Adattárházak

Gajdos Sándor, TMIT
2015. ősz

Döntéstámogató rendszerek (DSS: decision support systems)

- Kommunikáció-orientált
- Adat-orientált
- Dokumentáció-orientált
- Tudás-orientált
- Modell-orientált

Kommunikáció-orientált döntéstámogatás

- Abban hiszünk, hogy a kommunikáció hatékonyságának javítása vezet jobb döntésekhez.
- Telefonkonferencia, elektronikus hirdetőtábla, levelezési lista, dokumentum megosztás,...
- Ld. collaborative computing, groupware

Adat-orientált döntéstámogatás

- Abban hiszünk, hogy a sok tényadat alkalmas elemzése vezethet jobb döntésekhez
- Idősoros szemlélet jelentősége (v.ö. trend analízis)
- OLAP (OnLine Analytical Processing), EIS (Executive Information Systems), GIS (Geographic Information Systems), DW (Data Warehouses, adattárházak)

Dokumentáció-orientált döntéstámogatás

- Abban hiszünk, hogy a jó döntéshez szükséges információk már megvannak írott anyagokban, csak meg kell találni őket
- A dokumentum nemcsak papír lehet, hanem video vagy audio is
- Dokumentum-kezelő rendszerek, kulcsszavas keresés, information retieval, AI, fuzzy módszerek

Tudás-orientált döntéstámogatás

- Konkrét személy vagy személyek tudását, gondolkodásmódját másolják le. Képesek konkrét cselekvéseket javasolni
- A szakértelem egy adott területre vonatkozó tudásból áll és bizonyos problémák megoldásának képességéből
- Szakértő rendszerek, intelligens DSS

Modell-orientált döntéstámogatás

- Alkalmas modellek számos lehetséges kimenetel kiszámítását teszik lehetővé
- Statisztikai, pénzügyi modellek, időbeli szimulációk, „what-if” elemzések
- Más néven számítás-orientált döntéstámogatás

Gyakorlati döntéstámogatás

- Pénzfeldobás ☺
- Spreadsheet-döntéstámogatás
- ...
- Kombinált módszerek
- Vannak helyzetek, amikor „minden pénzt és módszert megér” a releváns információk megszerzése (ld. fegyverzetcsökkentési tárgyalások során látnokok, parafenomének alkalmazása)

Adatorientált döntéstámogatás fejlődése

- 60-as évek: batch riportok, nagy köteg nyomtatott papír, kérések lassú kiszolgálása, további adatfeldolgozás nehézkes
- 70-es évek: terminál alapú rendszerek, gyenge felh. interfész, adatforrások gyenge integrációja
- 80-as évek: PC alapú hozzáférés, EIS, GUI megjelenése, de: zavaros/inkonzisztens adatok, kevés historikus adat, bonyolult adatstruktúrák
- 90-es évek: adattárházak, amelyek a korábbi problémák legtöbbjét megoldják, trendanalízis, desktop OLAP, web-es rendszerek
- 2000 után: valós idejű döntéstámogatás, mobil interfészek

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, QED/Wiley, 1991.

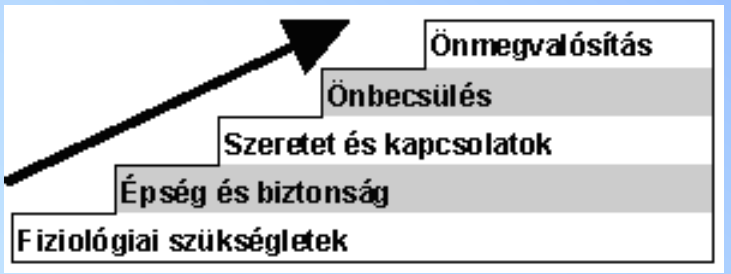
Üzleti intelligencia (BI)

Új 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)

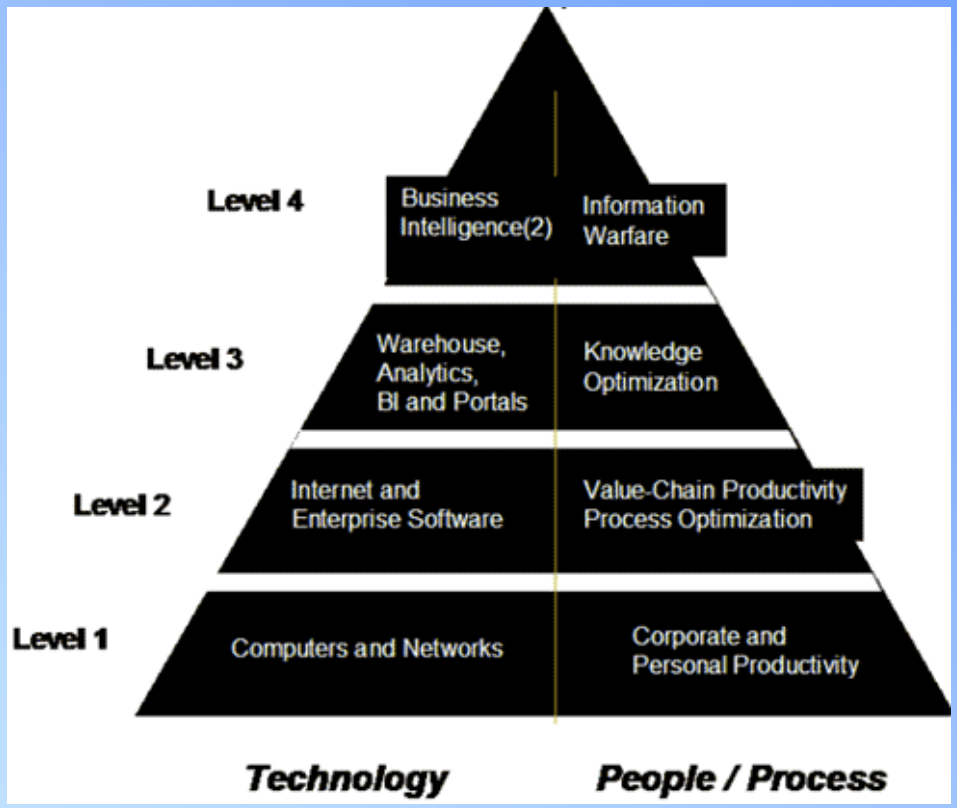
Avagy: mi „működteti” az embereket



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 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?



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

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

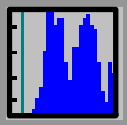
ROLAP/OLAP

4GL eszközök

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

M I D D L E W A R E

Metaadat

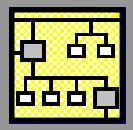
Adattárház-adminisztráció

Web böngésző
Adatbányászat

ROLAP/OLAP

4GL eszközök

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

Metaadat

Szakterületi adattár

Pénzügy

Metaadat

Szakterületi adattár

Marketing

Értékesítés

Web böngésző

Adatbányászat

ROLAP/OLAP

4GL eszközök

EIS/DSS

Riportok

Hozzáférés (API-k, Middleware)

Relációs lekérdezések optimalizálása dimenziós struktúrákon

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

Lekérdezés optimalizálás csillagsémákon

- Lényegében egy illesztés a ténytábla és a dimenziós táblák között
- Dimenziós táblákat sohasem join-olunk
- A lehetőség automatikus felismerése
- Hópihe 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 (‘2006-Q1',‘2006-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(‘2006-Q1',‘2006-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 2006. első negyedévi tényrekordokat
- time_id bitmap azonosítja a 2006. 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 reguláris 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...

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)