Quizs 1-6 & Caculate

1 Which of the following about the MongoDB query language is true?

(a) db.collection.find({“state": "TX", “pop” : 20000}) returns all document with key/value pair ("state", "TX") OR ("pop", 20000)

(b) db.collection.find({“state” : null}) will only returns document that does not have a "state" key

A. (a) and (b)

B. Neither (a) nor (b)

C. (b)

D. (a)

2 Consider modeling two entities ("Student" and "Faculty", each of them having multiple properties), and a relationship between them ("Advise"). Suppose we want to embed Faculty documents as part of the Student document (but not vice versa). Under what condition will this be a reasonable choice (as if without causing duplication and potential inconsistencies)?

(You can also assume these are the only entity/relationships that needs to be modeled)

(a) if Advise is a 1-1 relationship

(b) if Advise is a 1-n relationship (with 1 student having many advisers, but each faculty having only 1 advisee

(c) If Advise is a 1-n relationship (with 1 faculty having many advisers, but each student having only 1 advisor

(d) If Advise is a m-n relationship (with m, n > 1)

3 The uniqueness constraints in Cypher refer to:

A. Specify no two nodes can have the same number of properties

B. Specify no two nodes of the specifed label can be joined by an edge

C. Specify all nodes must have distinct labels

D. Specify all nodes that have a certain label must have distinct values in the property that is specified

4 For graph model databases, which of the following is/are correct?

(a) Subgraph pattern matching is not often implemented because it is NP-complete

(b) For path queries, one can restrict the types of edges for that path

Correct Answer

A. (a) and (b)

B. Neither (a) nor (b)

C. (b)

D. (a)

5 Which of the following about Solid State Drive (SSD) is correct?

(a) Seek time for a SSD is the same as rotational latency

(b) One cannot rewrite on the same spot of the SSD drive without erasing a larger block first

(c) There is not much difference in reading sequentially and randomly in an SSD

A. (b) and (c)

B. (c)

C. (a), (b) and (c)

D. (a) and (b)

6 Which of the following about RAID-1 is correct?

(a) Writing is always going to be at least twice at slow than RAID-0

(b) RAID-1 should be more reliable than RAID-0

(c) RAID-1 use parity bits

A. (c)

B. (a) and (b)

C. (b)

D. (a), (b) and (c)

7 Which is the following is an advantage of multi-table clustering file organization?

A. Joins involving those tables can be potentially speed up

B. Selection for one of the tables can be speed up

C. All of the other answers are correct

D. Projection for each of the tables can be speed up

8 Which of the following about external merge sort is correct?

(a) The first step of merge sort is to create list of one tuple each

(b) For external merge sort, one can merge more than 2 lists at a time

(c) With a file of N pages, each iteration takes 2N pages of read+write

A. (b) and (c)

B. (a), (b) and (c)

C. (a) and (b)

D. (c)

9 Which of the following about selection queries using indices are correct?

(a) A secondary index should always be used

(b) One cannot know whether a secondary index should be used without knowing how many tuples are going to be returned

(c) If the attribute that is used to build a secondary index is distinct, a query such as "SELECT * FROM A where A.att = 1" should (nearly) always benefit from a secondary index.

A. (b) and (c)

B. (a), (b) and (c)

C. (c)

D. (a) and (b)

10 Which of the following about non-clustering index is correct?

(a) A non-clustering index cannot be sparse

(b) A non-clustering index cannot only be used on an integer attribute

A. Neither (a) nor (b)

B. (a)

C. (a) and (b)

D. (b)

11 Which of the following statements about B+-tree is correct?

(a) Every node of the B+-tree has to be at least half full

(b) All data in a B+-tree are stored in internal nodes

A. Neither (a) nor (b)

B. (a) and (b)

C. (a)

D. (b)

12 Which of the following about the B-tree is correct?

(a) All leaf nodes are at the same level

(b) All elements that are stored are evenly distributed among the leaf nodes (to the best extent possible)

A. (a) and (b)

B. (b)

C. (a)

D. Neither (a) nor (b)

13 What do we mean by decorating a parse tree?

A. Deciding the algorithm to be used for each operation

B. Ordering the operations in the query

C. All of the other answers are correct

D. Translating the SQL statement into a set of relational algebra operation

14 For level k in linear hashing, how many buckets need to be split before we enter level k+1?

A. 2^k

B. 2^{k+1}

C. k

D. $2k$

15 In linear hashing, whenever a bucket overflows when an iterm is inserted to it, the next bucket to be splited is:

A. The bucket that just has the item inserted

B. Bucket 0

C. Pre-determined, unrelated to which bucket is full

D. The current bucket that has the most overflow buckets

16 Consider we have a table A that has attributes A.a, A.b, A.c, A.d. with non-clustering indices on attributes A.a, A.b, A.c separately. Consider the following query

```
SELECT *
FROM A
where <condition>
```

For which condition below we can use the "intersection of identifiers" option to execute the query? (Notice I do not claim that the option will be the best, just a feasible one).

