Miskolci Egyetem
Gépészmérnöki és Informatikai Kar
Informatikai Intézet
Alkalmazott Informatikai Intézeti Tanszék

Nappali tagozat
Tanév: 2018/2019. II. félév

Tantárgyi tematika

Adattárház rendszerek c. tárgyhoz gazdasági informatikus szak nappali tagozat, BSc szint

Tárgykód:

GEIAL32E-B

Tárgyfelelős előadó és gyakorlatvezető:

Dr. Sasvári Péter, egyetemi docens

A tárgy lezárásának módja:

aláírás és gyakorlati jegy

A tantárgy oktatásának időterve:

A tárgy egy féléves. A tárgy óraszáma: 2 óra előadás, 2 óra gyakorlat hetenként.

Hét	Tananyag
1	OLTP és OLAP rendszerek összehasonlítása, jellemzőik összefoglalása.
2	Egyváltozós adatelemzési módszerek.
3	Többváltozós adatelemzési módszerek.
4	Statisztikai tesztek.
5	Adatelemzés SQL és R statisztikai függvényekkel.
6	Adattárház rendszerek jellemzése és tervezése: adatbetöltés és ETL folyamatok.
7	A multidimenzionális adatmodell elemei. MD modell tervezésének lépései.
8	Az MD modell műveleti része, adatok lekérdezése.
9	Adattárház termékek: MS SQLServer Analysis Server és az MDX nyelv.
10	Adattárház termékek: Oracle DB with OLAP Option.
11	Esettanulmány: egy ingyenesen elérhető adattárház rendszer bemutatása.
12	Adatbányász módszerek az adatelemzésben.

Az évközi ellenőrzés:

Részvétel az előadások és gyakorlatok legalább 50%-án. 2 önálló feladat elkészítése.
Zárthelyi dolgozat legalább 50%-os teljesítése. A gyakorlati jegy kiszámítása:

0%-50%: elégtelen
50%-62%: elégséges
62%-75%: közepes
75%-88%: jó
88%-100%: jeles

Számonkérés módja:

A tárgy az eredményes évközi munkát elismerő aláírással, majd gyakorlati jeggyel zárul. A ZH értékelésére a félévközi dolgozat szabályai vonatkoznak.

Általános rendelkezések

Az ME SzMSz III. kötet 96§ alapján a tárgyakhoz kapcsolódó valamennyi számonkérési alkalomnál a nem engedélyezett segédeszközök használata (puskázás) vagy más munkájának sajátként történő feltüntetése (plagizálás) fegyelmi vétségnek minősül, mely tanulmányi szankciókat vagy fegyelmi eljárást von maga után.

Tanulmányi szankció az évközi számonkéréseknél a számonkérés sikertelen minősítése. A számonkérés ilyen esetekben nem pótolható.
A puskázás és/vagy plagizálás tényét a tanszék a hallgató tanulmányi ideje alatt nyilvántartja, és ismételt előfordulás esetén a ME SzMSz III. kötet 96§ által előírt fegyelmi eljárást kezdeményez.

Felhasznált irodalom:

Kötelező irodalom:

1. Kiadott előadás anyagok (Kovács László és Sasvári Péter diái)
2. Kovács László: Adatelemzési technikák és eszközök, Nemzeti Tankönyvkiadó
3. W. H. Inmon (2005): Building the Data Warehouse, Wiley, p. 576.

Ajánlott irodalom:

4. Ralph Kimball (2013): The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, Wiley, p. 600

Miskolc, 2019. február 1.

Dr. Sasvári Péter
tárgyjegyző

MISKOLCI EGYETEM
Gépészmérnöki és Informatikai Kar
Általános Informatikai Intézeti Tanszék

Név: .....
Neptun kód: .....

Zárthelyi dolgozat

Adattárház rendszerek
című tárgyból
……

Ponthatárok: 0-15 pont: elégtelen, 16-19 pont: elégséges, 20-23: közepes, 24-27 pont: jó, 28-30 pont: jeles.

1. Az adattárház koncepció (Az adattárház (data warehouse) fogalma, Néhány adattárház definíció, Inmon definíciója, A "Data Warehousing" fogalma, A "Business Intelligence" (BI) fogalma)

   /10 pont/
2. OLTP és OLAP rendszerek (OLTP rendszerek, Új követelmények - OLAP rendszerek, OLTP - OLAP rendszerek összehasonlítása)

   /10 pont/
3. MOLAP architektúrák (Adatstruktúrák, A többdimenziós tömb tárolás, Ritka mátrix kezelés, A multidimenzionális tárolás korlátai, MOLAP termékek)

   /10 pont/

MISKOLCI EGYETEM
Gépészmérnöki és Informatikai Kar
Általános Informatikai Intézeti Tanszék

Név: .....
Neptun kód: .....

