Advanced Data Management Technologies

Written Exam

28.01.2015

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 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 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
4. 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
5. At which granularity level should facts be stored in the multidimensional model?
   (a) finest granularity
   (b) depends on the specific application
   (c) coarsest granularity
6. 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
7. Junk dimensions are used to
   (a) group and store several degenerate dimensions
   (b) store complex hierarchical relationships between dimensional attributes
   (c) store measures that are not available for all facts
8. 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

9. A measure *discount rate* is always

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

10. 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

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

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. How many different groupings are created by `CUBE(a1,...,an)`?

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

15. What is the maximum number of result tuples of the following GROUP BY clause, if the attributes have the following cardinalities: $|a| = 2$, $|b| = 3$, $|c| = 1$, and $|d| = 4$?

```
SELECT   a, b, c, d, COUNT(*)
FROM     r
GROUP BY a, ROLLUP(b, c, d)
```

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

16. 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

17. 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

18. 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

19. Which type of aggregates can be efficiently computed by GMDJ but not in SQL?

    (a) Distributive aggregates

    (b) 1D cumulative aggregates

    (c) 2D cumulative aggregates

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

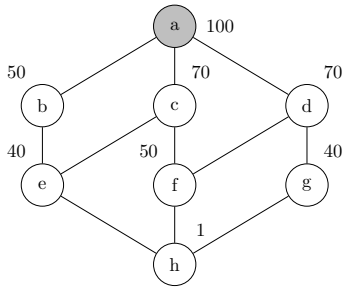
23. The greedy algorithm for pre-aggregate selection

    (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

24. 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

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) Search original table from the deleted tuple backward

    (b) Scan entire original table

    (c) Do a binary search on the original table

26. The compressed bitmap `10110011011` is the run-length encoding of

    (a) `00010010000000010000000`

    (b) `00010010000000010000001`

    (c) `00010010000000011111111`

27. 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

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 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

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. The data staging area is mainly used for

    (a) querying the DW

    (b) data transformations and cleansing

    (c) indexing dimensions

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

    (a) Fact table

    (b) Dimension tables

    (c) Indices

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. 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

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) column stores

    (b) key-value 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: $(k, v) \rightarrow (k, v')^*$, 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, the combiner function can be used to

    (a) combine intermediate tuples with the same key value across all mappers

    (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

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) Numerical summarization

    (b) Simple filtering (which eliminates uninteresting records)

    (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) all input records

    (b) $K \cdot M$ 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 true about unstructured P2P networks?

    (a) Data might not be found even if they are in the network

    (b) The network is inherently unstable

    (c) It is difficult to build and join the network

51. 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

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 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

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 node leaves the network

    (a) the keys of that node are assigned to the node's successor

    (b) the keys are removed

    (c) the keys of that node are re-assigned to nodes using a hash function

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