Advanced Data Management Technologies

Written Exam

31.01.2017

First name		Last name	
Student number		Signature	

Instructions for Students

- Write your name, student number, and signature on the exam sheet.
- This is a **closed book** exam: the only resources allowed are blank paper, pens, and your head. Use a pen, not a pencil.
- You have 2 hours for the exam.
- Each question has exactly **one** correct answer.
- You will get

  +1 points for each correct answer,

  −1 points for each wrong answer,

  0 points if you abstain.

  Advise: *if you are not sure about an answer, it is better to abstain.*

Good luck!

Reserved for the Teacher

Max. points	Plus Points	Minus Points	Sum
60			

BI and Multidimensional Modelling

1. What is offered by the three-layer DW architecture but not by the two-layer DW architecture?
   (a) A clear separation between analytical and transactional processing
   (b) DW is accessible even if the source systems are unavailable
   (c) A reconciled layer that forms a common reference data model for the whole enterprise
2. To which DW architecture corresponds query-driven data integration?
   (a) Single-layer DW architecture
   (b) Two-layer DW architecture
   (c) Three-layer DW architecture
3. What is true for query-driven data integration?
   (a) Query performance is high
   (b) Query is executed on the most up-to-date data
   (c) Query processing does not interfere with the local processing at the data sources.
4. The top-down approach of DW design
   (a) is based on a global picture of the goals
   (b) delivers a working system in the short term
   (c) is more flexible than the bottom-up approach with respect to changing requirements
5. The dimensional fact model is
   (a) a logical model against which the user can issue queries
   (b) a physical model to store a DW
   (c) a conceptual model with a graphical notation used for DW design
6. The multidimensional model
   (a) Is more flexible and general than the ER model
   (b) Serves one purpose and describes what is important and what describes the important things
   (c) Contains facts that describe important things and dimensions that are the important things
7. At which granularity level should facts be stored in the multidimensional model?
   (a) finest granularity, considering available resources and potential queries
   (b) finest granularity that is stored in production system
   (c) coarsest granularity to save disk space
8. What is a secondary event in a data warehouse?
   (a) The result of aggregating over a set of tuples in the fact table
   (b) The occurence of a fact, i.e., a tuple in the fact table
   (c) An entry in a dimension table

9. Junk dimensions are used to

   (a) store complex hierarchical relationships between dimensional attributes

   (b) store measures that are not available for all facts

   (c) group and store several degenerate dimensions

10. Surrogate keys

    (a) shall not be used if data is frequently consolidated or integrated from different sources

    (b) have performance advantages since they typically require much less space than operational keys

    (c) are important to store "intelligence" from the applications

11. A measure *quantity* that stores the number of sold items in a fact table with sales transactions is

    (a) additive

    (b) semi-additive

    (c) non-additive

12. Which measures are easiest to handle in a DW?

    (a) additive

    (b) semi-additive

    (c) non-additive

13. The use of shared dimensions helps to

    (a) increase the query performance

    (b) to break down the development process into small chunks

    (c) design data marts that can be easily integrated

14. Fact normalization means

    (a) All measures in the fact table are divided by the largest value in the corresponding domain to obtain a value between 0 and 1

    (b) All measures are collapsed into a single measure together with a special fact dimension that identifies the type of the measure

    (c) Split a fact table with more than one measure into several fact tables, each of which contains exactly one measure.

15. Compared to the snowflake schema, the star schema

    (a) requires no joins at query time

    (b) requires less space

    (c) has a better query performance

16. What are the advantages of using dimensions with many attributes?

    (a) Provides more flexibility for data analysis

    (b) Reduces the size of the fact table

    (c) Reduces the number of dimensions

Changing Dimensions and ETL

17. What happens if old values in a dimension table are overwritten?

    (a) Old facts point to incorrect information in the dimension table

    (b) New facts (inserted after changing the dimension table) point to incorrect information in the dimension table

    (c) Old and new facts point to correct information in the dimension table

18. What is a good strategy for ETL?

    (a) Implement all transformation in one single programm

    (b) Implement the transformations in a sequence of small operations/programms

    (c) Implement the transformations in the source database

19. Which of the following techniques does not help to tune the load step in the ETL process?

    (a) Sort the data before starting the load process

    (b) Disable the creation of log files

    (c) Use SQL-based updates

20. In the ETL process, what must be updated first?

    (a) Fact table

    (b) Indices

    (c) Dimension tables

Group-By Extensions, Window Functions, GMDJ

21. What is the correct processing order of an SQL statement?

    (a) FROM, WHERE, GROUP BY, HAVING, NTILE(4) OVER ()

    (b) FROM, WHERE, HAVING, GROUP BY, NTILE(4) OVER ()

    (c) NTILE(4) OVER (), FROM, WHERE, HAVING, GROUP BY

22. Which function can be used to programmatically determine the rollup level in SQL?

    (a) ROLLUP

    (b) GROUPING_ID

    (c) RANK

23. How many groupings are produced by the following GROUP BY clause?

```
GROUP BY ROLLUP(a, b), GROUPING SETS ((c,d),(e,f)), CUBE(g,h)
```