(a) condition is : A.a = 10 and A.b > 10 and A.c = 5

(b) condition is : A.a = 10 or A.b > 10 or A.d = 5

(c) condition is : A.a = 10 and A.b = 10

(d) condition is : A.a = 10 or A.b = 10

A. (a) and (c) and (d)

B. (a) and (c)

C. (c) and (d)

D. (a), (b), (c) and (d)

17 Consider set operations, which of the following statement(s) is/are correct?

(a) Union corresponds to joins which match every attributes

(b) Intersection corresponds to joins which match every attributes

A. (b)

B. (a)

C. Neither (a) nor (b)

D. (a) and (b)

18 One advantage of hash-join over sort-merge is:

A. Hash-joins runs in linear time, but sort-merge runs in O(n^2) time

B. Hash-join ordered the tuples based on the join condition, sort-merge does not

C. There is no advantage at all

D. The number of iteration is (potentially) bounded by the size of the smaller table in hash-join, but is definitely bounded by the larger table in sort-merge

19 Which of the following about the parse tree notation of a query is correct?

(a) Each node can have at most 2 children

(b) The leaves corresponds to tables

(c) For nested loop, the right child is the inner loop.

A. (b) and (c)

B. (a) and (b)

C. (b)

D. (a), (b) and (c)

20 Which of the following about pipeline/materialization is correct?

(a) The inner loop of a nested loop join need to be materialized

(b) An advantage of the left-deep tree is that pipelining is possible along the whole left-path of the tree.

A. (a)

B. (a) and (b)

C. Neither (a) nor (b)

D. (b)

21 Consider the first part of a multi-table query r1⋈ r2 (with the join condition r1.a = r2.a). Which of the following will make sort-merge a potential best-case scenario?

(a) At the end the tuples needed to be ordered by r1.a

(b) There is a subsequent join with the condition r1.b - r3.b

(c) There is a subsequent join with the condition r1.a - r3.c

22 Consider joining to table A and B with the condition A.a = B.b, and A.a is the primary key of A. Which of the following statement(s) is/are correct?-

(a) If A.a is the primary key of A, then the result cannot have more tuples than the number of tuples in A

(b) If A.a is the primary key of A, then the result cannot have more tuples than the number of tuples in B

23 Which of the following statements is/are correct?

(a) Equi-width histograms are better than equi-depth histograms to capture skew in data

(b) Suppose in a equi-width histogram there is an entry from 21.0-30.0 with frequency 100. Assuming uniformity, the best guess of number of items between 22.0 - 26.0 is 50.

24 Which of the following about atomicity for transaction is/are correct?

(a) It means that a transaction, once started, must finish

(b) It means that operations between different transaction cannot interleave

25 In terms of transaction processing, why do DBMS need to have mechanisms to ensure durability?

A. The statement is false. Durability is always automatically maintained.

B. Updated value may get rollback if transactions aborts

C. Updated value is always flushed to the disk immediately (even before the transaction commits)

D. Updated value by a transaction may still remains in the buffers in main memory after transaction commits.

26 Consider 2 transaction T1 and T2, with a database Which of the following pairs of transaction is in conflict?

(a) T1 Read(X), T2 Write(X)

(b) T1 Read(X), T2 Write(Y)

(c) T1 Write(Y), T2 Write(Y)

27 Which of the following about conflict serializability is correct?

(a) If a schedule is conflict serializable, it can be transformed into a serial schedule (by swapping adjacent operations in the schedule)

(b) A schedule that is NOT conflict serializable will never produce the same result as a serial schedule of the same set of transactions.

28 What is "two-phase" in two phase locking?

A. None of the other answers are correct

B. Once you start releasing locks you cannot obtain more locks

C. Once you start acquiring X-locks, you cannot acquire S-locks

D. Once you start acquiring X-locks, you cannot acquire S-locks

29 What is the motivation of having strict two-phase locking over basic two-phase locking?

A. Strict two phase locking avoid non-recoverable schedule

B. Strict two phase locking require only one type of lock

C. Strict two phase lockng allow for more concurrency

D. All of the other answers are correct.

30 Consider deadlock avoidance in two-phase locking. Suppose Ti is holding a S-lock on an item, and then a transaction Tj request an X-lock on the same item. Which of the following is correct? (Assume i < j)

A. Tj will be allowed to proceed because i > j

B. Tj will be allowed to wait if "wait-die" policy is in place

C. Tj will be allowed to wait if "wound-wait" policy is in place

D. No matter what policy is used, Tj will abort

31 For recovery purpose, why do we need to have log an operation before executing it?

A. The statement is wrong, one shuld exeuction the opeartion before logging it

B. Otherwise many transactions will have to wait

C. Otherwise we may not be aware of an operation being executed when the system restart

D. Otherwise it is going to be less efficient

32 Let say a transaction T holds an S-lock of an object. Which of the following statement(s) is/are true?

(a) If T wants to obtain an X-lock of the same object, it will be granted

(b) If a different transaction T' want to obtain an S-lock on the same obiect, it will be granted

