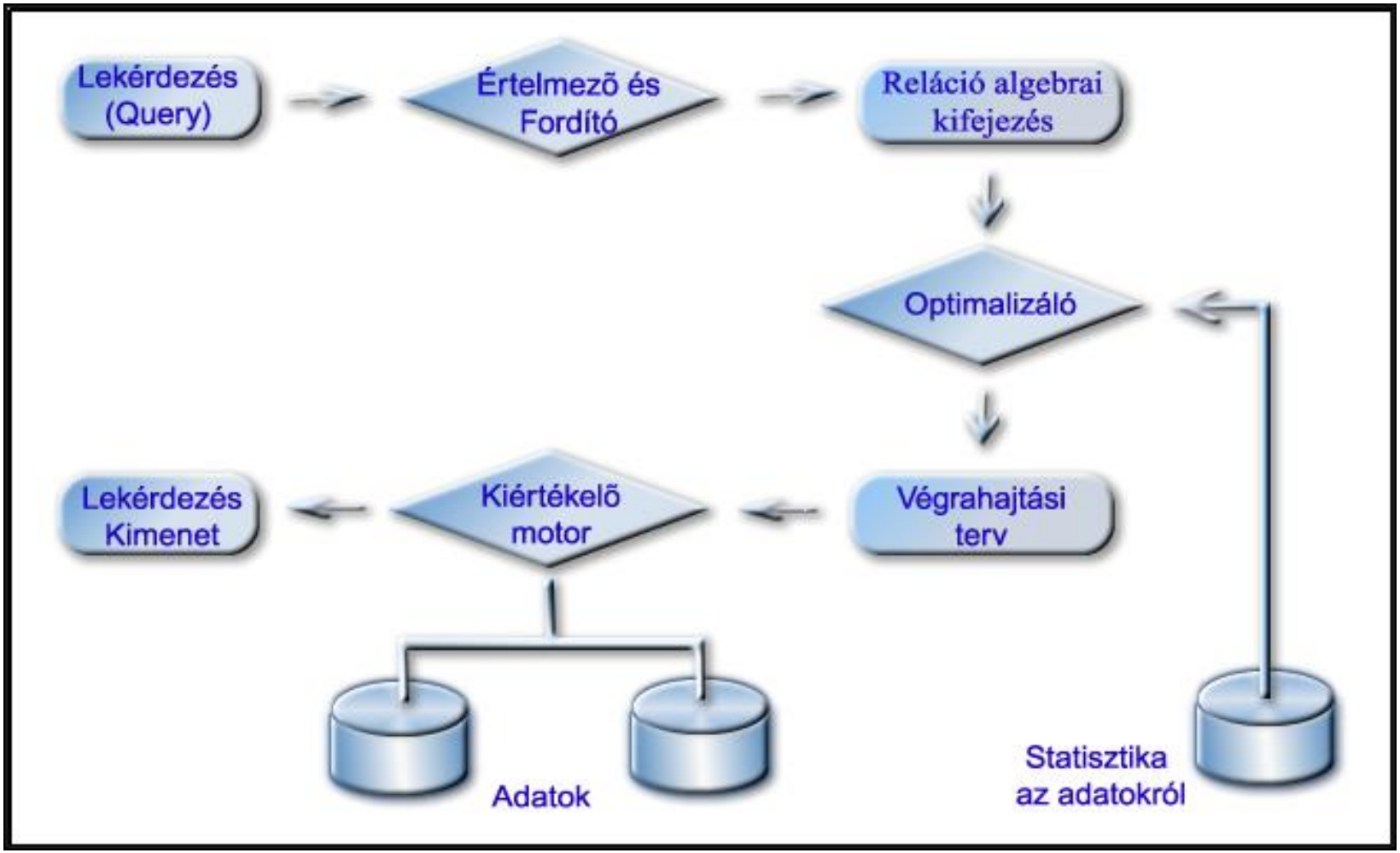
Végfelhasználói hozzáférési módok

- Riportok
  - Konzerv
  - Paraméterezett
- OLAP (ROLAP, MOLAP, HOLAP)
  - Drill down, rolling up, drill across, slice&dice
- Ad-hoc lekérdezések
  - Aggregátumnavigáció
- Adatbányászat

Lekérdezések támogatása VII. - optimalizálás

- Heurisztikus, szabály alapú optimalizálás (volt)
- Költség alapú optimalizálás (volt)
  - Katalógus költségbecslés
  - Operációk, műveletek áttekintése
  - Kifejezés-kiértékelés
  - Az optimális végrehajtási terv kiválasztása
- Lekérdezés optimalizálás csillagsémákon (most)

Optimalizálás - áttekintés



Lekérdezés optimalizálás csillagsémákon

- Lényegében egy illesztés a ténytábla és a dimenziós táblák között
- Dimenziós táblákat nem join-olunk
- “Snowflake” séma: gyenge browsing teljesítmény, relációk növekvő száma

Csillagséma optimális lekérdezése (feltételei, Oracle)

- Egyattribútumos bitmap index definiálása a tény valamennyi idegen kulcsára
- inicializáló paraméter beállítása (engedélyezés)
- költségalapú optimalizáló használata

Csillagtranszformáció

Transzparens a felhasználónak

Elve:

- 1. Dimenziós ID-k meghatározása
- 2. pontosan a szükséges tényrekordok kiolvasása bitmap segítségével
- 3. dimenziós rekordok illesztése a kiolvasott tényrekordokhoz (szükség esetén).

Csillagtranszformáció példa

```
SELECT ch.channel_class, c.cust_city, t.calendar_quarter_desc
FROM sales s, times t, customers c, channels ch
WHERE s.time_id = t.time_id
AND s.cust_id = c.cust_id
AND s.channel_id = ch.channel_id
AND c.cust_state_province = 'CA'
AND ch.channel_desc IN ('Internet','Catalog')
AND t.calendar_quarter_desc IN (‘2016-Q1',‘2016-Q2')

SELECT ch.channel_class, c.cust_city, t.calendar_quarter_desc
FROM sales WHERE
time_id IN
  (SELECT time_id FROM times WHERE calendar_quarter_desc
      IN(‘2016-Q1',‘2016-Q2'))
AND cust_id IN
  (SELECT cust_id FROM customers WHERE
cust_state_province='CA')
AND channel_id IN
  (SELECT channel_id FROM channels WHERE channel_desc IN
      ('Internet','Catalog'));
```

Működése

- a dimenziók általában kevés rekordot tartalmaznak
- dimenziók lekérdezése a dimenziós ID-kra
- time_id bitmap azonosítja a 2016. első negyedévi tényrekordokat
- time_id bitmap azonosítja a 2016. második negyedévi tényrekordokat
- hasonló bitmap-ek azonosítják a megfelelő customer-hez és channel-hez tartozó tényrekordokat
- a bitmap-eket kombináljuk logikai műveletekkel
- tényrekordok elővétele a diszkről
- dimenziós rekordok join-ja a tényrekordokhoz (módja hagyományos optimalizálás során dől el)

Mikor jó?

- Ha a where predikátuma kellően szelektív a tényrekordokra
- Ha sok tényrekord érintett az eredmény előállításában, akkor full table scan jobb lehet...

Inmon adattárház (DW) 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

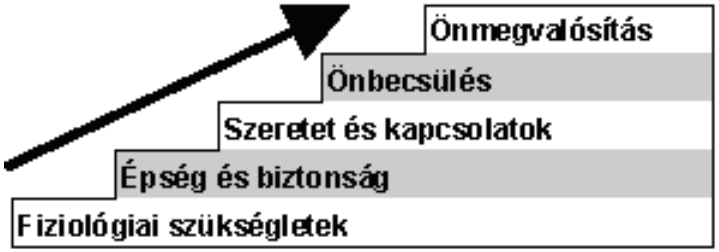
Üzleti intelligencia (BI)

Definíció (EPICOR, 2005):

„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.”