Zárthelyi dolgozat

Adattárház rendszerek
című tárgyból

.....

Ponthatárok: 0-15 pont: elégtelen, 16-19 pont: elégséges, 20-23: közepes, 24-27 pont: jó, 28-30 pont: jeles.

1. Az adattárház koncepció (Az adattárház (data warehouse) fogalma, Néhány adattárház definíció, Inmon definíciója, A "Data Warehousing" fogalma, A "Business Intelligence" (BI) fogalma)

Próbálkozzunk meg az adattárház definíciójával. Bár mára a kifejezés értelmezésében viszonylag nagy az egyetértés, kisebb értelmezésbeli, nézőpontbeli különbségek még mindig jelen vannak. Nézzük először, hogy az adattárház elmélet két evangelistájának mondott úttörőnk Ralph Kimball és Bill Inmon hogyan is fogja meg a fogalmat: (Az idézett könyvek (egyébként egyelőre kiállták az idő próbáját, és még ma is a két alapműként tartják számon a témában - 1996 óta, ami persze nem nagy idő.)
Néhány adattárház definíció
Idézet Ralph Kimballtól: Data Warehouse: "The conglomeration of an organization's data warehouse staging and presentation areas, where operational data is specifically structured for query and analysis performance and ease-of-use." [2] Az adattárház fogalma itt tehát egy adott szervezet azon adatgyűjtő és szolgáltató részeit foglalja magában, ahol a működési adatokat újrastrukturálják riportkészítési, jó teljesítményű és egyszerűen kezelhető elemzésekhez. Kimball ezen definícióját főleg azért szokták kedvelni és idézni, mert sok mindent nem határoz meg, pl. az adattárház nem feltétlenül döntéstámogatási célú.
Ebből átmenetként foghatjuk fel a következő változatot (már nem Kimballtól), mely az adattárházat technológiák gyűjteményének definiálja: "A data warehouse is a collection of technologies aimed at enabling the knowledge worker (executive, manager, analyst) to make better and faster decisions."
A Business Intelligence (BI), üzleti intelligencia fogalámát Howard Dresner (Gartner Group) definiálta 1989-ben, azóta általánosan elfogadott fogalommá vált. Olyan módszerek, fogalmak halmazát jelenti, melyek a döntéshozás folyamatát javítják adatok és ún. tényalapú rendszerek használatával. A "tényalapú rendszer" a következő alrendszereket foglalja magába:

- Vezetői információs rendszerek (Executive Information Systems)
- Döntéstámogató rendszerek (Decision Support Systems, DSS)
- Vállalati információs rendszerek (Enterprise Information Systems)
- Online Analitical Processing (OLAP)
- Adat- és szöveg-bányászat
- Adatvizualizáció
- Geográfiai információs rendszerek (Geographic Information Systems, GIS)

Az üzleti intelligencia fogalmát gyakran említik együtt az adattárházak fogalmával, mivel az lefedheti ezen részrendszereket, valamint kiszolgálhat ilyen rendszereket. Leginkább azonban tekinthetjük az adattárház megoldásokat az üzleti intelligencia megoldások egy szeletének.

/10 pont/

2. OLTP és OLAP rendszerek (OLTP rendszerek, Új követelmények - OLAP rendszerek, OLTP - OLAP rendszerek összehasonlítása)

Tulajdonságok	**OLTP**	**OLAP**
Orientáció	tranzakciók	adatanalízis
Felhasználó	vállalat adminisztációt végző alkalmazottai	döntéshozók és őket információval támogató alkalmazottak
Feladat	napi folyamatok követése	döntéstámogatás, hosszútávú információgyűjtés és szolgáltatás
Adatbázis tervezése	Egyed-Kapcsolat modell, alkalmazás orientált	tárgy-orientált, csillagséma
Adatok	aktuális, up-to-date	történeti adatok, időben archiválva
Aggregált adatok	nem jellemző; részletes felbontás	felösszegzett, egyesített adatok
Adatok nézete	részletezett, relációs	felösszegzett, multidimenzionális
Felhasználók hozzáférése	olvasás/írás	legtöbbször olvasás, adattárház adatait nem módosítják
Hangsúly	adatbevitelen	információkinyerésen
Feldolgozandó rekordszám	tizes nagyságrendű rekord alkalmanként	akár milliós rekordszám
Felhasználók száma	viszonylag sok	kevés, közép és felsővezetők ált.
Prioritás	állandó rendelkezésre állás és megbízhatóság	rugalmasság, felhasználói önállóság

