Advanced Data Management Technologies

Written Exam

06.07.2015

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,

  −0.5 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			

1. What is typical for OLAP?
   (a) A complex data model
   (b) The system is always available for updates and reads
   (c) Frequent read operations and infrequent updates
2. What is true for query-driven data integration?
   (a) Query performance is high
   (b) Query is executed on the most up-to-date data
   (c) Query processing does not interfere with the local processing at the data sources.
3. The top-down approach of DW design
   (a) delivers a working system in the short term
   (b) is based on a global picture of the goals
   (c) is more flexible than the bottom-up approach with respect to changing requirements
4. 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
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. Junk dimensions are used to
   (a) store complex hierarchical relationships between dimensional attributes
   (b) store measures that are not available for all facts
   (c) group and store several degenerate dimensions
7. 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 applications
8. The granularity of facts determines
   (a) the level of detail
   (b) the measures to be stored
   (c) the aggregation formula used for aggregating measures

9. Which measures are easiest to handle in a DW?

   (a) additive
   (b) semi-additive
   (c) non-additive

10. 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

11. 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

12. Compared to the star schema, the snowflake schema

    (a) has a better query performance
    (b) requires more joins at query time
    (c) uses more space

13. 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

14. What are the advantages of using dimensions with many attributes?

    (a) Provides more flexibility for data analysis
    (b) Reduces the size of the fact table
    (c) Reduces the number of dimensions

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

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

16. 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

17. 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

18. 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

19. 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) 23.3, 20.0, 30.0, 26.6

(c) 15.0, 20.0, 30.0, 35.0

20. 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

21. 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

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

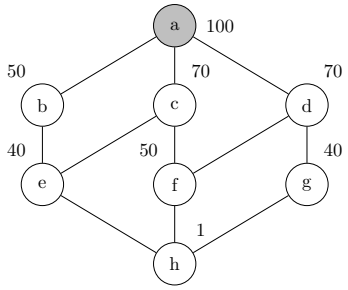
24. The greedy algorithm for pre-aggregate selection

    (a) is 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

25. 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

26. 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) Search original table from the deleted tuple backward

    (b) Scan entire original table

    (c) Do a binary search on the original table

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

    (a) Bitmap index
    (b) Hash index
    (c) B-tree index

28. The compressed bitmap index of 000100100000100 is

    (a) 10010010001
    (b) 10100010001
    (c) 10110011001

29. 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) $n^2 \log_2 n$

30. 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

31. Indices based on bit vectors can be used for

    (a) numeric attributes only
    (b) non-numeric attributes only
    (c) numeric and non-numeric attributes

32. 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

33. What is a good strategy for ETL?

    (a) Implement the transformations in a sequence of small operations/programms
    (b) Implement all transformation in one single programm
    (c) Implement the transformations in the source database

34. 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

35. The data staging area is mainly used for

    (a) querying the DW
    (b) data transformations and cleansing
    (c) indexing dimensions

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

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

37. 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

38. 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

39. 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

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

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

41. 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

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

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

43. In MapReduce, the combiner function can be used to

    (a) combine intermediate tuples from all mappers that have the same key value

    (b) divide up the intermediate key space for parallel reduce operations

    (c) 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

44. 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

45. 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);
```

46. 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) all input records

    (b) $K \cdot M$ records

    (c) K records

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 * FROM table WHERE A = null;
```

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

(a) Reduce side join

(b) Replicated join

(c) Composite join

49. 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

50. 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.

51. 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

52. 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

53. 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

54. 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

55. 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

56. 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)
$a = min(h_n(k), h_{n+1}(k))$;

57. 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

58. 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

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

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

(a) B-tree

(b) T-tree

(c) R-tree