A szükségletek hierarchiája (Maslow)

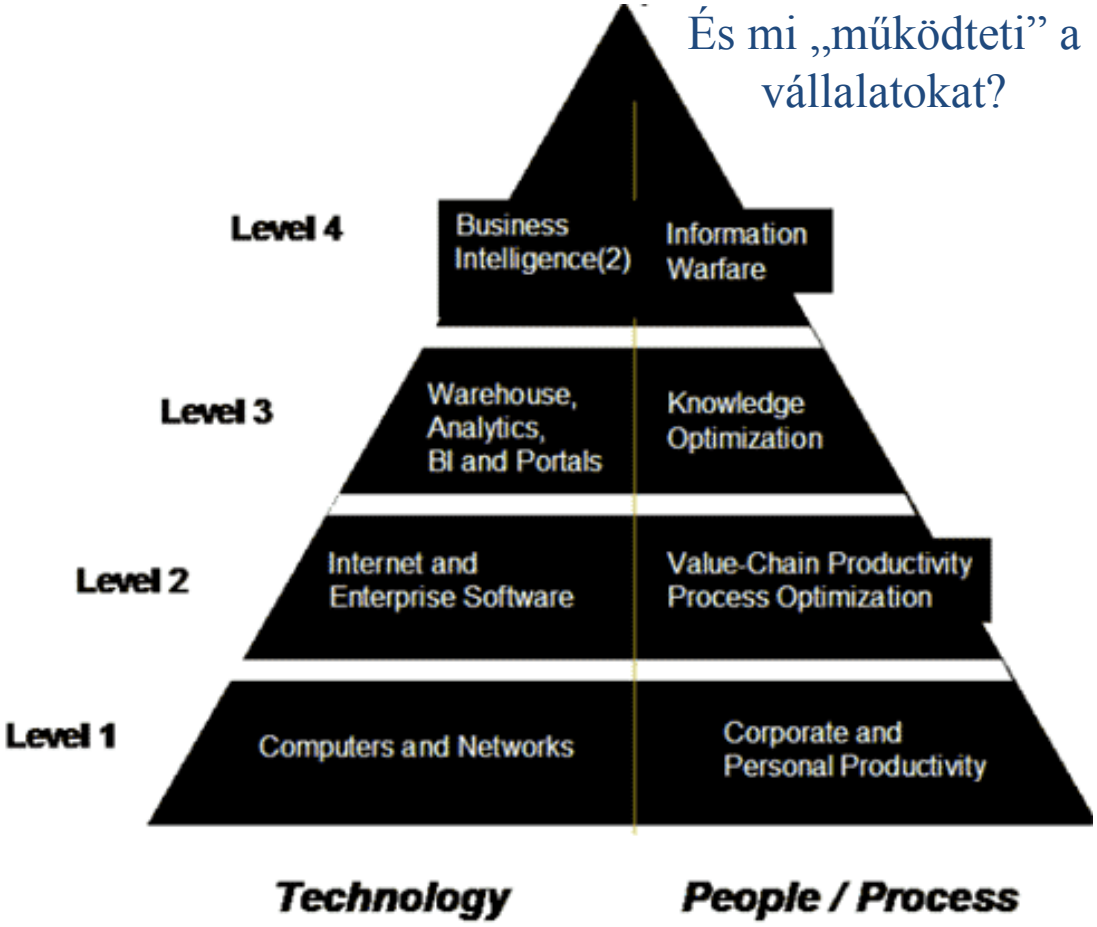
Avagy: mi „működteti” az embereket



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 észszerű 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.

És mi „működteti” a vállalatokat?

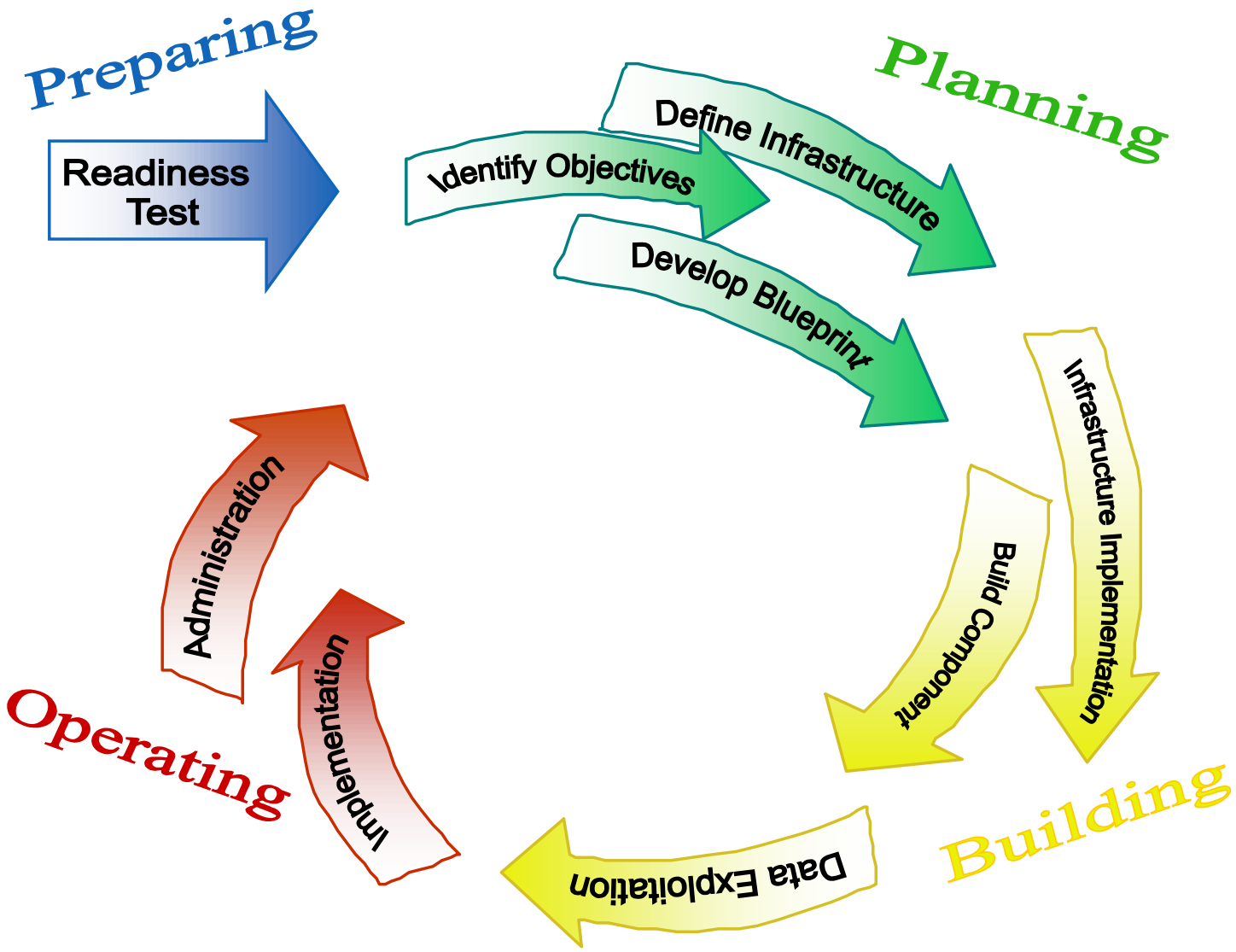


Adattárház módszertanok

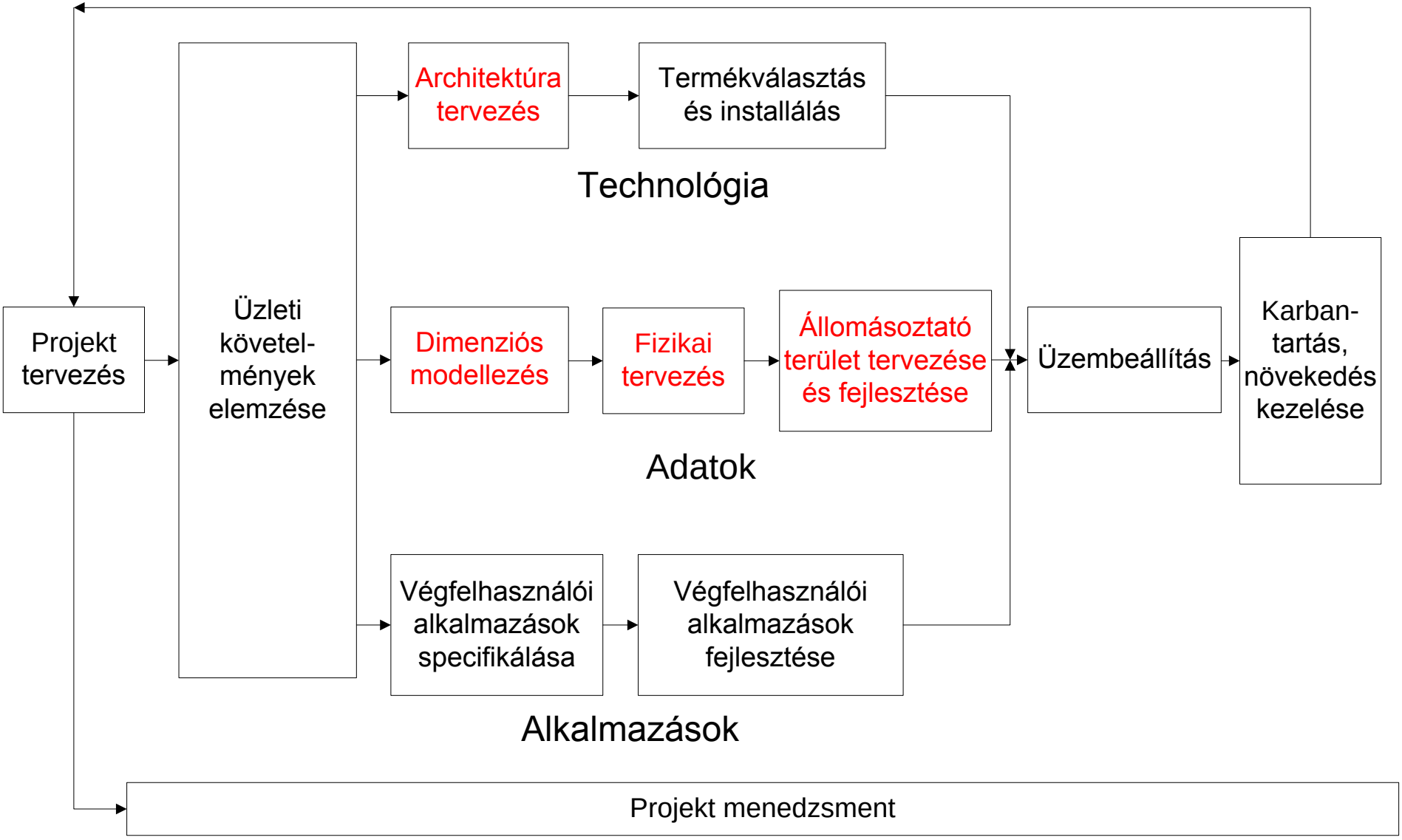
- Hadden-Kelly
- Oracle Warehouse Methodology
- Ralph Kimball
- SAS
- …