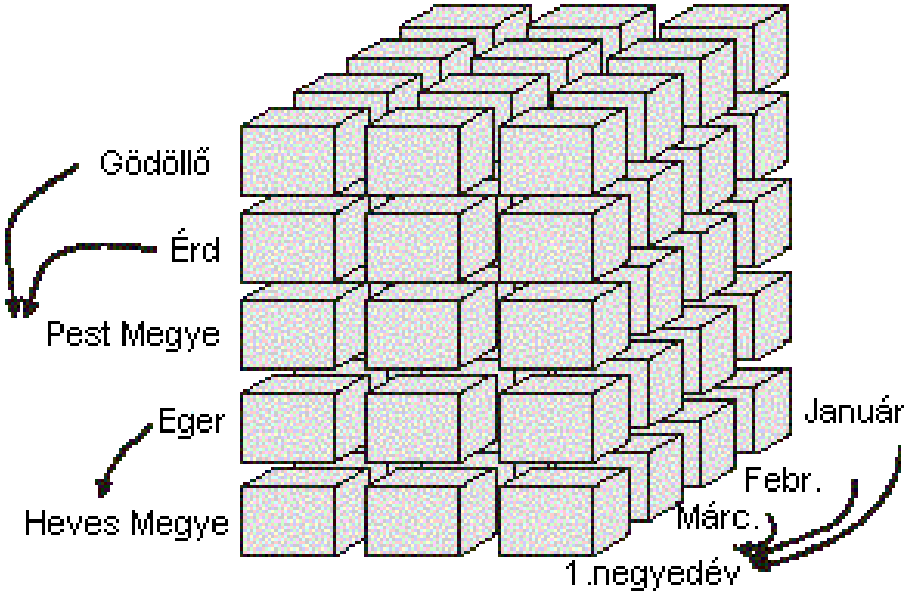
/10 pont/

3. MOLAP architektúrák (Adatstruktúrák, A többdimenziós tömb tárolás, Ritka mátrix kezelés, A multidimenzionális tárolás korlátai, MOLAP termékek)

Multidimenzionális OLAP alkalmazások esetében adatainkat speciális multidimenzionális struktúrában tároljuk.

MOLAP esetén külön kezeljük a dimenziók adatait és a tényadatokat.
Dimenziók: véges, rendezett lista a dimenziók elemeiről. Fontos, hogy a dimenzióelemek listája jól rendezett legyen.
Adatkocka: Egy n dimenziós kocka esetén az adatok egy n-dimenziós térben helyezkednek el, annak egy zárt részhalmazában a dimenzióértékek végessége miatt. Mivel a dimenziók diszkrét értékűek is, a dimenzióértékek a térben cellákat jelölnek ki. Ezek a cellák tartalmazzák a mutatószámokat. A dimenzióértékek által lehatárolt térrészt, azaz a kockát speciális indexstruktúrával látjuk el, aképp, hogy a a kocka minden egyes cellája be legyen indexelve. (Az index általában adott attribútumú sorok

gyors elérését szolgáló struktúra.) A kockát eltároljuk a háttértárunkon, felépítjük az indexet, ami általában elfér a memóriában is, és ezt használva az adatok elérése jelentősen gyorsabb, mint egy sima relációs adatbázisban. Előny továbbá, hogy a struktúra jól illeszkedik a koncepcionális modellhez, fordítás nélkül alkalmazhatóak rá az OLAP műveletek.
Dimenzió hierarchiák: A dimenziók hierarchikus felépítését úgy kezeljük, hogy a hierarchia csomópontjait elhelyezzük a dimenzióelemek között és összesített adatokat rendelünk hozzá.
Aggregált adatok: Aggregált, felösszegzett adatok kezelése MOLAP architektúra esetén akkor sem jár elfogadhatatlan válaszidővel, ha külön nem foglalkozunk összegek tárolásával, mégpedig a gyors adatelérés miatt. Ezen túlmenően előre is definiálhatunk a kockákban aggregációs szinteket, hasonlóképp mint a hierarchiák esetében, ekkor az összegek beépülnek a kockába. Fontos megjegyezni a rendszerek azon hiányosságát, hogy nem lehet aggregált adatokat tárolni dimenziók nem teljes értékkészletével. Például nem megoldható, hogy aggregált adatokat tároljunk csak Gödöllő és Eger adataival. Igaz ugyanakkor az is, hogy a felhasználói lekérdezések ritkán ilyen jellegűek.
A dimenzió hierarchiák és az aggregált adatok esetén nyújtott megoldások tulajdonképp az adatok rendundás tárolásához vezetnek. A teljesítmény növelésére használt redundáns tárolás nem csak a MOLAP megoldásokra, hanem általánosan jellemző az adattárházakra, a tárhely-takarékosságról áthelyeződött a hangsúly a kiértékelés gyorsaságára.
Attribútumok: Ebben az esetben attribútum alatt a dimenzió jellemzőit értjük. Például, tekintve a "Vevő" dimenziót, ennek attribútumai lehetnek a vevő címe, számlaszáma, kategóriája, szöveges leírása, és így tovább.
Virtuális adatkocka: olyan kocka, amely levezetett, számolt adatokat tartalmaz, melyek konkrétan nem szerepelnek fizikailag tárolt kockákban. Ilyen lehet például egy a nyereségből épített kocka, vagy egy tény-terv összehasonlító százalékos eltérést mutató kocka.



