Advanced Data Management Technologies
Written Exam

19.09.2018

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.

Good luck!

Reserved for the Teacher

Max. points	Plus Points	Minus Points	Sum
60			

BI and Multidimensional Modelling

1. What is the correct hierarchy of the BI pyramid (from lowest to highest)?

   (a) operational applications, what-if analysis, OLAP analysis, information exploration, data mining, decisions

   (b) operational applications, OLAP analysis, information exploration, data mining, what-if analysis, decisions

   (c) operational applications, information exploration, data mining, what-if analysis, OLAP analysis, decisions

2. 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

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) delivers a working system in the short term

   (b) is more flexible than the bottom-up approach with respect to changing requirements

   (c) is based on a global picture of the goals

5. What is a potential risk of supply-driven data mart design?

   (a) User requirements might not be sufficiently considered

   (b) The specification of the data sources is incomplete

   (c) It might be very time intensive since users do not have a clear understanding of the business goals

6. Which relationship between dimensional attributes is represented by a multiple arc in the dimensional fact model?

   (a) one-to-many relationship

   (b) many-to-one relationship

   (c) many-to-many relationship

7. 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

8. What is typical for OLAP?
   (a) A complex data model
   (b) The system is always available for updates and reads
   (c) Frequent read operations and infrequent updates

9. Why should facts in the multidimensional model be stored at the most detail level (considering available resources)?
   (a) Since drill-down queries can be answered more efficiently
   (b) Since this level determines the maximum detail level for querying the DW
   (c) Since disk space is never a problem

10. What are the advantages of using dimensions with many attributes?
   (a) Reduces the number of dimensions
   (b) Reduces the size of the fact table
   (c) Provides more flexibility for data analysis

11. 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

12. 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

13. 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 useful since they store “intelligence” from the applications

14. 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

15. In the inventory periodic snapshot model, a measure *quantity* that stores the quantity of a product is
   (a) semi-additive
   (b) additive
   (c) non-additive

16. 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 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.

17. Compared to the snowflake schema, the star schema genenerally

    (a) requires less aggregations at query time
    (b) requires less joins at query time
    (c) requires less storage space

18. Role-playing in the multidimensional model means that

    (a) a single dimension appears several times in the same fact table
    (b) a measure in the fact table represents different values
    (c) multiple hierarchies coexist in a dimension table

19. Bridge tables help to deal with

    (a) one-to-one relationships
    (b) one-to-many relationships
    (c) many-to-many relationships

    between dimensional attributes

Changing Dimensions and ETL

20. 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

21. Which of the following statements is correct?

    (a) ETL does not care about data quality but only efficiency
    (b) ETL is the most underestimated and time-consuming part of DW development
    (c) ETL must be done daily

22. Which of the following techniques helps to tune the load step in the ETL process?

    (a) Sort the data before starting the load process
    (b) Create indices
    (c) Use SQL-based update programs

23. Data cleansing

    (a) is extremely important since data almost never has decent quality
    (b) is only needed if data comes from many different sources
    (c) is rarely needed in DW

Group-By Extensions and Window Functions

24. What is the correct execution order of an SQL statement?

    (a) SELECT, FROM, WHERE, GROUP BY, HAVING, ORDER BY
    (b) FROM, WHERE, GROUP BY, HAVING, SELECT, ORDER BY
    (c) SELECT, FROM, WHERE, GROUP BY, ORDER BY, HAVING

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

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

    (a) 24
    (b) 28
    (c) 32

26. 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) 13
    (b) 23
    (c) 33

27. 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

28. What is a core characteristic of SQL analytic functions?

    (a) They allow for the first time to compute cumulative aggregates
    (b) They provide access to more than one row without a self join
    (c) They provide new possibilities for sorting the data

29. 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) 15.0, 20.0, 30.0, 35.0
    (b) 10.0, 20.0, 30.0, 35.0
    (c) 23.3, 20.0, 30.0, 26.6

Pre-Aggregates

30. Pre-aggregation in DW aims at

    (a) reducing space requirements

    (b) increasing query performance

    (c) reducing 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

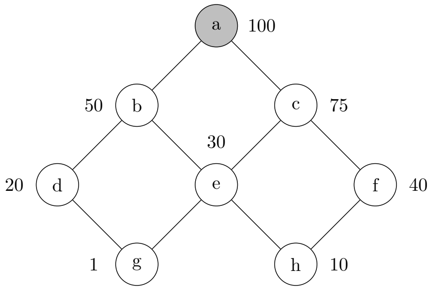
32. The greedy algorithm for pre-aggregate selection