Általánosan jellemző a fázisok definiálása és az iteratív szemlélet

Hadden-Kelly



Ralph Kimball módszertana



DW projekt definiálása

- érdekeltség vagy ellenérdekeltség?
- felkészültség értékelése

Pénzügyi megfontolások

- „A ROI (Return Of Investment) az Isten”
- Amibe kerül:
  - HW, SW, belső fejlesztési költségek, külső erőforrások költsége, support, növekedés költségei.
  - Mennyisége és eloszlása jól becsülhető
- Ami haszonként várható:
  - bevételnövekedés: pl. gyorsabb piacrajutás vagy forgalomnövekedés a termékek jobb pozícionálása miatt, ...
  - költségcsökkenés: költséghatékonyabb marketing kampányok, ...

DW projekt résztvevői

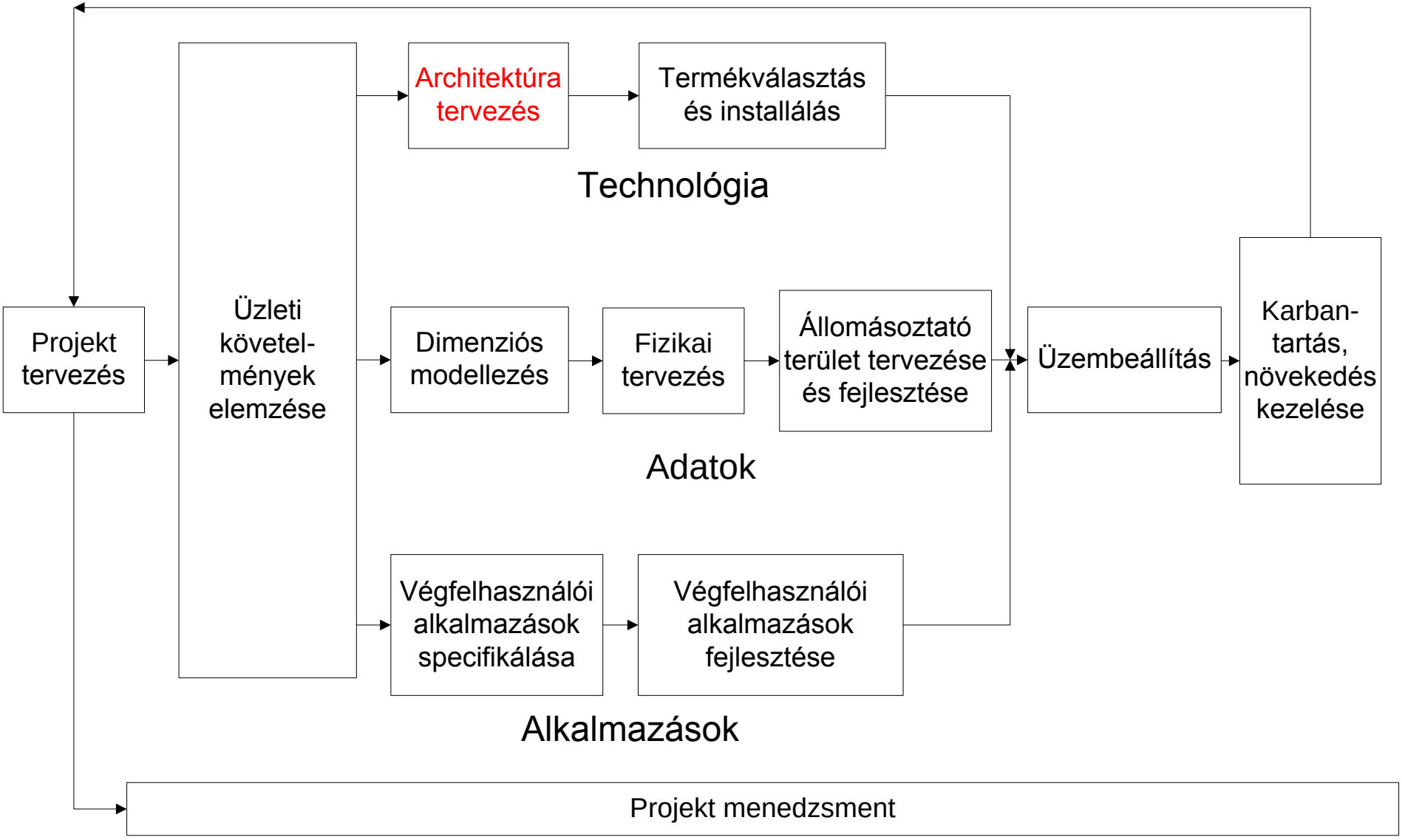
heterogén csapatra van szükség

- PM
- üzleti analitikus
- architect (főmérnök)
- adatmodellező
- betöltés tervező
- front-end tervező
- biztonsági tervező
- data steward (adatgondnok)
- DBA
- Oktatók
- Betöltés programozó
- Front-end fejlesztő
- üzemeltető
- (adat-)minőségbiztosító, tesztelő…

Követelmények összegyűjtése

- Alapelvek
- Előkészületek az interjúkhoz
- Interjúk lebonyolítása
- Sikerkritériumok meghatározása
- Konszolidálás, priorizálás, konszenzus kialakítása

Architektúra tervezés

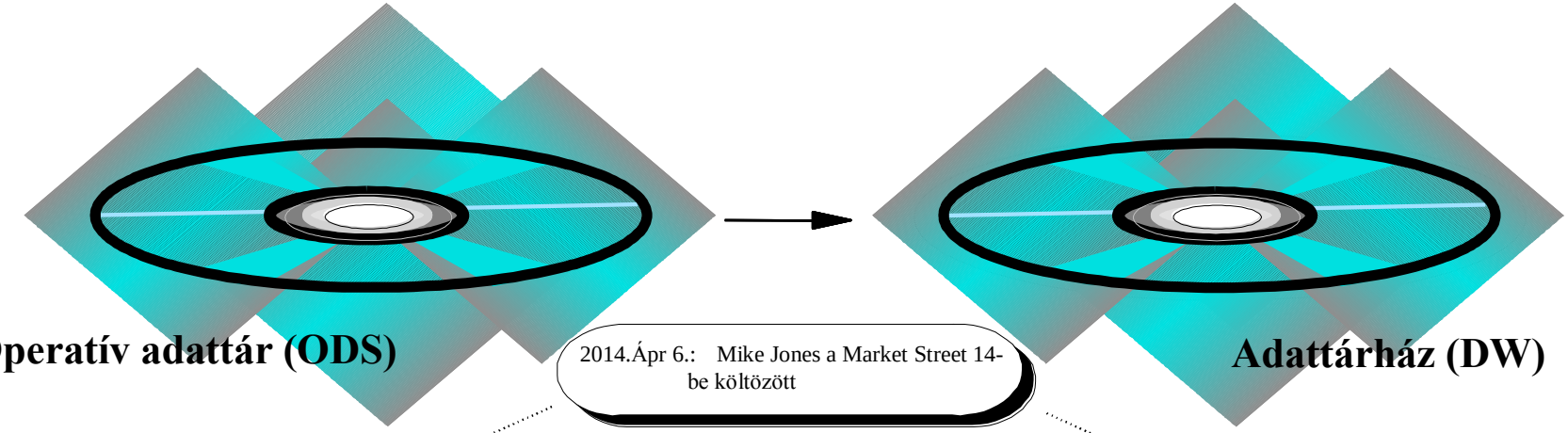


DW koncepcionális architektúra főbb elemei

- 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

ODS vs. DW

Adatváltozás hatása ODS ill. DW esetén



Operatív adattár (ODS)

Adattárház (DW)

