Advanced Data Management Technologies

Written Exam

30.01.2019

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.

- You need at least 31 points to pass the exam.

Good luck!

Reserved for the Teacher

Max. points	Plus Points	Minus Points	Sum
60			

BI and Multidimensional Modelling

1. Business Intelligence is used by companies
   (a) to improve the efficiency of operational systems
   (b) to reduce the storage costs of the data
   (c) to enable data-based decisions aimed at improving operative performance and incresing profitability
2. To which DW architecture corresponds query-driven data integration?
   (a) Single-layer DW architecture
   (b) Two-layer DW architecture
   (c) Three-layer DW architecture
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 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
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 conceptual model with a graphical notation used for DW design
6. 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
7. 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
8. The granularity of facts determines
   (a) the level of detail for querying the DW
   (b) the measures to be stored
   (c) the aggregation formula used for aggregating measures

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. In the multidimensional model, hierarchies from the root (finest granularity level) to the leaves (lowest granularity level) represent
   (a) many-to-one relationships
   (b) one-to-many relationships
   (c) many-to-many relationships
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. A measure *discount rate* is always
   (a) additive
   (b) non-additive
   (c) semi-additive
15. Which measures are easiest to handle in a DW?
   (a) additive
   (b) semi-additive
   (c) non-additive
16. 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
17. 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

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. What is an advantage of multidimensional OLAP (MOLAP) wrt. relational OLAP (ROLAP)?
    (a) More flexible
    (b) Standards are available
    (c) Faster query response times

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. 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
24. Which of the following techniques helps to tune the load step in the ETL process?
    (a) Sort the data before starting the load process
    (b) Create a small set of views for the most frequent DW queries
    (c) Use SQL-based updates
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 of the following equivalences is wrong?

    (a) `CUBE(a,b)` $\equiv$ `GROUPING SETS ((a,b), (a), (b), ())`

    (b) `GROUP BY GROUPING SETS((a,b,c))` $\equiv$ `GROUP BY a, b, c`

    (c) `GROUP BY GROUPING SETS(a,ROLLUP(b,c))` $\equiv$ `GROUP BY a UNION ALL GROUP BY b, c`

27. How many groupings are produced by the following GROUP BY clause?

```
GROUP BY a, ROLLUP(b, c), GROUPING SETS ((d,e),(f,g),(h)), CUBE(i,j)
```

    (a) 36

    (b) 10

    (c) 11

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. How can a hierarchical data cube be generated in SQL?

    (a) By using a sequence of `ROLLUP(A1,A2,...,Ak)`, `ROLLUP(B1,B2,...,Bk)`, ..., operations, where the attributes `Ai`, `Bi`, ... have to be listed according the dimensional hierarchies

    (b) By using one or more `ROLLUP(A1,A2,...,Ak)`, `ROLLUP(B1,B2,...,Bk)`, ..., operations, where the attributes `Ai`, `Bi`, ... can be specified in any order

    (c) By using a `CUBE(A1,A2,...,Ak)` operation

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 the following window function doing?
`NTILE(4) OVER (ORDER BY SUM(a))`

(a) Divides an ordered partition into buckets of size 4

(b) Divides an ordered partition into 4 equal sized buckets and assigns to each bucket a number

(c) Returns the first 4 tuples of the partition

Pre-Aggregates

32. 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

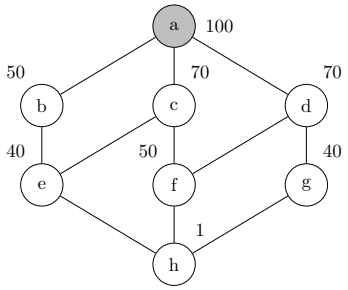
33. Which of the following assumptions in the greedy algorithm is unrealistic?

(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

34. 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, c

(b) b, d

(c) c, d

View Maintenance and Bitmap Indexes

35. Given a materialized view $\mathbf{v} = \pi_{A_1,\ldots,A_k}(\mathbf{r})$ with incremental view maintenance, which maintains tuples of the form $(a_1, \ldots, a_k, c)$. What is the correct code snippet to insert a set of tuples $\mathbf{r}_i$ in view $\mathbf{v}$?

