Advanced Data Management Technologies

Written Exam

10.02.2014

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 total exam (DW and DM part).
- Guidelines for answering the questions:
  - each question has exactly **one** correct answer
  - +1 for each correct answer
  - -1 for each wrong answer
  - 0 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 typical for OLAP?
   - A complex data model
   - The system is always available for updates and reads
   - Frequent read operations and infrequent updates
2. What is true for query-driven data integration?
   - Query performance is high
   - Query is executed on the most up-to-date data
   - Query processing does not interfere with the local processing at the data sources.
3. The multidimensional model
   - is less flexible and general than the ER model
   - serves many purposes and is very flexible
   - contains facts that describe important things and dimensions that are the important things
4. At which granularity level should facts be stored in the multidimensional model?
   - lowest (finest) granularity
   - depends on the specific application
   - highest (coarsest) granularity
5. Which statement about the multidimensional model is correct?
   - Dimensions should contain much information, which is then useful for the analysis
   - Dimensions should contain as little information as possible to save disk space
   - Dimensions can store at most one hierarchy
6. Which of the following statements is correct?
   - Surrogate keys produce larger fact tables
   - Surrogate keys make the DW independent from operational changes
   - Surrogate keys contain "intelligence" which is helpful for data analysis
7. Which measures are easiest to handle in a DW?
   - additive
   - semi-additive
   - non-additive
8. A data warehouse bus matrix specifies
   - which dimensions are used by which business processes
   - the attributes of the dimension tables
   - the hierarchies in the dimension tables
   - the measures

9. The use of shared dimensions helps to
   - increase the query performance
   - to break down the development process into small chunks
   - design data marts that can be easily integrated
10. Compared to the star schema, the snowflake schema
    - has a better query performance
    - requires more joins at query time
    - uses more space
    - hides the hierarchies
11. Role-playing in the multidimensional model means that
    - a single dimension appears several times in the same fact table
    - a measure in the fact table represents different values
    - multiple hierarchies coexist in a dimension table
12. 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)
    ```

    - 24
    - 32
    - 48
13. Which function can be used to programmatically determine the rollup level in SQL?
    - GROUPING_ID
    - ROLLUP
    - RANK
14. What is the correct processing order of an SQL statement?
    - FROM, WHERE, GROUP BY, HAVING, NTILE(4) OVER ()
    - FROM, WHERE, HAVING, GROUP BY, NTILE(4) OVER ()
    - NTILE(4) OVER (), FROM, WHERE, HAVING, GROUP BY
15. How many result tuples are produced by the following SQL statement, if a, b and c have 4, 5 and 2 different values, respectively?

    ```
    SELECT a, b, SUM(c)
    RANK() OVER (PARTITION BY a ORDER BY SUM(c) DESC)
    FROM r
    GROUP BY a, b
    ```

    - 9
    - 11
    - 20
    - 40

16. A composite column in the SQL GROUP_BY extensions

    - allows to skip aggregation across certain levels
    - is a shorthand for a set of columns
    - is a compact way to generate all possible groupings among individual columns

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

    - one
    - two
    - an arbitrary number

18. What is a core feature of the Generalized MD-Join?

    - Always sorts the data in the result table
    - The base table is automatically derived from the detail table
    - Allows to compute several complex aggregates with a single scan of the detail table

19. Which of the following statements is not correct?

    - SQL window functions can efficiently compute 1D and 2D cumulative aggregates
    - The GMDJ operator can efficiently compute 2D cumulative aggregates
    - The GMDJ operator can efficiently compute distributive and algebraic aggregates

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

    - a combination of JOIN and CASE clauses
    - WINDOW functions
    - GROUP BY extensions and WINDOW functions

21. Pre-aggregation in DW aims to

    - reduce space requirements
    - increase query performance
    - reduce the update cost