2014.Ápr 6.: Mike Jones a Market Street 14-be költözött

Ügyfél	– Mike Jones
Cím	– 123 Main Street
Hitel	– AA
Utolsó módosítás	– 2014. jan. 12

Ügyfél	– Mike Jones
Cím	– 14 Market Street
Hitel	– AA
Utolsó módosítás	– 2014. ápr. 6

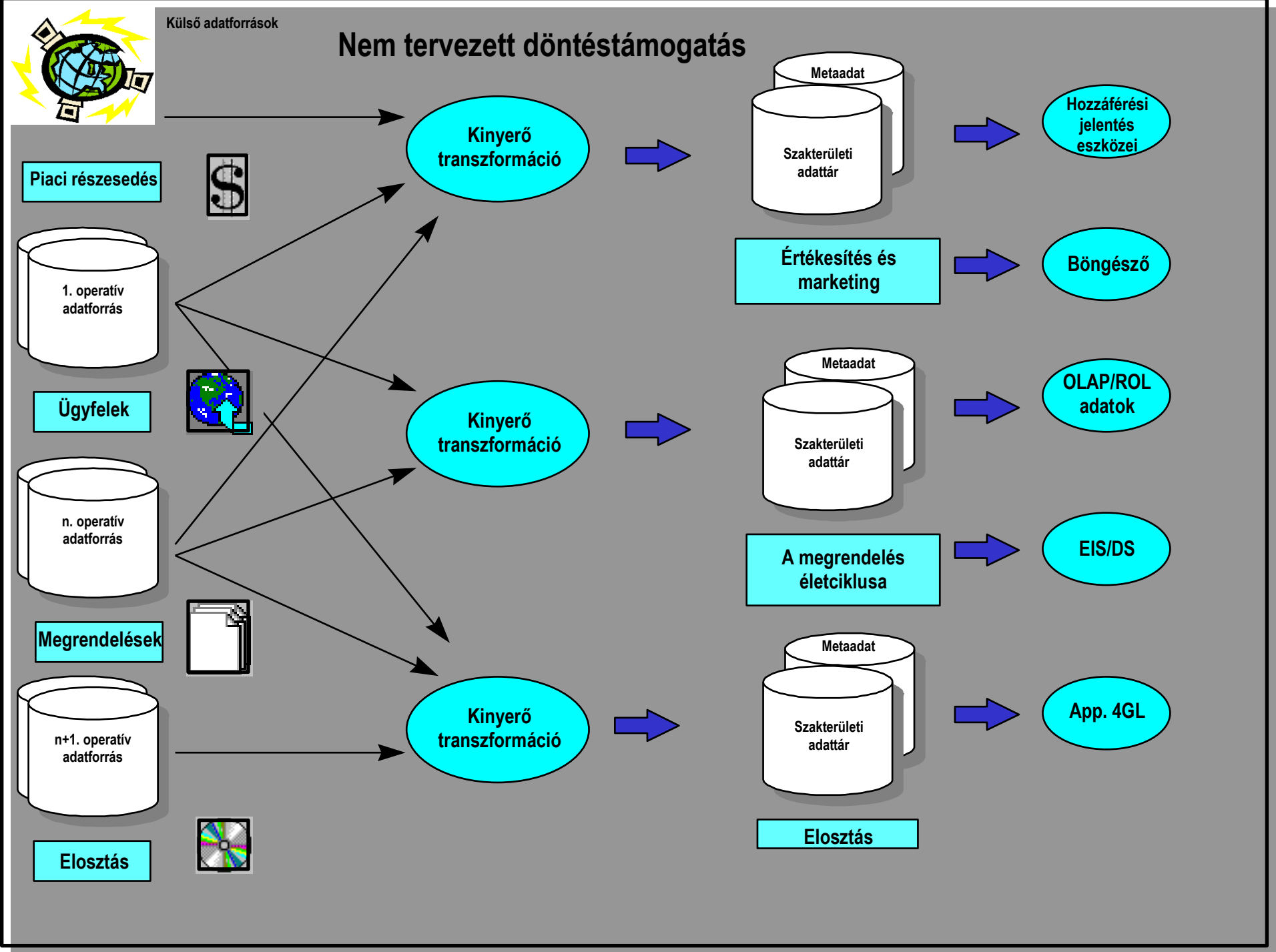
Ügyfél	– Mike Jones
Cím	– 123 Main Street
Hitel	– AA
Mettől:	– 2014. jan. 12
Meddig:	– jelen

Ügyfél	– Mike Jones
Cím	– 123 Main Street
Hitel	– AA
Mettől:	– 2014. jan. 12
Meddig:	– 2014. ápr. 5.

Ügyfél	– Mike Jones
Cím	– 14 Market Street
Hitel	– AA
Mettől:	– 2014. ápr. 6.
Meddig:	– jelen

Konc. architektúra összehasonlítási szempontok

- Költségek
- Megvalósítási idő
- Rugalmasság
- Funkcionalitás
- Adatkonzisztencia
- Illeszkedés a vállalati hierarchiához


Külső adatforrások
Nem tervezett döntéstámogatás
Piaci részesedés
1. operatív adatforrás
Ügyfelek
n. operatív adatforrás
Megrendelések
n+1. operatív adatforrás
Elosztás
Kinyerő transzformáció
Kinyerő transzformáció
Kinyerő transzformáció
Metaadat
Szakterületi adattár
Értékesítés és marketing
Metaadat
Szakterületi adattár
A megrendelés életciklusa
Metaadat
Szakterületi adattár
Elosztás
Hozzáférési jelentés eszközei
Böngésző
OLAP/ROL adatok
EIS/DS
App. 4GL

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

ODS

Metaadat

Szakterületi adattár

Megrendelés életciklusa

Metaadat

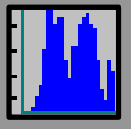
Szakterületi adattár

Elosztás

Web böngésző
Adatbányászat

ROLAP/OLAP

EIS/DSS
Riportok
Hozzáférés
(API-k, Middleware)



Virtuálisan integrált szakterületi adattárak

Külső Adatforrások

Piaci részesedés

1. operatív adatforrás

Ügyfelek

n. operatív adatforrás

Megrendelések

n+1. operatív adatforrás

Elosztás

Szemantikai integrációs folyamat

Metaadat

Szakterületi adattár

Értékesítés és marketing

ODS

Metaadat

Szakterületi adattár

Megrendelés életciklusa

Metaadat

Szakterületi adattár

Elosztás

MIDDLEWARE

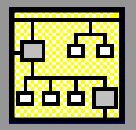
Metaadat

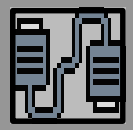
Adattárház-adminisztráció

Web böngésző
Adatbányászat

ROLAP/OLAP

EIS/DSS
Riportok
Hozzáférés
(API-k, Middleware)





Függő szakterületi adattár (hub-and-spoke architektúra)

Külső adatforrások

Piaci részesedés

1. operatív adatforrás

Ügyfelek

n. operatív adatforrás

Megrendelések

n+1. operatív adatforrás

Elosztás

Szemantikai és fizikai integrálás

Intermediate Data Store

(Operatív adattár)

