Advanced Data Management Technologies

Written Exam

18.06.2014

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 exam.
- 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 Business Intelligence?
   (a) A combination of processes, technologies, and applications used to support decision making
   (b) A system that makes intelligent decisions for the user
   (c) An method to store huge amounts of data in a central repository
2. What is true for warehouse-driven data integration?
   (a) The query performance is high
   (b) The most current data is available
   (c) Query processing competes with local processing at the sources.
3. 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
4. At which granularity level should facts be stored in the multidimensional model?
   (a) lowest (finest) granularity
   (b) depends on the specific application
   (c) highest (coarsest) granularity
5. Which statement about the multidimensional model is correct?
   (a) Dimensions should contain much information, which is then useful for the analysis
   (b) Dimensions should contain as little information as possible to save disk space
   (c) Dimensions can store at most one hierarchy
6. 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 application
7. 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
8. A data warehouse bus matrix specifies
   (a) which dimensions are used by which business processes
   (b) the attributes of the dimension tables
   (c) the hierarchies in the dimension tables
   (d) the measures

9. 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

10. Compared to the star schema, the snowflake schema
    (a) has de-normalized dimension tables
    (b) has a better performance
    (c) is less efficient at query time due to many joins

11. 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

12. How many result groups are produced by the following GROUP BY clause, if a has 2, b has 3, c has 1 and d has 4 different values?

```
GROUP BY a, ROLLUP(b, c, d)
```

    (a) 24
    (b) 38
    (c) 39

13. Which function can be used to programmatically determine the rollup level in SQL?
    (a) GROUPING_ID
    (b) ROLLUP
    (c) RANK

14. What is a correct execution order?
    (a) SELECT, FROM, WHERE, GROUP BY, HAVING, ORDER BY
    (b) SELECT, FROM, WHERE, GROUP BY, ORDER BY, HAVING
    (c) FROM, WHERE, GROUP BY, HAVING, SELECT, ORDER 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
```

    (a) 9
    (b) 11
    (c) 20
    (d) 40

16. A composite column in the SQL GROUP_BY extensions

    (a) allows to skip aggregation across certain levels
    (b) is a shorthand for a set of columns
    (c) 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?

    (a) one
    (b) two
    (c) an arbitrary number

18. Which kind of aggregates cannot be computed by SQL window functions?

    (a) Distributive aggregates
    (b) 1D cumulative aggregates
    (c) 2D cumulative aggregates

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

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

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

    (a) a combination of JOIN and CASE clauses
    (b) WINDOW functions
    (c) GROUP BY extensions and WINDOW functions

21. How are algebraic aggregate functions evaluated with the Generalized MD-Join?

    (a) Are natively supported
    (b) Reduction to distributive aggregates in combination with a pre- and post-processing step
    (c) Reduction to holistic aggregates

22. Pre-aggregation in DW aims to

    (a) reduce space requirements
    (b) increase query performance
    (c) reduce the update cost

23. 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 all views
    (c) on the set of already selected views and the views that depend on v

24. 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

25. 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) Take the previous element in the view in sort order

    (c) Search the original table from the current position til the end

26. The compressed bitmap index of 000100100000100 is

    (a) 10010010001

    (b) 10110011001

    (c) 10100010001

27. 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) $n^2 \log_2 n$

28. Indices based on bit vectors can be used for

    (a) numeric attributes only

    (b) non-numeric attributes only

    (c) numeric and non-numeric attributes

29. Which of the following statements is correct?

    (a) ETL is the most underestimated and time-consuming part of DW development

    (b) ETL does not care about data quality but only efficiency

    (c) ETL must be done daily

30. 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

31. 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

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

    (a) Fact table
    (b) Dimension tables
    (c) Indices

33. Which is the most advanced solution to handle slowly changing dimensions?

    (a) Versioning of rows with changing attributes
    (b) Versioning of rows with changing attributes plus timestamping of rows
    (c) Create two versions of each changing attribute

34. 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

35. What does “Availability” mean in the CAP theorem?

    (a) The system is “always on”, no downtime
    (b) All clients see the same data
    (c) The system continues to function even when split into disconnected subsets due to network errors

36. 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

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 in MapReduce?

    (a) $map : (k, v) \rightarrow (k', v')^*, \quad reduce : (k', v'[]) \rightarrow (v'')^*$
    (b) map: $(k, v) \rightarrow (k, v')^*$, reduce: $(k, v'[]) \rightarrow (v'')^*$
    (c) map: $(k, v) \rightarrow (k', v')^*$, reduce: $(k', v') \rightarrow (v'')^*$

40. 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

41. 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

42. 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);`

43. Which of the following MapReduce design patterns has both a mapper and a reducer?

    (a) Filtering
    (b) Numerical summarization
    (c) Replicated join

44. 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

45. 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 DISTINCT * FROM table;`
(b) `SELECT * FROM table;`
(c) `SELECT * FROM table WHERE A = null;`

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 lookup 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. 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.