Advanced Data Management Technologies

Written Exam

15.09.2015

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,

  −0.5 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			

1. What is Business Intelligence?
   (a) A system that makes intelligent decisions for the user
   (b) A combination of processes, technologies, and applications used to support decision making
   (c) An method to store huge amounts of data in a central repository
2. 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
3. What is true for warehouse-driven data integration?
   (a) The most current data is available
   (b) Query processing competes with local processing at the sources.
   (c) The query performance is high
4. The top-down approach of DW design
   (a) delivers a working system in the short term
   (b) is based on a global picture of the goals
   (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 graphical conceptual model for DW design
6. At which granularity level should facts be stored in the multidimensional model?
   (a) finest granularity
   (b) coarsest granularity
   (c) depends on the specific application
7. What is a primary event in a data warehouse?
   (a) A particular occurence of a fact, i.e., a tuple in the fact table
   (b) The result of aggregating over a set of tuples in the fact table
   (c) The selection of a single tuple from the fact table
8. Junk dimensions are used to
   (a) store complex hierarchical relationships between dimensional attributes
   (b) group and store several degenerate dimensions
   (c) store measures that are not available for all facts

9. 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

10. 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

11. A data warehouse bus matrix specifies

    (a) the attributes of the dimension tables

    (b) the hierarchies in the dimension tables

    (c) which dimensions are used by which business processes

12. The use of shared dimensions helps to

    (a) design data marts that can be easily integrated

    (b) increase the query performance

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

13. 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.

14. Compared to the star schema, the snowflake schema

    (a) has a better query performance

    (b) uses more space

    (c) requires more joins at query time

15. Compared to the snowflake schema, the star schema

    (a) has a better query performance

    (b) requires more joins at query time

    (c) requires generally less space

16. 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

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

    (a) Reduces the size of the fact table

    (b) Reduces the number of dimensions

    (c) Provides more flexibility for data analysis

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

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

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

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

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

    (a) ROLLUP

    (b) GROUPING_ID

    (c) RANK

20. 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

21. What is the maximum number of result tuples of the following GROUP BY clause, if the attributes have the following cardinalities: $|a| = 2$, $|b| = 3$, $|c| = 1$, and $|d| = 4$?

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

    (a) 24

    (b) 38

    (c) 39

22. 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

23. 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

24. Which type of aggregates can be efficiently computed by GMDJ but not in SQL?

(a) Distributive aggregates

(b) 1D cumulative aggregates

(c) 2D cumulative aggregates

25. 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

26. What is the correct way to incrementally compute an algebraic aggregate function F over a set A?

(a) $F(A) = G(F_1(A), \ldots, F_m(A))$ with G is super-aggregate and F_i are aggregate functions

(b) $F(A) = G(F(A_1), \ldots, F(A_k))$ with $A_1 \cup \cdots \cup A_k = A$ and $A_i \cap A_j = \emptyset$ and G is super-aggregate

(c) $F(A) = F(A_1 \cup \cdots \cup A_k)$ with $A_1 \cup \cdots \cup A_k = A$

27. Computing the optimal number of pre-aggregates in a DW

(a) is NP-complete

(b) can be done by a simple greedy algorithm

(c) is provided in any commercial DW system

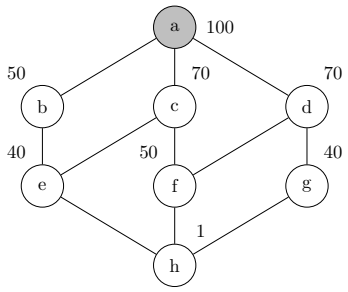
28. In the greedy algorithm for pre-aggregate selection, the benefit of a view v depends

    (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

29. 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, d

    (b) b, c

    (c) c, d

30. 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

31. What is an efficient index structure for attributes with low cardinality?

    (a) Hash index

    (b) B-tree index

    (c) Bitmap index

32. The compressed bitmap `10110011011` is the run-length encoding of

    (a) `00010010000000010000001`

    (b) **`00010010000000010000000`**

    (c) `00010010000000011111111`

33. 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) $n \log_2 2n$

34. 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

35. Indices based on bit vectors can be used for

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

36. Which of the following statements is correct?

    (a) ETL is the least time-consuming part of DW development
    (b) The most important aspect of ETL is efficiency
    (c) Data extracted in ETL almost never has decent quality

37. 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

38. 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

39. 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

40. 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

41. Which of the following NoSQL data models offers high performance, scalability and flexibility?

    (a) column stores
    (b) key-value stores
    (c) graph databases

42. 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

43. In MapReduce, the combiner function can be used to

    (a) to merge tuples with the same key value inside each mapper in order to reduce the number of tuples that are shuffled to the reducer

    (b) combine intermediate tuples from all mappers that have the same key value

    (c) divide up the intermediate key space for parallel reduce operations

44. What is a meaningful map function in MapReduce for the word count example?

    (a)
    ```
    map( String key, String value);
    ForEach w in value do EmitIntermediate("1",w);
    ```

    (b)
    ```
    map( String key, String value);
    ForEach w in value do EmitIntermediate(w,"1");
    ```

    (c)
    ```
    map( String key, String value);
    ForEach w in value do EmitIntermediate(w,w);
    ```

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 * FROM table WHERE A = null;`

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

(a) Job tracker pushes tasks to Task tracker

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

(c) Task tracker requests tasks from the Job tracker

48. Which mechanism is provided in Hadoop to deal with an error of the master node?

(a) One of the slave nodes takes the role of the master node

(b) The slaves run without a master until a new master is started

(c) No mechanism is provided

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

50. What is true about unstructured P2P networks?

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

(b) The network is inherently unstable

(c) It is difficult to build and join 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. 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

53. 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.

54. Which is the correct lookup function for centralized linear hashing (p = split pointer, h_n, h_{n+1} are 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)
    $a = min(h_n(k), h_{n+1}(k))$;

55. 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

56. 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 node (predecessor or successor)

57. In consistent hashing, if a node leaves the network

    (a) the keys of that node are assigned to the node's successor
    (b) the keys are removed
    (c) the keys of that node are re-assigned to nodes using a hash function

58. 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) $O(n \log n)$

59. Which is a critical aspect for data representation in main memory databases?

    (a) Access locality
    (b) Compressing the size of the data
    (c) Variable length data fields

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