Közbülső adattár

Részletes Adattárház

Metaadat

Szakterületi adattár

Hálózat

Értékesítés

Metaadat

Szakterületi adattár

Pénzügy

Web böngésző

Adatbányászat

ROLAP/MOLAP

EIS/DSS

Riportok

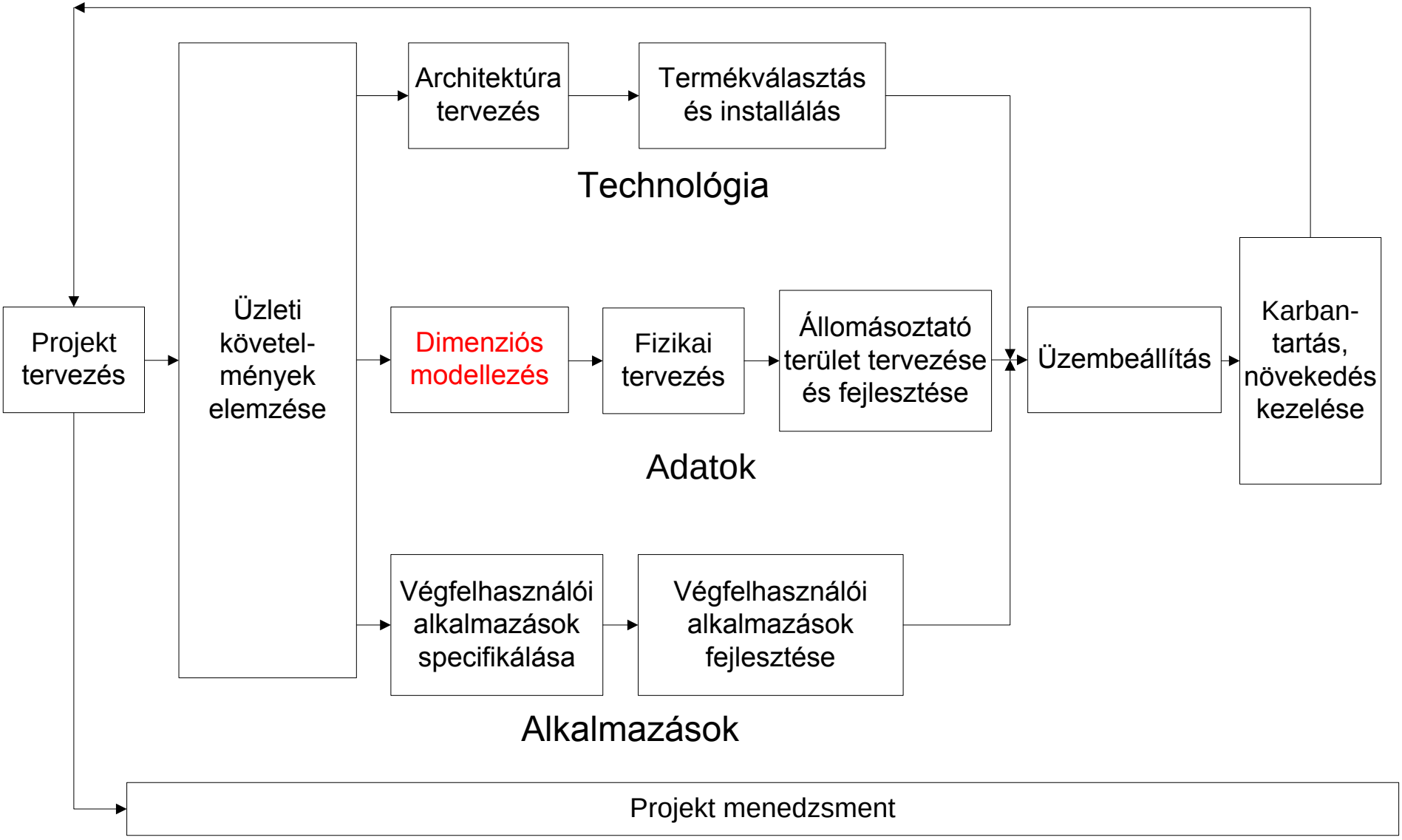
Hozzáférés (API-k, Middleware)

Metaadat

Szakterületi adattár

Marketing

Ralph Kimball módszertana



Dimenziós modellezés

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

Négylépéses dimenziós modellezés

1. Üzleti folyamat azonosítása
2. Tényadat granularitásának megválasztása (üzleti szinten)
3. Dimenziók (és attribútumaik) azonosítása
4. Tény attribútumok 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
- nem elég részletes: elemzéseket akadályozhat meg
- LE KELL ÍRNI A TÉNYREKORD PONTOS JELENTÉSÉT

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

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

Dimenziós tervezési elvek

- A pontosan ismerni és értelmezni kell tudni az adatokat
- Dimenziós táblák: leíró attribútumuk, 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

Idő dimenzió

IDOSZAKOK_DIMENZIO		
IDOSZAK_ID	<pk>	NUMBER(4)
NAPTARI_DATUM		DATE
NAP_MEGNEVEZESE		CHAR(10)
NAP_MEGNEVEZESE_ANGOL		CHAR(9)
NAP_ROVID_BETUJELE		CHAR(3)
NAP_ROVID_BETUJELE_ANGOL		CHAR(3)
HET_HANYADIK_NAPJA		NUMBER(1)
HONAP_HANYADIK_NAPJA		NUMBER(2)
EV_HANYADIK_NAPJA		NUMBER(3)
PENZUGYI_NEGYEDEV_NAPJA		NUMBER(3)
HONAP_HANYADIK_HETE		NUMBER(2)
EV_HANYADIK_HETE		NUMBER(2)
HONAP_ROVIDITESE		CHAR(5)
HONAP_ROVIDITESE_ANGOL		CHAR(3)
EV_HANYADIK_HONAPJA		NUMBER(2)
NAPTARI_NEGYEDEV		NUMBER(1)
NEGYEDEV_HONAPJA		NUMBER(1)
NEGYEDEV_HETE		NUMBER(2)
NEGYEDEV_NAPJA		NUMBER(3)
PENZUGYI_NEGYEDEV		NUMBER(1)
PENZUGYI_NEGYEDEV_HONAPJA		NUMBER(1)
PENZUGYI_NEGYEDEV_HETE		NUMBER(3)
HANYADIK_FELEV		NUMBER(1)
HONAP_MEGNEVEZESE		CHAR(10)
HONAP_MEGNEVEZESE_ANGOL		CHAR(9)
EVSZAM		NUMBER(4)
ROVID_EVSZAM		NUMBER(2)
PENZUGYI_EVSZAM		NUMBER(4)
PENZUGYI_ROVID_EVSZAM		NUMBER(2)
IDOSZAK_MEGNEVEZESE		CHAR(40)
IDOSZAK_MEGNEVEZESE_ANGOL		CHAR(40)
IDOSZAK_ROVID_NEVE		CHAR(3)
IDOSZAK_ROVID_NEVE_ANGOL		CHAR(3)
NAPOK_SZAMA_FIX_IDOPONTTOL		NUMBER(4)
KARACSONY_JELZO		CHAR(1)
HUSVET_JELZO		CHAR(1)
ALAPERTELMEZETT_IDOSZAK		CHAR(1)
NAPTIPUS		NUMBER(1)
NAPTIPUS_MEGNEVEZES		CHAR(9)

Ténytábla tervezés

Tényadatok a lehető legkisebb granularitásban (vö.: hiányzó "vásárlói kosár" analízis).

- Additív tényadatok
  - Hacsak lehetséges, összegezhetőnek kell választani.
- Nem additív tényadatok
  - Egyáltalán nem összegezhetők, egyetlen dimenzió mentén sem.
- Szemi-additív tényadatok
  - minden dimenzió szerint összegezhető, kivéve az időt. (általánosabban: bizonyos dimenziók szerint összegezhetők, mások szerint nem)

Dimenziós tervezési minták I.

Ténynélküli ténytáblák

- pl. diákok óralátogatási szokásai (idő, tárgy, terem, diák, tanár függvényében)
- (kampány) lefedettségi táblák
  Pl. az eladás ténye termék, bolt, idő, kampányjellemzők függvényében. Nem ad választ arra, hogy mit NEM adtak el abból, amiről a kampány szólt!
  Megoldás: egy másik ténytábla rekordja jelentse a kampányban való részvételt

tényrekord jelentése: van olyan...

Valójában klasszikus több-több kapcsolatok

Dimenziós tervezési minták II.

Állapot- és esemény-tények

- Esemény-tény: egyetlen időpont
- Állapot-tény: két időpont
  - 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!

Dimenziós tervezési minták III.

Role-playing dimenziók

- pl. idő, cím,... többféle jelentést is hordozhat a tényadathoz kapcsolódóan
- egyetlen fizikai dimenzió, amely több idegen kulccsal kapcsolódik a tényrekordhoz, ezek értelmezése különböző

Degenerált dimenziók

Számla, tételekkel. A tételek lesznek a tényadatok.

Mi legyen a számlaszámmal?

- Vannak olyan leíró (rövid, dimenziós jellegű) adatok, amelyeket a ténytáblában helyezünk el kapcsolódó dimenzió nélkül.
- Pl.: dokumentum egyedi azonosító száma
- A forrásrendszerben lehet könnyen azonosítani velük valamit
- Egyedi megfontolás. Normálisak, várhatók, hasznosak

Junk dimenziók

- Flag-ek és szöveges leírók nem mindig szervezhetők értelmes dimenziókba
- Ténytáblában nem célszerű elhelyezni
- Egy vagy néhány jelentés nélküli dimenziót alkothatnak.

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

Pl. az ügyfél elköltözik, címe megváltozik

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

1. felülírás

Pl.: az ügyfelek címei változhatnak, ha elköltözik.

Ügyfél ID	**Ügyfél neve**	**Ügyfél címe**
123	Gipsz Jakab	Budapest, Tó u. 15.

1. felülírás

Ügyfél ID	**Ügyfél neve**	**Ügyfél címe**
123	Gipsz Jakab	Debrecen, Fő u. 3.

Egyszerű, de nincs history.

2. “old” mező létrehozása

Ügyfél ID	Ügyfél neve	Ügyfél címe
123	Gipsz Jakab	Budapest, Tó u. 15.

2. A jelenlegi és az előző állapot jellemzésével

Ügyfél ID	Ügyfél neve	Ügyfél előző címe	Ügyfél jelenlegi címe
123	Gipsz Jakab	Budapest, Tó u. 15.	Debrecen, Fő u. 3.

egyszerű, de korlátozottak a lehetőségei.

3. Új dim. rekord készítése

Ügyfél ID	Ügyfél neve	Ügyfél címe
123	Gipsz Jakab	Budapest, Tó u. 15.

