Adatbázisok analitikus környezetben

Adatbázisok elmélete 4. előadás

Gajdos Sándor

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...

- 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
- 2000- valós idejű
- 2010- mobil

Lekérdezések támogatása I.

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

Lekérdezések támogatása I.

Hogyan????

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.

Lekérdezések támogatása IV.

Teljes modell

- A ténytáblák csak dimenziókat, a dimenziók csak tényeket kapcsolnak össze
- Adattárház busz
- Konform dimenziók
  - Definíciója
  - Jelentősége
- Implementációs lehetőségek
  - Relációs
  - Natív multidimenziós
  - OO,...

Lekérdezések támogatása V.

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

Lekérdezések támogatása VI.

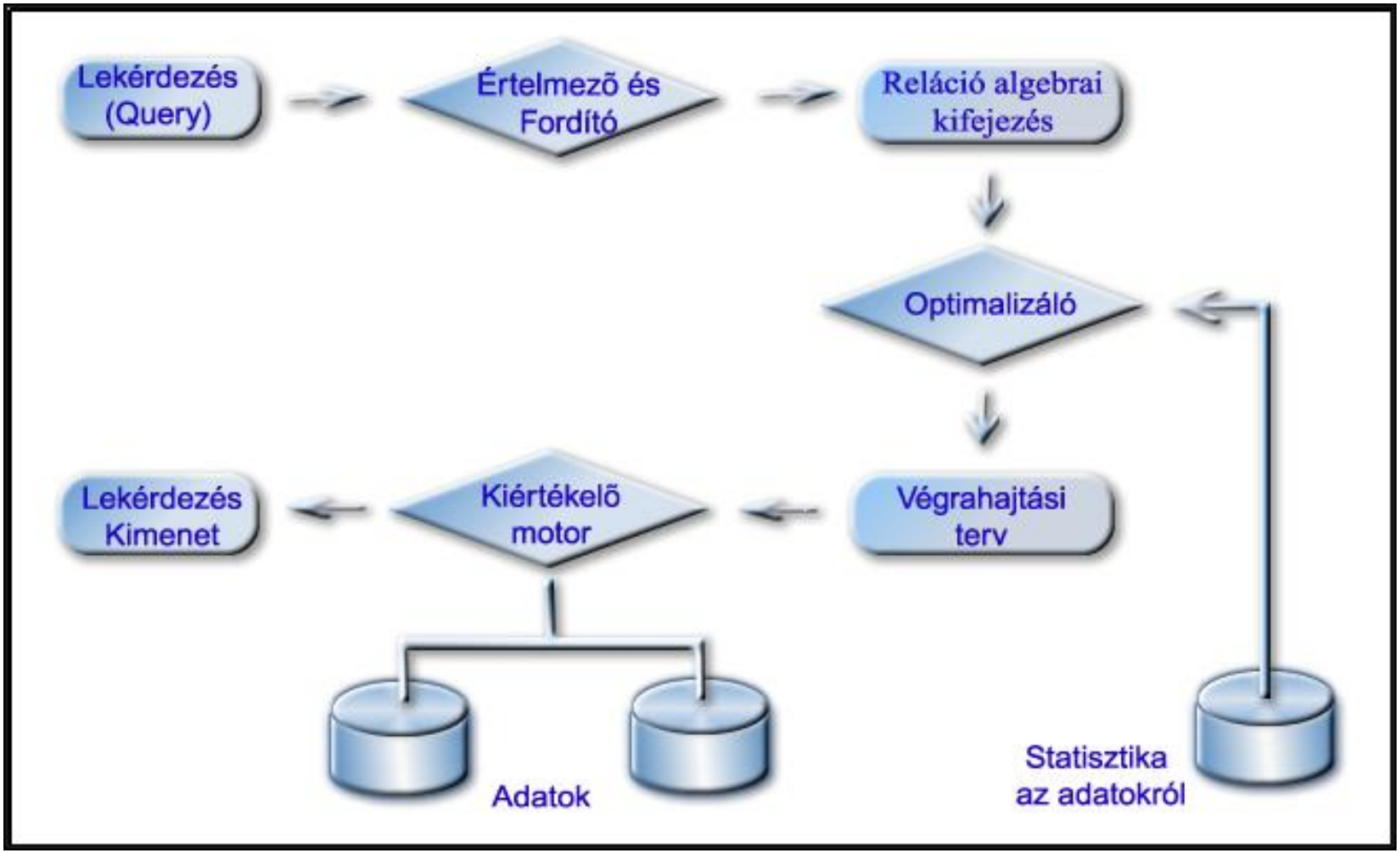
Végfelhasználói hozzáférési módok

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

Lekérdezések támogatása VII. - optimalizálás

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

Optimalizálás - áttekintés



Lekérdezés optimalizálás csillagsémákon

- Lényegében egy illesztés a ténytábla és a dimenziós táblák között
- Dimenziós táblákat sohasem 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 tényrekordokhoz.

Csillagtranszformáció példa

```
SELECT ch.channel_class, c.cust_city, t.calendar_quarter_des
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 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.”

Dimenziós modellezés

- **Dimenziós modellezés előnyei:**
  - lekérdezése könnyen optimalizálható
  - a modell bővítése egyszerű, nem kell átstrukturálni az adatbázist, ha bővül a modell
  - 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 érteni 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

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.
123	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:

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

FIZIKAI TERVEZÉS

1. ld. fizikai adatbázis tervezésről eddig tanultak
2. összegzések 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> 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)	Össze-gzés 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 8i-től)