Advanced Data Management Technologies

Written Exam

30.01.2018

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.

  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 Business Intelligence?
   (a) A system that processes huge amounts of data and makes intelligent decisions for the user
   (b) A set of tools to store huge amounts of data in a central repository
   (c) A combination of processes, technologies, and applications used to support decision making
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. To which DW architecture corresponds query-driven data integration?
   (a) Single-layer DW architecture
   (b) Two-layer DW architecture
   (c) Three-layer DW architecture
4. 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
5. The bottom-up approach of DW design
   (a) requires huge initial investments
   (b) gives managers a quick feedback about the actual benefits of a data warehouse
   (c) requires to analyze and integrate all data sources at the beginning
6. Which relationship between dimensional attributes is represented by a multiple arc in the dimensional fact model?
   (a) many-to-many relationship
   (b) one-to-many relationship
   (c) many-to-one relationship
7. The multidimensional model
   (a) serves many purposes and is very flexible
   (b) is less flexible and general than the ER model
   (c) contains facts that describe important things and dimensions that are the important things
8. What is typical for OLAP?
   (a) Fewer, but "bigger" queries that typically need to scan a huge amount of records and do some aggregation
   (b) Many "small" queries on a small number of tuples from many tables that need to be joined
   (c) A mix of small and big queries that are evaluated over many relations

9. At which granularity level should facts be stored in the multidimensional model?
   (a) coarsest granularity to save disk space
   (b) finest granularity that is stored in production system
   (c) finest granularity, considering available resources and potential queries
10. 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
11. What is a primary event in a data warehouse?
   (a) The result of aggregating over a set of tuples in the fact table
   (b) A particular occurence of a fact, i.e., a tuple in the fact table
   (c) A single entry in a dimension table.
12. Which type of facts yield a dense cube?
   (a) Event facts
   (b) Fact-less facts
   (c) Snapshot facts
13. What is true for a degenerate dimension?
   (a) Contains only one attribute
   (b) Contains at most one hierarchy
   (c) Stores information that is not useful for querying
14. Which of the following statements is correct?
   (a) Surrogate keys produce larger fact tables
   (b) Surrogate keys make the DW independent from operational changes
   (c) Surrogate keys contain "intelligence" which is helpful for data analysis
15. A measure *discount rate* is always
   (a) additive
   (b) non-additive
   (c) semi-additive
16. Which measures are easiest to handle in a DW?
   (a) additive
   (b) semi-additive
   (c) non-additive
17. 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

18. Fact normalization collapses all measures into a single measure. This makes sense if
    (a) the fact table is sparsely populated
    (b) comparisons between different measures are frequent
    (c) all measures are additive
19. Compared to the star schema, the snowflake schema
    (a) has de-normalized dimension tables
    (b) hides the hierarchies
    (c) is less efficient at query time due to many joins
20. 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

Changing Dimensions and ETL

21. 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
22. 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
23. The data staging area is mainly used for
    (a) querying the DW
    (b) data transformations and cleansing
    (c) indexing dimensions
24. In the ETL process, what must be updated first?
    (a) Fact table
    (b) Indices
    (c) Dimension tables
25. Which of the following techniques for improving data quality during ETL is typically the most difficult one to apply?
    (a) Data stewards
    (b) DW-controlled improvements
    (c) Source-controlled improvements

Group-By Extensions and Window Functions

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

    (a) ROLLUP
    (b) GROUPING_ID
    (c) RANK

27. How many different groupings are created by `CUBE(`a_1`,...,`a_n`)`?

    (a) n^n
    (b) n^3
    (c) 2^n

28. How many result tuples are produced by the following SQL statement, if $|a| = 4$, $|b| = 5$ and $|c| = 2$?

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

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

29. A composite column in the SQL GROUP_BY extensions

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

30. Consider the query

```
SELECT type,
       SUM(amount) OVER () AS sales
       SUM(SUM(amount)) OVER () AS total_sales
FROM r
GROUP BY type
```

    and the following result table, where the third column is missing:

type	sales	total_sales
Direct	10.000	
Internet	30.000	
Partners	15.000	

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

    (a) 55.000, 55.000, 55.000
    (b) 10.000, 40.000, 55.000
    (c) The third column is not defined, i.e., syntactically not correct