A többdimenziós tömb tárolás

A kocka indexeléséhez szokás olyan indexstruktúrát használni, ahol a kocka celláit valamilyen adott algoritmussal sorbarendezzük, majd az indexek sora ennek a sorbarendezésnek felel meg. Ennek legegyszerűbb módja, ha a kocka adott (x1, x2, .. xn) koordinátájú pontjához a kordinátákból alkotott x1 + (x2-1)*|{1.dimenzió elemszáma}|+...+(xn-1)*|{(n-1)..dimenzió elemszáma}| sorszámú indexet rendeljük.

Az indexeket és magát a fizikai tárolóhelyet a háttértárakon előre elkészítjük, így az adatok anélkül írhatóak, hogy az indexstruktúrát módosítani kellene. Ugyanakkor mivel a rendezés egyértelműen azonosítja az adott cellát, nem kell az adatokkal együtt a kulcsokat is eltárolni.

Ritka mátrix kezelés

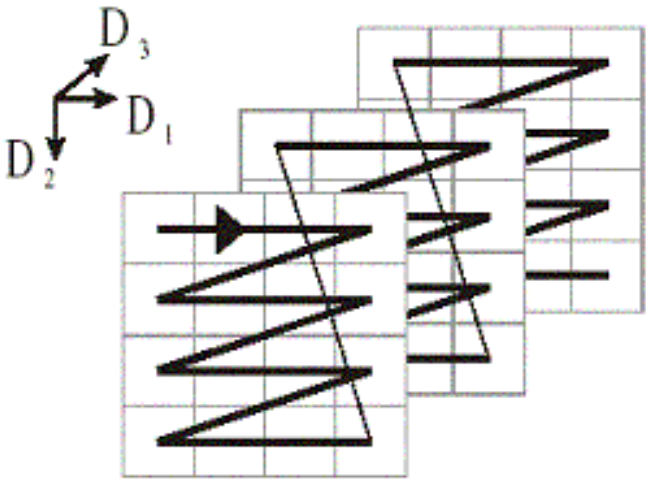
Amennyiben az adatmátrixunk ritka, tehát az adatok a kockán belül szétszórtan helyezkednek el, a kocka a hasznos adat mennyiségéhez képest nagy területet foglalhat el. Ennek oka, hogy az előre felépített indexstruktúra miatt a háttértáron előre helyet kell foglalni a kocka egészének. Sok dimenzió és nagy kiterjedésű dimenziók esetén ez akár oda is vezethet, hogy az adatbázis használhatatlanul naggyá válik. (A konkrét adatok egy becsléséhez ld. 6.ábra.) A ritka mátrix probléma kezelésére egyes multidimenzionális adatbáziskezelők tartalmaznak ún. ritka mátrix algoritmust, amely a kocka szerkezetéből megpróbálja a nem használt részeket kiszűrni, és a nekik fenntartott helyet felszabadítani, így elkerülve a mátrix kezelhetetlen naggyá válását.

A multidimenzionális tárolás korlátai

A már említett ritka mátrix probléma mellett meg kell említenünk még, hogy a strukturális változtatások ebben a modellben rendkívül költségesek. Emellett ezek a rendszerek általában nehezen skálázhatók, nincs általánosan elfogadott szabványuk, minden gyártó saját utakon jár.

MOLAP termékek

A MOLAP termékek széles skálán mozognak az asztali, néhány 10 Mb mennyiségű adat kezelésére alkalmas alkalmazásoktól kezdve a vállalati "high end" szoftverekig. Az első csoportba tartoznak pl. a Cognos PowerPlay-e, az Andyne PaBLO-ja és a Business Objects Mercury-ja. Az utóbbi kategóriában a Kenan Acumulata ES-e, az Oracle Express családja, a Planning Sciences Gentium-a és a Holistic System Holos-a olyan termékek, melyek nem csupán a multidimenzionális adattárolást, hanem rengeteg más kapcsolódó feladatot is megoldanak. Tisztán multidimenzionális adatbázis motorok az Arbor Essbase-e, a D&B/Pilot Ship Servere és a TM/1 a Sinper-től.



/10 pont/