

A kérdéseket magyarul tettük fel, de magyar és angol választ is elfogadunk.

A. Adja meg a táblázatban, hogy a következő állítások igazak vagy hamisak!

1. Egy normalizált séma pontosabban leírja a valóság modellezni kívánt részét, mint egy nem (kevésbé) normalizált.
2. A sémanormalizálás során több relációt (táblát) kapunk, mint eredetileg volt.
3. A veszteséges dekompozíció során a kapott relációkat természetes összekapcsolással (natural join) összekapcsolva a kiindulási relációhoz képest több sort kapunk vissza.
4. A bináris fa mélysége a fában levő elemek számának 2-es alapú logaritmusával arányos.
5. A levelek száma határozza meg egy bináris keresőfában a műveletek (pl. keresés) költségét.
6. Egy B+ fa alapú adatbázis index csomópontjának méretére egy szokásos méret a 8kB.
7. Egy T(A,B,C,D,E) táblán létrehozott I(A,B,C) indexet (hatékonyan) lehet használni B-re és C-re vonatkozó, és-sel (AND) összekapcsolt lekérdezési feltételek esetén.
8. Nagyon szelektív lekérdezés esetén igazán hasznosak az indexek.
9. Egy SQL lekérdezést sokféleképpen tud végrehajtani az adatbáziskezelő.
10. Az adatbáziskezelő csak a lekérdezést, az SQL parancsot elemzi, abból egyértelmű, hogyan kell végrehajtani.
11. A tranzakciókezelés kizárja, hogy több felhasználó is használjon egy adatbázist.
12. Ha eredményesen lezárult egy tranzakció, áramszünet esetén sem vesznek már el az adatok.

Kérdés	Igaz	Hamis
1		X
2	X	
3	X	
4	X	
5		X
6	X	
7		X
8	X	
9	X	
10		X
11		X
12	X	

13. A tranzakciók, bármilyen izolációs szint is van beállítva, egymástól tökéletesen függetlenül hajtódnak végre.
14. Az XML egy, a gépi feldolgozásra optimalizált, bináris formátum.
15. Az XML formátumban az elemek egymásba is ágyazhatóak, egy XML dokumentum logikailag egy fának feleltethető meg.
16. Egy XML dokumentumban egy elem gyerekeinek a száma a százaz nagyságrendben van.
17. Egy jól formázott (well-formed) XML dokumentum mindenképpen érvényes (valid) is.
18. Egy érvényes (valid) XML dokumentumhoz mindenképpen tartozik egy sémaleírás (DTD vagy XML séma).
19. Egy JSON (JavaScript Object Notation) formátumú fájl kisebb, mint a megfelelő XML dokumentum.
20. A csillagséma az OLAP rendszerek jellegzetesége.
21. A ténytábla tartalmazza a tény (fact) attribútumokat, melyek jellegzetesen numerikusak.
22. Az információvisszakeresés pontossága (precision) 100%, ha az eredményhalmaz valamennyi releváns dokumentumot tartalmazza.
23. A sorbarendezett eredményhalmazt szolgáltató (ú.n. vektor-tér) információvisszakereső modell egy szót annál nagyobb súllyal vesz figyelembe, minél több dokumentumban szerepel az.
24. A sorbarendezett eredményhalmazt szolgáltató (ú.n. vektor-tér) információvisszakereső modellben a dokumentumvektorok hossza meghatározó a relevancia meghatározásánál.
25. A sorbarendezett eredményhalmazt szolgáltató (pl. ú.n. vektor-tér) információvisszakereső modell minőségét a pontosság-fedés görbével (precision-recall curve) szokták jellemezni.

Kérdés	Igaz	Hamis
13		X
14		X
15	X	
16		X
17		X
18	X	
19	X	
20	X	
21	X	
22		X
23		X
24		X
25	X	

B. Egészítse ki a szöveget a megfelelő kifejezésekkel, ill. válassza ki a megfelelő választ bekarikázással. Az előbbi esetben, ahol kiegészítésre van szükség, használhatja az angol vagy a magyar kifejezést is. (56 pont)

A három Armstrong axioma és azok neve! (3 pont)

1. Reflexivitás - ha Y részhalmaza X , akkor $X \rightarrow Y$.
2. Bővítés - ha $X \rightarrow Y$, akkor $X \cup Z \rightarrow Y \cup Z$.
3. Transzitivitás - ha $X \rightarrow Y$ és $Y \rightarrow Z$, akkor $X \rightarrow Z$.