3. új dimenziós rekord minden változáshoz

Ügyfél ID	Ügyfél neve	Ügyfél címe	Tól	Ig
123	Gipsz Jakab	Budapest, Tó u. 15.	1989. júl. 15.	2005. szept. 6.
196	Gipsz Jakab	Debrecen, Fő u. 3.	2005. szept. 7.	???????

particionálja a history-t, nehézkesebb a lekérdezés

Gyakorlat: Reklámkampány analízis

1. Mi a korreláció bizonyos oksági tényezők (engedmények, kiállítás módja, kuponok) és a pezsgősvödrök eladása között (darabban és forintban) szupermarketenként, termékenként és 4 hetes eladási periódusonként?
2. Változik-e a pezsgősvödrök árérzékenysége üzletenként?

Szükség van továbbá az alábbi standard riportokra:

- Piaci részesedés termékkategóriákként, szupermarketenként és időszakonként
- A legjobban fogyó márkák szupermarketenként és időszakonként

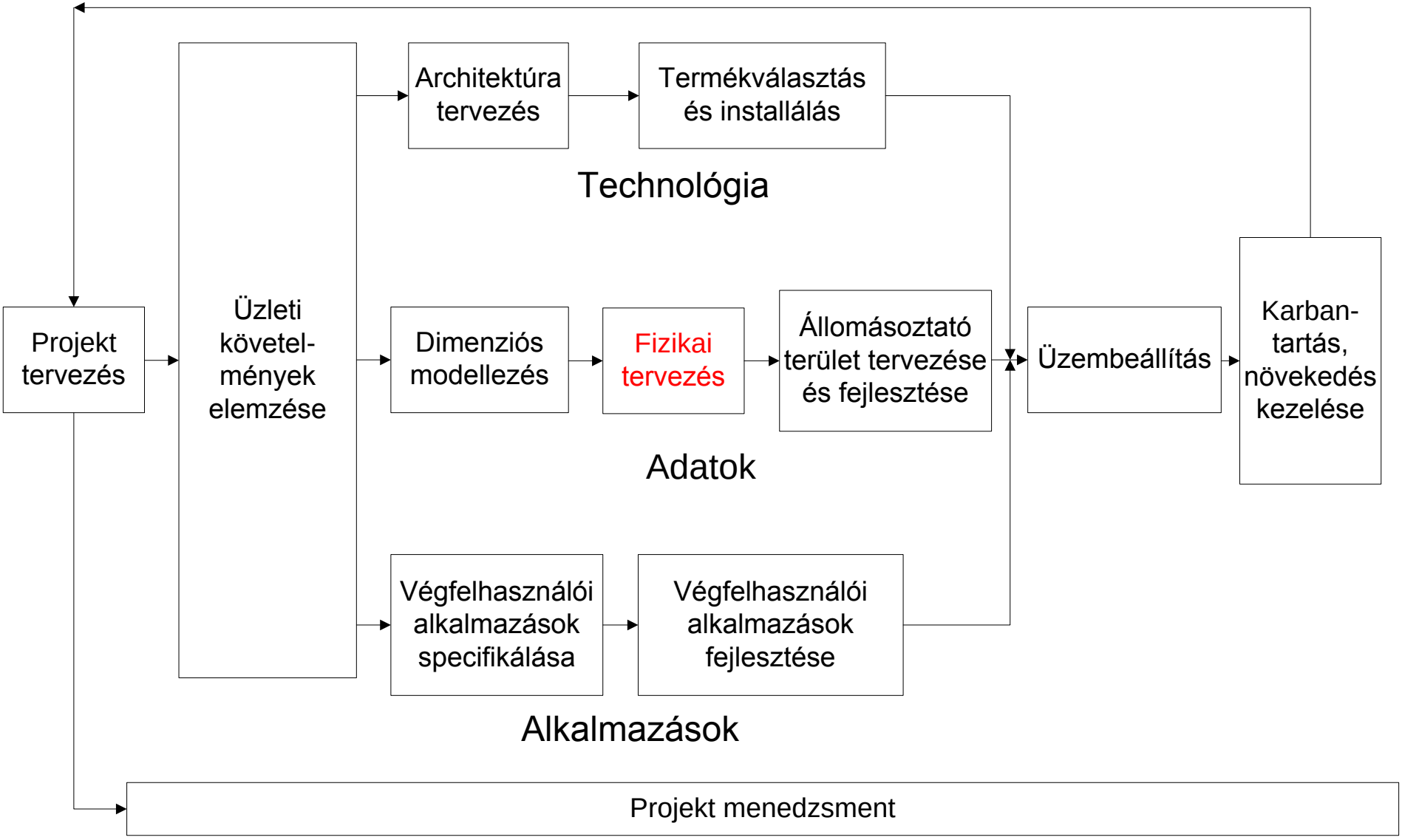
Az adatforrások:

- a szupermarketek eladási adatai 4 hetes összesítésekben termékkódokként és szupermarketenként
- az így kapott file tartalmaz információt az alkalmazott kedvezményekről, a kiállítás módjáról, a kuponokról, az eladott darabszámról, az eladási árról, az átlagos kiskereskedelmi árról és a kereskedelmi hierarchiáról.

Attribútumlista:

Márka, kategória, kuponok, szín, kiállítás módja, eladási ár, íz, üzlet, csomagolás, költség, év, évszak, termékkód, darabszám, hét, cím (üzlet), dátum, kedvezmények, átlagos kiskereskedelmi ár

Ralph Kimball módszertana



FIZIKAI TERVEZÉS

1. ld. fizikai adatbázis tervezésről eddig tanultak
2. lekérdezés optimalizálás kérdései
3. összegzések (aggregációk) tervezése

Összegzések tervezése

- DEF.: előre kiszámított speciális lekérdezés, amikor a ténytábla tényadatait összegezzük bizonyos feltételek mentén.
- Másképpen: a dimenziókban lévő hierarchiák "összenyomása" és a tényadatok ennek megfelelő összeadása. (Ezért fontos a tényadatok additivitása.)
- Legfontosabb eszköz a teljesítmény kézbentartására
- Akár 1000 összegzés is létezhet egyidejűleg!

Összegzések tárolása

Új tényrekordokra van szükség, amelyhez új dimenziós táblák kellenek és új mesterséges kulcs.

Az új rekordok kétféleképpen tárolhatók:

- új ténytáblában
- új szintjelző mezők segítségével (kevésbé jellemző)

Összegzés új ténytáblában

- Az összegzett tényrekordokat új táblában helyezzük el (Praktikusan a meglévő ténytáblából is képezhetjük a szerkezetét).
- Hasonlóképpen az új dimenziós táblákat is képezhetjük a meglévő dimenziósakból, a granularitás csökkentésével
- Példa:
  - eredeti tény: termékek megrendelése, dimenzió: termék
  - aggregátum tény: márkák megrendelése, dimenzió: márka
- A tényrekordokat összegeztük márkák szerint, új kulcsot definiáltunk a márka dimenzióhoz.

Összegzések méretezése 1.

- Elv: legalább 10:1 mértékű rekordszámcsökkenés
- A választás szempontjai a (**dimenzió) kompressziója** és az **együttes előfordulási gyakoriság** (density).
- A kompresszió: ha egy márkához átlagosan (!) 50 termék tartozik, akkor a márkára definiált összegzés 50-szeres kompressziójú.
- Termék-bolt-nap előfordulási gyakorisága: ha egy boltban egy nap eladják a termékek 10%-át (átlagosan)
- Márka-bolt-nap előfordulási gyakorisága: ugyanakkor egy boltban egy nap eladják a márkáknak az 50%-át (átlagosan)

Összegzések méretezése 2.