(a)
foreach *tuple* $(a_1, \ldots, a_k) \in \pi_{A_1,\ldots,A_k}(\mathbf{r}_i)$ **do**
  Let c_i be # occurrences of the tuple;
  if $(a_1, \ldots, a_k, c) \in \mathbf{v}$ **then** $c = c + c_i$;
  else Insert $(a_1, \ldots, a_k, c_i)$ into $\mathbf{v}$;

(b)
foreach *tuple* $(a_1, \ldots, a_k) \in \pi_{A_1,\ldots,A_k}(\mathbf{r}_i)$ **do**
  Let c_i be # occurrences of the tuple;
  Let $(a_1, \ldots, a_k, c) \in \mathbf{v}$;
  $c = c + c_i$;

(c)
foreach *tuple* $(a_1, \ldots, a_k) \in \pi_{A_1,\ldots,A_k}(\mathbf{r}_i)$ **do**
  Let c_i be # occurrences of the tuple;
  Remove $(a_1, \ldots, a_k, c)$ from view $\mathbf{v}$;
  Insert $(a_1, \ldots, a_k, c_i)$ into view $\mathbf{v}$;

36. 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)`

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

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

37. The run-length encoding of the bitmap vector `000100100000100` is

    (a) `10010010001`

    (b) `10100010001`

    (c) `10110011001`

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

    (a) `00010010000000010000001`

    (b) `00010010000000010000000`

    (c) `00010010000000011111111`

39. 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

40. Which of the following bitmap-based indices allows to encode dimensional hierarchies?

    (a) Bittmapped join index

    (b) Bit-sliced index

    (c) Bitmap-encoded index

NoSQL and MapReduce

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

    (a) at least 2 of the 3 properties must be satisfied at any time

    (b) at most 2 of the 3 properties can be achieved at any time

    (c) exactly 2 of the 3 properties are satisfied at any time

42. 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

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. 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));
```

45. Which MapReduce program implements the SQL query `SELECT * FROM table WHERE val < x`?

(a)
```
map(key, record) {
  if record.val < x then emit(key, record)
}

reduce(key, records) {
  emit(records)
}
```

(b)
```
map(key, record) {
  if record.val < x then emit(key, record)
}
```

(c)
```
map(key, record) {
  if record.val < x then emit(key, record)
  else emit(key, null)
}
```

46. The DistributedCache in Hadoop can be used

(a) to store and share input splits

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

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

47. 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

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

P2P Networks and Distributed Hash Index

50. What is true in P2P networks?

    (a) Nodes can be both clients and servers

    (b) The network is quite static

    (c) Nodes are controlled by a central authority

51. Which is not a benefit of P2P networks?

    (a) Easy to obtain consistency

    (b) Easy scalability

    (c) High reliability

52. 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

53. Which replication policy in a P2P network avoids that read requests by clients have to wait?

    (a) Eager replication with primary copy

    (b) Eager replication without primary copy

    (c) Lazy replication with primary copy

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

    (a) Strong consistency

    (b) Weak consistency

    (c) Eventual 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 are stored that have been read before

    (c) Information about where the local data is replicated

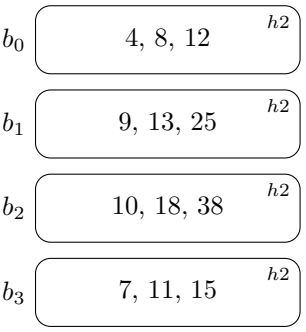
56. A naive solution for a distributed hash-based index is to assign each bucket of the hash file to one of the participating servers and to share the hash function among all nodes. What is a problem of such a solution?

    (a) When a new object is inserted, all nodes need to be informed

    (b) Distributed systems are highly dynamic, i.e., data sets evolve over time and nodes are added/deleted

    (c) If a bucket overflows, the hash value of all objects in that bucket need to be recomputed and then distributed to other buckets

57. 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 $p = 1$

58. In distributed linear hashing, the so-called forward algorithm

    (a) handles bucket overflows by forwarding data to other peers

    (b) copes with lookup errors due to outdated local information

    (c) forwards a lookup request to a central server

59. 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 predecessor or successor node

60. 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