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

    - 2^n
    - n^2
    - $2n$
    - $\sqrt{n}$

23. In the greedy algorithm for pre-aggregate selection, the benefit of a view v depends
    - only on the views w that depend on v, i.e., $w \leq v$
    - on the set of all views
    - on the set of already selected views and the views that depend on v
24. The greedy algorithm for pre-aggregate selection
    - is optimal if all benefits are equal
    - is optimal if the benefit of the first view is much larger than the other benefits
    - is never optimal
25. Incremental view maintenance for the min/max aggregate functions needs to scan the base table
    - if the current min/max is deleted
    - if a new tuple is inserted in the base table
    - only at the beginning when the view is created
26. The compressed bitmap index of 000100100000100 is
    - 10010010001
    - 10100010001
    - 10110011001
27. What is the maximal space consumption of a compressed bitmap index for a table with n records?
    - $2n$
    - $2n \log_2 n$
    - $n^2 \log_2 n$
28. Indices based on bit vectors can be used for
    - numeric attributes only
    - non-numeric attributes only
    - numeric and non-numeric attributes
29. Which of the following statements is correct?
    - ETL is the most underestimated and time-consuming part of DW development
    - ETL does not care about data quality but only efficiency
    - ETL must be done daily
30. What is a good strategy for ETL?
    - Implement all transformation in one single programm
    - Implement the transformations in a sequence of small operations/programms
    - Implement the transformations in the source database

31. Data cleansing

    - is extremely important since data almost never has decent quality
    - is only needed if data comes from many different sources
    - is rarely needed in DW

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

    - Fact table
    - Dimension tables
    - Indices

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

    - Old facts point to incorrect information in the dimension table
    - New facts (inserted after changing the dimension table) point to incorrect information in the dimension table
    - Old and new facts point to correct information in the dimension table

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

    (a) Lack of efficient index structures
    (b) ACID properties
    (c) XML data cannot be stored in relational tables

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

    (a) at most 2 of the 3 properties can be achieved at any time
    (b) at least 2 of the 3 properties must be satisfied at any time
    (c) exactly 2 of the 3 properties are satisfied at any time

36. 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

37. Wich of the following NoSQL data models offers high performance, scalability and flexibility?

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

38. 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

39. What is the correct signature of the map and reduce functions?

    (a) map: $(k, v) \to (k', v')^*$, reduce: $(k', v'[]) \to (v'')^*$

    (b) map: $(k, v) \to (k, v')^*$, reduce: $(k, v'[]) \to (v'')^*$

    (c) map: $(k, v) \to (k', v')^*$, reduce: $(k', v') \to (v'')^*$

40. 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

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

    (a) intermediate key-value pair

    (b) intermediate key and set of values with that key

    (c) intermedidate value

42. In MapReduce, a combiner function can be used to

    (a) merge the output of all map tasks together before sending to the reduce tasks

    (b) store the output of the reduce tasks into a single file

    (c) minimize the data that is shuffled between map and reduce tasks

43. 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

44. 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);
```

45. In the MapReduce Top Ten pattern, how many records are sent to the reducer if Top-K is computed and M mappers are used?

    (a) $K \cdot M$ records

    (b) all input records

    (c) K records

46. Which is the most flexible join pattern in MapReduce?

    (a) Reduce side join

    (b) Replicated join

    (c) Composite join

47. The DistributedCache in Hadoop can be used

    (a) to share data among map tasks that is different from the input data

    (b) to store and share input splits

    (c) to cache the intermediate results before sending them to the reducers

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

    (a) Job tracker pushes tasks to Task tracker

    (b) Task tracker requests tasks from the Job 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 for slow tasks (stragglers)

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

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

50. What is not true for P2P networks?

    (a) Nodes can be both client and server, but not at the same time

    (b) Nodes enter and leave the network frequently

    (c) Nodes have widely varying capabilities

51. Which replication policy should be used if throughput should be maximized?

    (a) Eager replication with primary copy

    (b) Eager replication without primary copy

    (c) Lazy replication with primary copy

    (d) Lazy replication without primary copy

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

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

    (b) A part of the global file system namespace

    (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) If the hash function changes, the hash value of most objects changes too.
    (b) The data are not evenly distributed among the available peers
    (c) Lookup is slow

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) has to cope with lookup errors due to outdated local information
    (b) handles bucket overflows by forwarding data to other peers
    (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 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

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

    (a) $O(1)$
    (b) $O(n \log n)$
    (c) $O(\log n)$

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

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

60. Which is the most important index structure in main memory databases?

    (a) B-tree
    (b) T-tree
    (c) R-tree