    (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

33. 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, f

    (c) c, d

View Maintenance and Bitmap Indexes

34. Incremental view maintenance for the min/max aggregate functions needs to scan the base table

    (a) if the current min/max is deleted

    (b) if a new tuple is inserted in the base table

    (c) only at the beginning when the view is created

35. Given is the following view:

```
SELECT   a, b, MIN(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, min)`
(b) `(a, b, min, sum)`
(c) `(a, b, min, count)`

36. The compressed bitmap of 0000001011000010000000000 using run-length encoding is

(a) 11010010011000
(b) 11011010011001
(c) 11011010011010

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

(a) $2n$
(b) $2n \log_2 n$
(c) $n \log_2 2n$

38. A bitmap-encoded index for an attribute C consists of a

(a) bit matrix only
(b) bit matrix and a conversion table
(c) a bit matrix for each value in the domain of the attribute C

39. Indices based on bit vectors can be used for

(a) numeric attributes only
(b) non-numeric attributes only
(c) numeric and non-numeric attributes

NoSQL and MapReduce

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

(a) Lack of efficient index structures
(b) ACID properties
(c) No mechanism is provided for the concurrent execution of queries

41. What does "Availability" mean in the CAP theorem?

(a) All clients need always stay connected
(b) The system is "always on", no downtime
(c) The system continues to function even when split into disconnected subsets due to network errors

42. Which of the following is a BASE property?

    (a) An application can be considered to work in isolation
    (b) An application must always be consistent
    (c) An application does not have to be consistent all the time

43. 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

44. In MapReduce, the reducer is called once for each

    (a) intermediate key-value pair
    (b) intermedidate value
    (c) intermediate key and set of values with that key

45. The following reduce function computes the relative word frequency across a set of documents:

```
reduce(String key, Iterator values);
if key == "" then
    ...;
else
    int word_count = 0;
    foreach v in values do
        word_count += ParseInt(v);
    Emit(key, AsString(word_count / total_word_count));
```

Which code snippet is missing in the if-block?

(a) total_word_count = 0;
    ForEach v in values **do** total_word_count += ParseInt(v);
(b) **ForEach** v in values **do** total_word_count += ParseInt(v);
(c) total_word_count += ParseInt(values);

46. Given is the following MapReduce program:

```
map(key, record):
  emit(record, null)

reduce(key, records):
  emit(key)
```

Which is the corresponding SQL statement?

(a) SELECT * FROM table;
(b) SELECT DISTINCT * FROM table;
(c) SELECT A FROM table; // where A is the primary key of table

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

    (a) at any time
    (b) when the first map task has completed
    (c) after all map tasks have completed

48. The hearbeat message of a TT sent to the JT

    (a) may contain a request for a map or a reduce task
    (b) always contains a request for a map or a reduce task
    (c) is only sent when a task is finished to return the result

49. In MapReduce, how can a crash of the master node (job tracker) be handeled?

    (a) Start a redundant master node as soon as problems with the current master node are discovered
    (b) There is no mechansims; the master node need to be restarted and all jobs need to be resubmitted
    (c) Assign the role of the master node to one of the slave nodes

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. What is “flooding” in unstructured P2P networks?

    (a) A way to connect and disconnect from the network
    (b) A mechanism to distribute data among the peers
    (c) A search technique to locate data

52. What distinguishes structured P2P networks from unstructured networks?

    (a) Any node can efficiently search the network for data
    (b) Joining and leaving the network becomes easier
    (c) Worse performance and stability

53. Which replication policy should be used in a P2P network if throughput should be maximized?

    (a) Eager replication with primary copy
    (b) Lazy replication with primary copy
    (c) Lazy replication without primary copy

54. What is natively achieved in distributed file systems for very large data by using chunks (rather than files) as basic storage units?

    (a) Reliability
    (b) Fair load balancing
    (c) Availability

55. The client image in the GFS

    (a) solves scalability issues by caching the location of previously read file chunks
    (b) improves the performance by creating an index on the data chunks
    (c) reduces the required disk space by avoiding redundant storage of data chunks

56. 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.

57. What is true about linear hashing (LH)?

    (a) LH provides a logarithmic growth of the hash directory

    (b) A large part of the hash directory remains unchanged when the hash function is modified

    (c) Whenever a bucket overflows, this bucket is immediately split

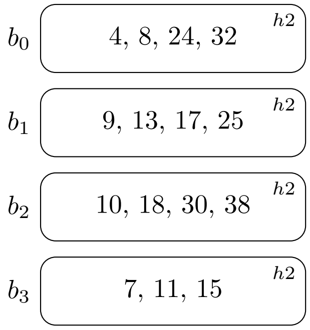
58. Which is the correct lookup function for centralized linear hashing (p is the split pointer, h_n, h_{n+1} are the hash functions)?

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

59. 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 $p = 0$

60. In consistent hashing, if a new node joins the network

    (a) all keys need to be reassigned

    (b) no keys need to be reassigned

    (c) some keys of the new node's successor need to be reassigned