    (a) 24

    (b) 32

    (c) 48

24. What is the number of result tuples of the following GROUP BY clause, if $|a| = 1$, $|b| = 2$, $|c| = 3$, and $|d| = 4$?

```
SELECT   a, b, c, d, COUNT(*)
FROM     r
GROUP BY a, ROLLUP(b, c, d)
```

(a) 24

(b) 33

(c) 38

25. A composite column in the SQL GROUP_BY extensions

(a) is a shorthand for a set of columns

(b) allows to skip aggregation across certain levels

(c) is a compact way to generate all possible groupings among individual columns

26. How many different rankings over a data set can be computed in a single (unnested) SQL query using window functions?

(a) one

(b) two

(c) an arbitrary number

27. Consider the centered aggregate query:

```
SELECT Day, SUM(A) AS Sum,
       AVG(SUM(A)) OVER ( ORDER BY T RANGE BETWEEN INTERVAL '1' DAY PRECEDING
                          AND INTERVAL '1' DAY FOLLOWING ) AS CAvg
FROM r
```

and the partial result table:

Time	Sum	CAvg
1-JAN-2015	10	
2-JAN-2015	20	
3-JAN-2015	30	
4-JAN-2015	40	

Which are the correct values of the last column (first value corresponds to first tuple, etc.)?

(a) 10.0, 20.0, 30.0, 35.0

(b) 15.0, 20.0, 30.0, 35.0

(c) 23.3, 20.0, 30.0, 26.6

28. The GMDJ can be systematically transformed to SQL by using

(a) WINDOW functions

(b) GROUP BY extensions and WINDOW functions

(c) a combination of JOIN and CASE clauses

29. Which aggregate function can be incrementally computed as $F(A) = G(F(A_1), \ldots, F(A_k))$ with $A_1 \cup \cdots \cup A_k = A$ and $A_i \cap A_j = \emptyset$ and G is super-aggregate?

(a) Algebraic aggregate function

(b) Distributed aggregate function

(c) Holistic aggregate function

Pre-Aggregates

30. Pre-aggregation in DW aims to

    (a) reduce space requirements

    (b) increase query performance

    (c) reduce the update cost

31. How many pre-aggregates can be computed in an n-dimensional data cube?

    (a) $\sqrt{n}$

    (b) n^2

    (c) 2^n

32. In the greedy algorithm for pre-aggregate selection, the benefit of a view v depends

    (a) only on the views w that depend on v, i.e., $w \leq v$

    (b) on the set of already selected views and the views that depend on v

    (c) on the set of all views

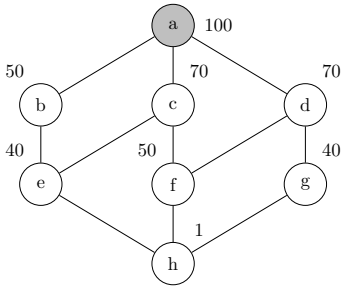
33. The greedy algorithm for pre-aggregate selection

    (a) is optimal if all benefits are equal

    (b) is optimal if the benefit of the first view is much larger than the other benefits

    (c) is never optimal

34. Given is the following lattice with the indicated costs, and view a is already materialized:



    If two other views shall be materialized, which ones would be selected by the greedy algorithm?

    (a) b, c

    (b) b, d

    (c) c, d

View Maintenance and Bitmap Indexes

35. Incremental maintenance of aggregation views require to store additional book-keeping information, e.g., tuples of the form $(group, minimum, count)$ for the MIN aggregate function. Assume an entry $(g, 1000, 1)$ in a view. How is the new MIN value determined when the tuple $(g, 1000)$ is deleted from the original table?

    (a) Scan entire original table

    (b) Search original table from the deleted tuple backwards

    (c) Do a binary search on the original table

36. Given is the following view:

```
SELECT   a, b, SUM(c)
FROM     r
GROUP BY a, b
```

To make the view self-maintainable and support incremental view maintenance, the tuples of the view must have the form