Az indexek a háromrétegű adatbázisarchitektúra legfelső/középső/alsó szintjére tartoznak, melynek neve fizikai szint. Az adatbáziskezelőkben használt indexek többnyire B+ fa (adatstruktúra; minél pontosabban) alapúak. Az adatbáziskezelő az adatokat a merevlemezen blokkokba rendezve tárolja, ezek szokásos mérete 16 B/1kB/8kB/256kB/1MB. (5p)

Az indexben adatok csak a fa levelei-ben találhatóak. A levelek az értékeken kívül pointereket tartalmaznak. A fában a csomópontok gyerekeinek száma jellemzően 10/500/10000/100000, így a fa mélysége jellemzően nagyon kicsi, szokásosan 3-5/10-15/20-50/>100. (5p)

A költség alapú lekérdezés optimalizálás célja, hogy a deklaratív (jellegű) SQL nyelvben megadott feladat relatív optimális (jellegű) végrehajtását meghatározza. A lekérdezés egy helyesség-ellenőrzés után egy belső formába, az ún. operátorfá-ba transzformálódik. Ezen különböző transzformációkat lehet végrehajtani, melyek során a(z) végeredmény nem változik, de a végrehajtás költsége igen. Az ökölszabály, hogy a közbülső eredmények kicsik legyenek. Például a szelekció és a projekció műveleteket érdemes lehetőség szerint korán, a keresztszorzás és a összekapcsolás műveleteket lehetőség szerint későn végrehajtani. A költségek becslésének alapja az adatokról vezetett statisztika, például a táblákban levő rekordok száma. (14p)

A tranzakciókezelés két kihívásra ad megoldást: (1).külső hibák tolerálása (robosztusság)
(2) konkurrens hozzáférés..... Egy tranzakció jellemzően több utasításból áll. A tranzakciókezeléssel kapcsolatos szokásos követelmények:

- (1) atomicitás.....
- (2) konzisztencia.....
- (3) izoláció.....
- (4) tartósság.....

A fenti követelmények teljes betartása a sebesség csökkenéséhez vezet. Ezért a izoláció..... követelményt szokták lazítani. Jelenségek, melyeket a szabvány bizonyos beállítások között engedélyez:

- (1) phantom read.....,
- (2) non-repeatable read.....,
- (3) dirty read.....,

(10 pont)

jelentési és

Az adattárház egy olyan adatbázis, amit elemzési..... célra használnak. Az adattárházak aktuális és historikus adatokat is tárolnak. Egy adattárházba jellemzően sok forrásból töltenek be adatot, sok adatforrást a source data system.....néven (vagy betűszóval) szoktak hívni. A betöltőfolyamatok neve extract, transform, load... , rövidítve: ETL, ezek az adatok harmonizálásán, historizálásán kívül az adatok tisztítását, az adathibák kezelését is elvégzik. Az üzleti adattárházak jellegzetes sémájának (sémarészletének) neve: csillagséma..... Az ehhez tartozó lekérdező-alkalmazás: Data Cube..... (6 pont)

A processzorok órajel-frekvenciája..... a korábbi dinamikus növekedés után az elmúlt években stagnál. Ugyanakkor a processzormagok.....száma növekedésnek indult. A memória ára folyamatosan esik, jelenlegi szerverek 4-16 GB méretű (nagyságrend) memóriával érhetőek el. Az új hardver fejlődési trendek lehetővé teszik új architektúrájú (relációs) adatbáziskezelők építését, ezek egy nevezetes képviselője a SAP.....(cég ill. termék) (4 pont)

Az információ visszakeresés (information retrieval) nem struktúrált adatokban való kereséssel foglalkozik, a felhasználó információs igénye szerinti relevanciát figyelembe véve. A legjellemzőbb struktúrátlan adat a
szöveges dokumentum....., további példák: (1)
.....kép....., (2)audio/video.....

A bináris információvisszakeresés eredményhalmaza általában eltér a releváns dokumentumok halmazától. A bináris információvisszakeresés minőségének két elterjedt jelzőszáma: (1)0 - nem szerepel....., (2)
.....1 - szerepel.....

A bináris információvisszakeresés fő nehézsége, hogy a megfogalmazott kérdés/lekérdezés vagy túl szűk, nem kapunk eredményt, vagy túl tál, és használhatatlanul nagy lesz az eredményhalmaz. A sorbarendezéses információvisszakeresés (ranked retrieval) erre ad megoldást.

A sorbarendezéses információvisszakeresésnél (az ú.n. vektor-tér modellben, vector-space model/ranking) a dokumentumokat egy-egy vektor írja le. A dimenziókat a „szavak” adják. A dokumentumvektorok értékét egy-egy dimenzióban két dolog határozza meg:

- (1)..... kifejezés-gyakoriság (term frequency)
.
(2)..... dokumentum-gyakoriság (document frequency)
.

Egy lekérdezést ebben a modellben egy vektor... reprezentálja. A lekérdezés és a dokumentumvektorok közelségét.... hasonlítjuk össze, ez alapján rangsoroljuk a dokumentumokat. (9pont) (szögét)