- A várható rekordok száma az összegzés ténytáblájában = <sorok száma a dimenziókban összesen> szorozva <előfordulási gyakoriság>
- Az együttes előfordulási gyakoriságok előre általában nem ismertek…
- Megoldás: becslések, ill. tapasztalati méretezés (ha elég jó, akkor meghagyjuk ☺)

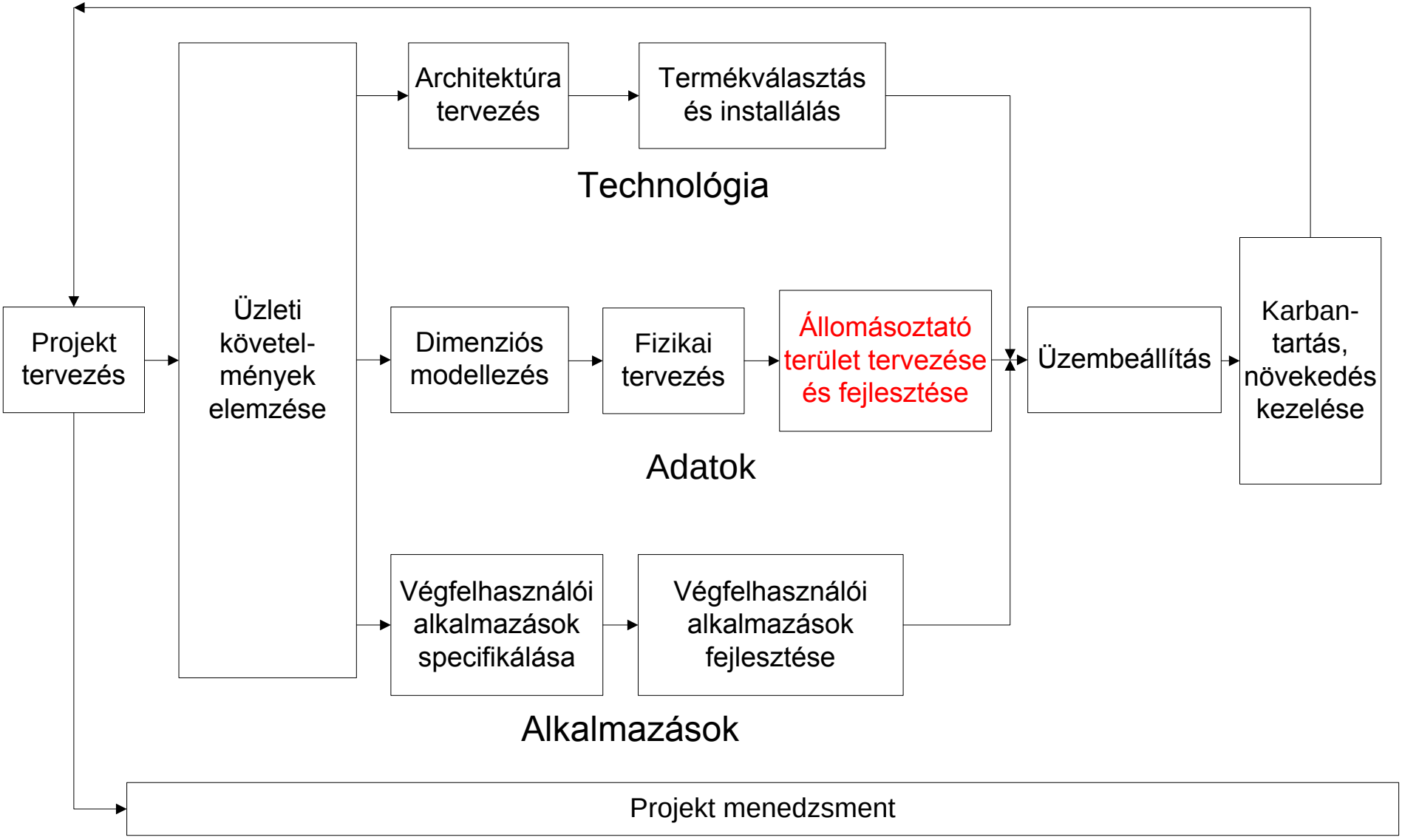
Összegzések méretezése 3.

way	**Termék dim.**	**Üzlet dim.**	**Időszak dim.**	Termék	Üzlet	Időszak	Gyakori-ság	Rekord-szám (millio)	Összeg-zés komp-resszió
0	SKU	üzlet	nap	10000	1000	90	0.1	90,000,000	
1	márka	üzlet	nap	2000	1000	90	0.5	**90,000,000**	**1**
1	SKU	kerület	nap	10000	100	90	0.5	**45,000,000**	**2**
1	SKU	üzlet	hónap	10000	1000	3	0.5	**15,000,000**	**6**
2	márka	kerület	nap	2000	100	90	0.8	**14,400,000**	**6**
2	márka	üzlet	hónap	2000	1000	3	0.8	**4,800,000**	**19**
2	SKU	kerület	hónap	10000	100	3	0.8	**2,400,000**	**38**
3	márka	kerület	hónap	2000	100	3	1	**600,000**	**150**
Dimenzió kompressziók:									
	Termék-márka		5:1						
	Üzlet-kerület		10:1						
	Nap-hónap		30:1						

Összegzés navigáció

- Új réteg. Nyilvántartja a létező összegzéseket és meghatározza, hogy melyik a legalkalmasabb a felhasználói lekérdezés kiszolgálására.
- Teljesítőképesség és kényelmes használat
- Nagy a veszélye a túl sok összegzés definiálásának
- Nem mindegyik összegzés csökkenti jelentősen a sorok számát, ezeket futási időben kell kiszámolni.
- Számos adatbáziskezelőnek része (pl. Oracle)

Ralph Kimball módszertana



ÁLLOMÁSOZTATÓ TERÜLET TERVEZÉSE
(Az ETL egyes fontosabb kérdései)

Back-room: data acquisition & staging

Data acquisition (adat kinyerés) I.

- forrásrendszer minimális terhelése
- adat nem veszhet el és/vagy sérülhet
- teljes vs. inkrementális
- kezdeti ill. rendszeres
- flat file vs. adatbázis kapcsolat vs. hordozható táblatér
- metaadat gyűjtés/vezérlés
- gyakoriság (ODS esetén sajátos megfontolások)
- tipikus források (pl. SAP) esetén „dobozos” interfész

Data acquisition (adat kinyerés) II.

- DW fejlesztési erőfeszítések 60%-a
- adatelemek kiválasztása
- változások érzékelése (tranzakciós napló)
- full extract, ha:
  - változás nem jól követhető
  - szinkronizációhoz
  - kis táblák esetén

Data acquisition (adat kinyerés) III.

Adatkinyerés módja	Előnyök	Hátrányok
Adott időnként egyedi tábla másolatok (Full snapshot)	Egyszerű Nem kell a forrásrendszert módosítani Forrásrendszer terhelése átidőzíthető	Erőforrásigényes mindkét oldalon Információvesztés lehetséges Nagy késleltetés
Adott időnként egyedi tábla változásadatok	Forrásrendszer terhelése átidőzíthető Információvesztés valószínűsége kisebb	A forrásrendszer módosításával járhat Nem mindig megvalósítható Nagy késleltetés
Változásadatok eseményvezérelt kinyerése egyes táblákból	Kitüntetett adatokra kis késleltetés Információvesztés valószínűsége még kisebb	Viszonylag költséges Folyamatos többletterhelés a forrásnak Nem mindig megvalósítható Forrásrendszer módosításával jár együtt
Változásadatok kinyerése teljes tranzakció kontextusra, eseményvezérelten	Információvesztés nincs Késleltetés nincs	Költséges Bonyolult technológia Folyamatosan nagyobb többletterhelés a forrásnak Nem mindig megvalósítható Forrásrendszer módosításával jár együtt

Staging (állomásoztatás) I.

- Itt keletkezik a legtöbb hozzáadott érték
- Tervezése időigényes
- dimenziós és ténytáblák előállítása (bulk load!)
- flat file vs. relációs vs. egyedi struktúrák
- C, Cobol, utility-k, ill. adatbázis műveletek (sok overhead)
- archiválás
- adatmodell: teljesítmény és könnyű fejlesztés

Staging (állomásoztatás) II.

- **metaadat vezérelt elv**: a folyamatok a metaadattárból vezéreltek, mintsem beágyazottak az ETL programokba
- aktív-passzív metaadat (utasítás ill. dokumentál)
- változások a metaadattáron keresztül megvalósíthatók

Staging (állomásoztatás) III.

- adattípus konverziók
- adatforrások integrálása
  - surrogátum kulcsok generálása
- referenciális integritás kezelése
- cleansing: (duplikátumok, hibás-hiányzó adatok) pontos specifikálás
- NULL: sok rendszerben nincs kódja

Adatminőség javítása

- Minőségi standard-ok definiálása (pontosság, teljesség, ellentmondás-mentesség, egységesség, frissesség)
- Javítás
  - spec. karakterek ellentmondásos/változó használata (F N M F m f y n …)
  - mező használat dokumentálatlan célra
  - mező használat többféle célra
  - adat fejlődés - jelentésváltozás
  - hiányzó - hibás - dupla értékek
- Javításuk a forrásrendszerben kívánatos, de nem mindig lehetséges
- A nevek és címek javítása külön tudomány

Staging (állomásoztatás) IV.

Job vezérlés

- ütemezés: idő és/vagy eseményvezérelt
- monitorozás
- naplózás (adatbázisba, segít optimalizálni is)
- kivételkezelés (visszautasított rekordok kezeléséhez hely, idő, paraméterek kellenek)
- hibakezelés (crash recovery, stop, restart, állítható commit set jól jöhet)
- értesítés eseményről (mail, SMS)

Staging (állomásoztatás) V.

Mentések

- még UPS, mirroring, redundáns HW esetén is kell biztonsági mentés
- nagy teljesítmény
- egyszerű adminisztráció
- nagyfokú automatizmus
- szoros kapcsolat a rendelkezésreállással

Staging (állomásoztatás) VI.