(c) If a different transaction T' want to obtain an X-lock on the same object, it will be granted

A. (a), (b) and (c)

B. (b)

C. (a) and (b)

D. (a)

1 How long does one sector to transfer?

one sector is 512 bytes

data-transfer is 200 MB per second

2 Consider the following numbers, seek time = 7 ms, rotational latency = 5ms. Data transfer rate = 50 MB per second

a. Reading one 4KB block

Time =?

b. Reading one block

Time = ?

c. Reading 10 consecutive block same track

Time = ?

d. Reading 10 blocks on different tracks

Time = ?

3 Two tables `Department(dept_name, building, budget)`. Assume each tuple is 40 bytes, `dept_name` is key. `Instructor(id, name, dept_name, salary)`. Assume each tuple is 50 bytes. `id` is key, `dept_name` is foreign key (referencing Department). Assume data are stored on disk. Each page has 1050 bytes. Assume 50 bytes are needed for overhead information. Assume the pages are fully filled

a. Now assume Department has 100,000 tuples, So number of pages =?

b. Assume Instructor has 400,000 tuples, So number of pages =?

c. Now consider the following query

```
SELECT * FROM Instructor
where id = “1997”
```

Now if you have a heap file, you have to look for each tuple.

i. You can stop when you find the tuple (why?) Worst case is what ? How many pages you have to search, i.e. what the total cost?

ii. However, if Instructor table is sorted via id?

d. Now consider the following query

```
SELECT * FROM Instructor
where id = “1997”
```

However, for magnetic disk, we need to worry about seek/rotation , for binary search, subsequent searches are not on consecutive pages, thus need rotate (or even seek). Now suppose reading a page take s seconds, and a rotate/seek take $100 \times s$ second

i. Worst case. Then time for heap file = ?

ii. Worst Time for sequential file = ?

e. Now consider the following query

```
SELECT * FROM Instructor
where id = “1997”
```

What is the file is smaller (e.g. 400 pages)

i. Worst case. Then time for heap file = ?

ii. Worst Time for sequential file = ?

4 Merge sort

Suppose you have a file with 12,800 pages, Assume you have 200 page of memory available.

The total number of iterations?

So the total page read/written?

Write the the process.

5 Hm1.1

6 B+-tree Hm 1.2

7 Dynamic hashing

- Items to be stored are numbers
- Assume each bucket store 2 numbers

  insert 14, 7

  insert 8

  insert 15

  insert 11

  insert 1

  insert 5, 9

8 Extensible hashing

- Items to be stored are numbers
- Assume each bucket store 2 numbers

  insert 14, 7

  insert 8

  insert 15

  insert 11

  insert 1

  insert 5, 9

9 Linear hashing

- Items to be stored are numbers
- Assume each bucket store 2 numbers

  insert 14, 7

  insert 8

  insert 15

  insert 11

  insert 1

insert 5, 9

10 Suppose we want to represent a ternary relationship between three entities. Assume each entity has a unique key attribute, and other attributes that need to be stored. Now suppose you are only allowed to use embedding to store the information. Suggest what can be a problem. Use an example to illustrate. Is there any restrictions on the relationship that will make this at least a feasible way of representing the relationship?

11hw 2.1

12 hw 2.2

13 Consider joining to tables R (1000 pages), S (100000 pages), assume we have 10 pages of buffers.

a. Consider sorting R write the iterations and total number of read/write

b. Consider sorting S write the iterations and total number of read/write

c. Total cost?

d. if not sorting table first?

14 Hash joins

Consider joining to tables R (1000 pages), S (100000 pages), Assume we have 11 pages of buffers. Assume all hash function evenly distribute the tuples for both tables.

What is the iteration and the running time ?

15

```
SELECT S.id, I.id, I.salary
From Student S, Instructor I, Advise A
WHERE S.id = A.s_id AND I.id = A.i_id AND S.dept = "CS" AND S.gpa >= 3.5
```

$$\Pi_{S.id,I.id,I.salary}(\sigma_{S.gpa\ And\ S.dept='CS'}(Student \bowtie_{A.sid=S.id''} Advisor \bowtie_{A.iid=I.id''} Instructor)) \tag{1}$$

Build a parse tree

16

$$\sigma_{gpa>3.0}(Student) \bowtie Department \tag{2}$$

Suppose

- 1000 pages for Department

- 5000 pages for Student
- $\sigma_{gpa>3.0}(Student)$ return 2500 pages
- Suppose 100 buffers, split 50 each

a. Consider Student in the inner loop (no pipeline)

cost is what?

b. Consider Student in the outer loop (no pipeline)