(a) `(a, b, sum)`

(b) `(a, b, sum, count)`

(c) `(a, b, sum, count, avg)`

37. What is the correct run-length encoding of the bitmap `00000010110000100000000000000`?

(a) `11011010011011`

(b) `11010010011000`

(c) `11000110011000`

38. What is the maximal space consumption of a compressed bitmap index for a table with n records?

(a) $2n$

(b) $n \log_2 2n$

(c) $2n \log_2 n$

39. How is the growth of a bit-sliced index for a numeric attribute C?

(a) logarithmically in the size of the domain of C

(b) linear in size of the domain of C

(c) linear in the number of tuples of the relation

40. A well-defined coding function in a bitmap-encoded index minimizes

(a) the number of bit vectors

(b) the number of bit vectors to be accessed for a selection predicate

(c) the number of index entries

NoSQL and MapReduce

41. What is a major problem for RDBMs to scale to big data?

(a) Lack of efficient index structures

(b) XML data cannot be stored in relational tables

(c) ACID properties

42. The CAP theorem states about the 3 properties Consistency, Availability, and Partition tolerance:

(a) at least 2 of the 3 properties must be satisfied at any time

(b) at most 2 of the 3 properties can be achieved at any time

(c) exactly 2 of the 3 properties are satisfied at any time

43. Which of the following is not a BASE property?

    (a) an application works basically all the time

    (b) an application does not have to be consistent all the time

    (c) an application will always be in a consistent state

44. Which of the following NoSQL data models is known for high performance, scalability and flexibility?

    (a) key-value stores

    (b) column stores

    (c) graph databases

45. In MapReduce, the programmer

    (a) must only specify a map and a reduce function

    (b) must also specify how to distribute the data

    (c) must also specify how to partition intermediate key-value pairs

46. Which of the following statements about the map function is wrong?

    (a) Can do something to each individual key-value pair, but cannot look at other key-value pairs

    (b) Can emit only one intermediate key-value pair for each incoming key-value pair

    (c) Can emit data with specific keys to all reducers

47. In MapReduce, the reduce tasks can start to work

    (a) when a map task produces the first output

    (b) when the first map task has completed

    (c) only after all map tasks have completed

48. How does the pull-scheduling strategy of MapReduce work?

    (a) Task tracker requests tasks from the Job tracker

    (b) Job tracker pushes tasks to Task tracker

    (c) Map tasks are requested by the task tracker, whereas reduce tasks are pushed by the job tracker

49. Speculative execution in Hadoop means that

    (a) a redundant task is started if an error occurs

    (b) a redundant task is started for slow tasks (stragglers)

    (c) a task is aborted and restarted again if it does not send a heartbeat meassage for a given time

P2P Networks and Distributed Hash Index

50. What is true about unstructured P2P networks?

    (a) The network is very stable

    (b) It is difficult to build and join the network

    (c) Data might not be found even if they are in the network

51. Which replication policy should be used if data consistency has the highest priority?

    (a) Eager replication with primary copy

    (b) Lazy replication with primary copy

    (c) Lazy replication without primary copy

52. Which of the following consistency levels leads to the best performance in P2P systems?

    (a) Strong consistency

    (b) Weak consistency

    (c) Eventual consistency

53. What is stored in the client image in the GFS?

    (a) A part of the global file system namespace

    (b) Meta-information about where the chunks of a file that has been read before are stored

    (c) Information about where the local data is replicated

54. What is a major problem with a naive solution of a distributed hash index, where each hash key is assigned to a different peer?

    (a) Lookup is slow

    (b) The data are not evenly distributed among the available peers

    (c) If the hash function changes, the hash value of most objects changes too.

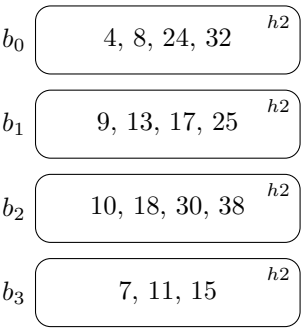
55. Which is the correct lookup function for centralized linear hashing (p is the split pointer, h_n, h_{n+1} are the hash functions)?

    (a) Lookup(k)
    $a = h_n(k)$;
    if $(a < p)$ **then** $a = h_{n+1}(k)$;

    (b) Lookup(k)
    $a = h_n(k)$;
    if $(a \geq p)$ **then** $a = h_{n+1}(k)$;

    (c) Lookup(k)
    $a = min(h_n(k), h_{n+1}(k))$;

56. Given is the following LH structure with $h_2(k) = k \mod 4$, $p = 0$, and each bucket can hold at most four tuples:



What steps are executed if a tuple with key 5 is added?

(a) Bucket b_1 is split and the keys of b_1 and the new key 5 are distributed among b_1 and the new bucket b_4, split pointer is set to $p = 1$

(b) An overflow bucket is added to b_1 storing 5, bucket b_0 is split and 4 is moved to the new bucket b_4, split pointer is set to $p = 1$

(c) An overflow bucket is added to b_1 storing 5, bucket b_0 is split, but no keys are moved to the new bucket b_4, split pointer remains $p = 0$

57. In distributed linear hashing, the so-called forward algorithm

(a) handles bucket overflows by forwarding data to other peers

(b) has to cope with lookup errors due to outdated local information

(c) forwards a lookup request to a central server

58. Which statement about consistent hashing is not correct?

(a) Nodes and data keys are mapped to the same range

(b) Peers are arranged in a logical ring

(c) A key is stored at the closest predecessor or successor node

59. With the help of finger tables the lookup performance in Chord is improved from $O(n)$ to

(a) $O(1)$

(b) $O(\log n)$

(c) $O(n \log n)$

60. Concurrency control in main-memory databases

(a) is almost not needed

(b) is more important than in traditional disk-based databases

(c) requires a complicated lock table data structure