Betöltés lépésenként

- Céltáblánként és célattribútumonként transzformációk leírása,végrehajtása
- Kivételkezelés megvalósítása
- Dimenzió betöltése (kis statikus, kis változó, nagyméretű)
- Ténytáblák töltése
- Összegzések készítése
- Automatizmusok kialakítása

Prezentációs szerver

- Ahol az adatokat a végfelhasználók elérik
- Eleinte nem vált el a részletes adattártól
- Minél magasabb rendelkezésreállás

Közel valósidejű adattárházak

- Eddig: kötegelt (batch) feldolgozás
- Oka: a tipikus igényeket kielégíti és rel. olcsó
- Következmény: jelentős késleltetés az eredményekben
- Cél: a késleltetés csökkentése

Valósidejűség értelmezése I.

- felhasználói szempont: a lekérdezés eredménye álljon gyorsan elő
- műszaki/üzleti szempont: az adatok legyenek minél frissebbek

Az igazi kihívás a kettő együttes teljesítése.

Eredmény: stratégiai, taktikai és operatív döntések támogatása

Valósidejűség értelmezése II.

- Szigorú valós idejű működés – akár folyamatirányításra
- Puha valós idejű működés – döntéstámogatásra (near-time, soft real-time, right-time, on-time)
  - Tipikusan mp-es, ill. nagyobb késleltetések
  - Romlik a hatásfok, ha csökken a késleltetés

Hol lehet rá szükség?

Általában:

- Ahol sok adat alapján gyors és automatizált döntésekre van szükség
- Ahol gyors döntéseket kell hozni példányokra vonatkozóan
- Ahol a jelenlegi működést a historikus viselkedéssel kell összehasonlítani

Tipikus területek:

- korai riasztások („early warning”)
- KPI számítása
- kritikus folyamatok állandó követése
- vételi ajánlat készítés kártyás vásárlásnál
- CRM - gyors ügyfélinformáció
- hitelkártyacsalások felderítése

RTDW definíció

Egy vállalatot átfogó (folytonos és többpontos) adatfolyam históríkus és analitikus része (Haisten, 1999.)

„Real Time is anything that is too fast for your current ETL” (Kimball, 2005.)

Valósidejűség a gyakorlatban (kompromisszumok)

Nyers erő helyett paradigmaváltás

- snapshot-ok helyett változásadatok
- frissítés gyakoriságának korlátozása
- statikus és dinamikus adatok szétválasztása
- lassan változó adatok kezelése továbbra is kötegelten

Komponensek

- Adatstruktúrák (tartalom)
- Megjelenítés
- Interfészek, adatmozgatás
  - ETL helyett CTF (Capture, Transform, Flow)

CTF

- Adatkinyerés (Capture)
  - folyamatosság ($\Leftrightarrow$ kötegelt)
  - teljes kontextus megragadása
  - eseményvezérelt
- Transzformációk (Transform)
  - konverziók, mezők szétválasztása, kódok feloldása
  - összegzés, multidimenziós átalakítás
- Adattovábbítás (Flow)

Adatok kinyerése

- Full snapshot-ok: csak batch esetén
- Megváltozott adatok kinyerése
  - Közvetlen támogatás nincs
    - partícionálás
    - időbélyegek
    - triggerek
  - Közvetlen támogatás a forrásrendszerben
    - pl. CDC (Change Data Capture)

Oracle CDC (Change Data Capture)

- módosított adatok (U, I, D) azonnal elérhetők egy külön táblában
- publisher-subscriber(s)
- minden előfizetőnek saját nézete van a megváltozott adatokra
- Az előfizető kezeli a nézet hosszát és törli belőle az adatokat

Adatok mozgatása

- egyre szorosabb együttműködés az ETL és EAI eszközök között:
  - Acta Works: Java JMS
  - Informatica PowerCenterRT: IBM WebSphere MQ, TIBCO Rendezvous
  - DataStage XE: IBM WebSphere MQ
  - IBM Warehouse Manager: IBM WebSphere MQ

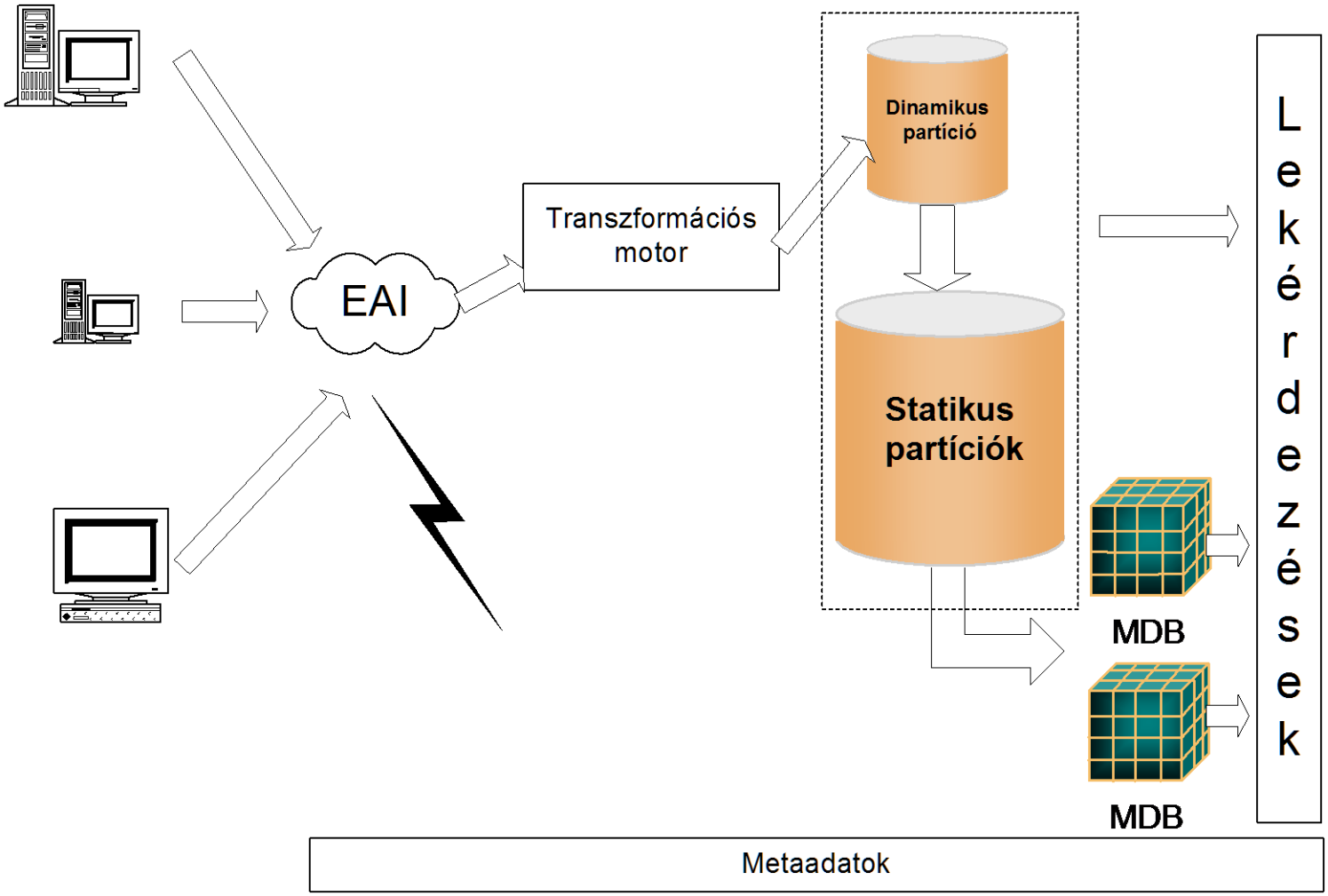
A valósidejű adattár

- Hogyan töltsünk lekérdezés közben?
  - Valós idejű partíció
    - kritikus mérőszámok gyors korrekciójára
    - IMDB
  - Statikus partíciók
    - hagyományos adattárház
    - szakterületi adattárak ettől függenek
    - valós idejű partíció ürítése holtidőben

Összegzések

- Ismert, hogy
  - csökkenti a válaszidőket
  - helyet takarít meg
  - elfedi a részletes adatokat
- szerepe átértékelődik
- trendek követése időérzékeny tevékenységeknél
- inkrementális összegzés

Egy lehetséges felépítés



Analógiák

- Tranzakciós rendszer: (digitális) fénykép a jelen állapotról
- Hagyományos adattárház: (digitális) fényképek sorozata - film
- Valósidejű adattárház: (digitális) film, ahol csak a megváltozott képtartalmat rajzoljuk újra

Rövid történet

- 1998: első publikációk
- 1999: első fejlesztések
- 2001: első CTF eszközök megjelenése
- 2002: Oracle 9iR2 CDC
- 2003: BME-TMIT kísérleti RTDW v1.0
- 2004: Első Oracle alapú ipari implementáció (Euronext)
- 2006: Első Mo.-i ipari alkalmazás a MAVIR-nál (BME-TMIT)