Advanced Data Management Technologies

Written Exam

02.02.2016

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 the correct hierarchy of the BI pyramid (from lowest to highest)?

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

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

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

3. 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) A reconciled layer that forms a common reference data model for the whole enterprise

   (c) DW is accessible even if the source systems are unavailable

4. To which DW architecture corresponds query-driven data integration?

   (a) Single-layer DW architecture

   (b) Two-layer DW architecture

   (c) Three-layer DW architecture

5. The bottom-up approach of DW design

   (a) requires huge initial investments

   (b) gives managers a quick feedback about the actual benefits of the system being built

   (c) requires to analyze and integrate all data sources at the beginning

6. 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 conceptual model with a graphical notation used for DW design

7. Which relationship between dimensional attributes is represented by a multiple arc?

   (a) many-to-many relationship

   (b) one-to-many relationship

   (c) one-to-one relationship

8. The multidimensional model
   (a) is less flexible and general than the ER model
   (b) serves many purposes and is very flexible
   (c) contains facts that describe important things and dimensions that are the important things

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

10. 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) A single entry in a dimension table.

11. Which type of facts yield a dense cube?
    (a) Event facts
    (b) Fact-less facts
    (c) Snapshot facts

12. 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

13. 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

14. In the inventory periodic snapshot model, a measure *quantity* to store the quantity of each product is
    (a) additive
    (b) semi-additive
    (c) non-additive

15. 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

16. 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

17. Fact normalization collapses all measures into a single measure. This makes only sense if

    (a) the fact table is sparsely populated
    (b) comparisons between different measures are frequent
    (c) all measures are additive

18. Compared to the star schema, the snowflake schema

    (a) is less efficient at query time due to many joins
    (b) has de-normalized dimension tables
    (c) hides the hierarchies

19. 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

20. 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

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 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

23. Which of the following techniques does not help to tune the load step in the ETL process?

    (a) Sort the data before starting the load process
    (b) Disable the creation of log files
    (c) Use SQL-based updates

24. 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

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, Window Functions, GMDJ

26. 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

27. 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) 32

    (c) 48

28. 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) 38

    (b) 33

    (c) 24

29. 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

30. 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

31. 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) 10.0, 20.0, 30.0, 35.0

(b) 15.0, 20.0, 30.0, 35.0

(c) 23.3, 20.0, 30.0, 26.6

32. Which of the following statements is not correct?

(a) SQL window functions can efficiently compute 1D and 2D cumulative aggregates

(b) The GMDJ operator can efficiently compute 2D cumulative aggregates

(c) The GMDJ operator can efficiently compute distributive and algebraic aggregates

33. 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

Pre-Aggregates

34. Pre-aggregation in DW aims to

(a) reduce space requirements

(b) increase query performance

(c) reduce the update cost

35. 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

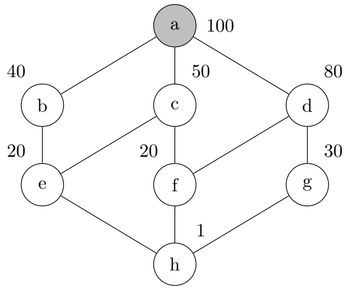
36. The greedy algorithm for pre-aggregate selection

(a) is never optimal

(b) is optimal if all benefits are equal

(c) is optimal if the benefit of the first view is much larger than the other benefits

37. 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, g

(b) b, d

(c) e, d

View Maintenance and Bitmap Indexes

38. 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

39. Given is the following view:

```
SELECT   a, b, SUM(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, sum, count, avg)`

(b) `(a, b, sum, count)`

(c) `(a, b, sum)`

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

(a) 11011010011001

(b) 11011010011010

(c) 11010010011000

41. 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

NoSQL and MapReduce

42. What is a major problem for RDBMs to scale to big data?
    - (a) Lack of efficient index structures
    - (b) XML data cannot be stored in relational tables
    - (c) ACID properties

43. 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

44. 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

45. What is the correct signature of the map and reduce functions in MapReduce?
    - (a) $\mathit{map} : (k, v) \rightarrow \mathit{list}(k', v'), \quad \mathit{reduce} : (k', \mathit{list}(v')) \rightarrow \mathit{list}(v'')$
    - (b) map: $(k, v) \rightarrow \mathit{list}(k, v')$, reduce: $(k, \mathit{list}(v')) \rightarrow \mathit{list}(v'')$
    - (c) map: $(k, v) \rightarrow \mathit{list}(k', v')$, reduce: $(k', v') \rightarrow \mathit{list}(v'')$

46. Complete the following map function to compute the relative word frequency across a set of documents with the correct code snippet:

```
map(String key, String value);
int word_count = 0;
```

(a)

```
foreach word w in value do
    EmitIntermediate(w, "1");
    word_count++;
EmitIntermediateToAllReducers(w, AsString(word_count));
```

(b)

```
foreach word w in value do
    EmitIntermediateToAllReducers(w, "1");
    word_count++;
EmitIntermediate(w, AsString(word_count));
```

(c)

```
foreach word w in value do
    EmitIntermediate(w, "1");
    word_count++;
EmitIntermediateToAllReducers("", AsString(word_count));
```

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. Which is the most flexible join pattern in MapReduce?

(a) Reduce side join

(b) Replicated join

(c) Composite join

49. 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

50. 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

51. 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

52. What is true for structured P2P networks with respect to 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) Eager replication without primary copy
    (c) Lazy replication with primary copy
    (d) Lazy replication without primary copy

54. Which of the following consistency levels leads to the best performance in P2P systems?

    (a) Strong consistency
    (b) Eventual consistency
    (c) Weak consistency

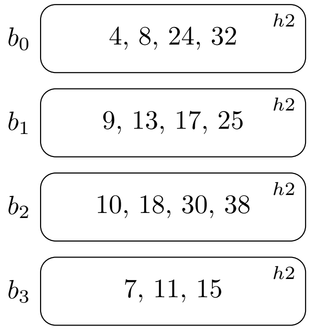
55. What is stored in the client image in the GFS?

    (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 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

57. 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) 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 $p = 1$
    (b) 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 $p = 1$
    (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 $p = 0$

58. 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

59. 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

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) $O(n \log n)$