31. 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

Pre-Aggregates

32. 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

33. 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

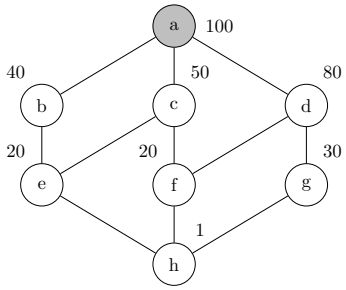
34. Which of the following assumptions in the greedy algorithm is unrealistic?

    (a) All views in the lattice have the same probability of being requested in a query

    (b) The user queries are identical to some views in the lattice

    (c) The time (or cost) to answer a query is equal to the size of the view from which the query is answered

35. 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) c, d

    (b) c, g

    (c) e, d

View Maintenance and Bitmap Indexes

36. 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

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

    (a) Hash index

    (b) B-tree index

    (c) Bitmap index

38. The (compressed) bitmap `10110011011` is the run-length encoding of

    (a) `00010010000000010000001`

    (b) `00010010000000010000000`

    (c) `00010010000000011111111`

39. Which of the following indices grows linearly with the number of distinct attribute values?

    (a) Bitmap index

    (b) Bit-sliced index

    (c) Bitmap-encoded index

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 does "Partition tolerance" mean in the CAP theorem?

    (a) The data need to be stored in different partitions

    (b) Nodes in different partitions see different data

    (c) The system continues to function even when split into disconnected subsets, e.g., due to network errors

42. 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

43. 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

44. What is the correct signature of the map and reduce functions in MapReduce?

    (a) $map : (k, v) \to list(k', v'), \quad reduce : (k', list(v')) \to list(v'')$

    (b) map: $(k, v) \to list(k, v'), \quad$ reduce: $(k, list(v')) \to list(v'')$

    (c) map: $(k, v) \to list(k', v'), \quad$ reduce: $(k', v') \to list(v'')$

45. 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

46. 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 EmitIntermediateToAllReducers(w,"1");
    ```

47. 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`)

48. 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

49. 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

50. 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

51. What is the "time to live" in unstructured P2P networks?

    (a) The maximum time that is available to process and complete a query

    (b) The minimum time a node/peer has to stay connceted to the network

    (c) The number of times a query is forwarded before being discarded to avoid using too much resources

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 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

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. 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

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.

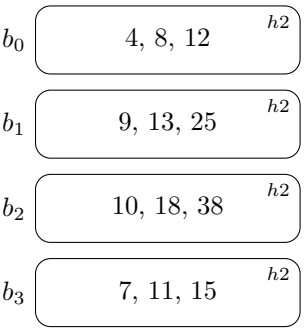
57. What is true about linear hashing (LH)?

    (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. Given is the following LH structure with $h_2(k) = k \mod 4$, split pointer $p = 0$, and each bucket can hold at most 3 tuples:



What happens if a tuple with key 42 is added?

(a) Bucket b_2 is split; the keys of b_2 and the new key 42 are distributed among b_2 and a new bucket b_4; split pointer is set to $p = 3$

(b) An overflow bucket is added to b_0; split pointer is set to $p = 1$; a new hash function $h_3(k) = k \mod 8$ applies to bucket b_2; keys in b_2 are distributed according to h_3

(c) An overflow bucket with 42 is added to b_2; bucket b_0 is split and $h_3(k) = k \mod 8$ applies to b_0; 4 and 12 are moved to a new bucket b_4; split pointer is set to $p = 1$

59. 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 of that node are assigned to the node's predecessor

(c) the keys of that node are distributed among all active nodes

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