CS614 QUIZ 1

MADE BY: ARSL SHANI,AINA MALIK & MCS FAMILY GROUP TEAM

1. _____ is known as state of being only one of its kinds or being without an equal or parallel.
   a) Completeness
   b) Uniqueness 185
   c) Accessibility
   d) Consistency

Uniqueness	The state of being only one of its kind or being without an equal or parallel.

2. One of the preconditions to decide about operations to be parallelized is that;
   a) Operations can be implemented independent of each other. Page 201
   b) Output of one operation become input of other
   c) Operations share same memory location
   d) Operations share same namespace

Every operation can not be parallelized, there are some preconditions and one of them being that the operations to be parallelized can be implemented independent of each other. This means that there will be no interference between the operations while they are

3. In context of joining tables, the join condition is specified in _____ clause
   a) FROM
   b) SELECT
   c) WHERE page 239
   d) GROUP BY

- Tables identified by FROM clause and condition by WHERE clause.
- Will cover different types of joins.

Join commands are statements that retrieve data from multiple tables. A join is identified by multiple tables in the FROM clause, and the relationship between the tables to be joined is established through the existence of a join condition in the WHERE clause. Because

4. _____ do not (typically) keep the index values in sorted order

a) Dense index
b) Sparse index
c) B- Tree index
d) Hash based index page 227

In contrast to B-tree indexing, hash based indexes do not (typically) keep index values in sorted order.

5. In _____ index , the ith bit is set to "1" if the ith row of the base table has the value for the indexed column.
   a) Inverted index
   b) Bitmap index page 233
   c) Cluster index
   d) Join index

27.2 Bitmap Indexes: Concept

- Index on a particular column
- Index consists of a number of bit vectors or bitmaps
- Each value in the indexed column has a corresponding bit vector (bitmaps)
- The length of the bit vector is the number of records in the base table
- The *i*th bit is set to 1 if the *i*th row of the base table has the value for the indexed column

6. Parallelism can be exploited , if there is:
   a) Symmetric multi-processors (SMP)
   b) Sufficient I/O bandwidth
   c) Underutilized or intermittently used CPUs
   d) All of the given options page 202

Parallelism can be exploited, if there is…

- Symmetric multi-processors (SMP), clusters, or Massively Parallel (MPP) systems AND
- Sufficient I/O bandwidth AND
- Underutilized or intermittently used CPUs (for example, systems where CPU usage is typically less than 30%) AND
- Sufficient memory to support additional memory-intensive processes such as sorts, hashing, and I/O buffers

7. _____ is the degree to which data accurately reflects the real-world object that the data represents.
    a) Intrinsic Data Quality page 181
    b) Realistic Data Quality
    c) String Data Quality
    d) Weak Data Quality

Intrinsic data quality, simply stated, is data accuracy. Intrinsic data quality is the degree to which data accurately reflects the real-world object that the data represents. If all facts

8. _____ is making all efforts to increase effectiveness and efficiency in meeting accepted customer expectations.
    a) Quality assurance
    b) Quality improvement page 183
    c) Quality maintenance
    d) Quality establishment

Quality improvement is making all efforts to increase effectiveness and efficiency in meeting accepted customer expectations.

9. "More resources means proportionally less time for given amount of data". The statement refers to:
    a) Scale-up
    b) Speed-up page 203
    c) Size-up
    d) Over-utilized system

10. Which of the following is NOT one of the parallel hardware architecture?
    a) Symmetric Multi-Processing
    b) Massively Parallel Processing
    c) Non-uniform Memory Access
    d) Shared Memory page 206

- Parallel Hardware Architectures
    - Symmetric Multi Processing (SMP)
    - Distributed Memory or Massively Parallel Processing (MPP)
    - Non-uniform Memory Access (NUMA)

11. _____ is the extent to which data is regarded as true and credible.

ARSL SHAN, MALIK & MCS FAMILY GROUP TEAM

a) Believability page 188
b) Completeness
c) Accessibility
d) Consistency

Believability is the extent to which data is regarded as true and credible. Among other factors, it may reflect an individual's assessment of the credibility of the data source,

12. The degree to which the values are present in the attributes that require them is known as ____.
a) Completeness page 185
b) Uniqueness
c) Accessibility
d) Consistency

Completeness	The degree to which values are present in the attributes that require them.

13. In B-Tree index, the lowest level index blocks are called leaf blocks, and these blocks contain:
a) NULL value to make the terminal node
b) Every indexed data value and a corresponding ROWID page 225
c) Every indexed data value and pointer to next level block
d) Every indexed data value and pointer to root block

B-tree indexes are stored in a tree structure that has branch blocks which point to lower-level blocks. The lowest level index blocks are called leaf blocks, and these blocks contain every indexed data value and a corresponding ROWID. The ROWID is a pointer

14. The optimizer uses a hash join to join two tables if they are joined using a equijoin and
a) Outer tables has less number of rows
b) inner tables has less number of rows
c) cardinality of tables is equal
d) large amount of data needs to be joined page 245

The optimizer uses a hash join to join two tables if they are joined using an equijoin and if either of the following conditions are true:

- A large amount of data needs to be joined.
- A large portion of the table needs to be joined.

15. As the number of processor increase, the speedup should also increase. Thus we should have linear speedup. Which is the following is one of the barriers to achieve this linear speed-up?
a) AMDAHL' Law page 219
b) Startup

c) No interference
d) Skew

As I said in the first lecture on parallelism, intuitively we feel that as the number of processors increase, the speedup should also increase. Thus theoretically there should be a linear speedup. However, in reality this is not the case. The biggest hurled is the Amdahl's law which we have discussed in detail. The other problem is the startup cost, it

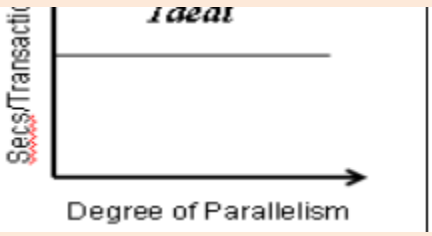
16. Which of the following is not Data Quality Validation Technique?
    a) Referential integrity
    b) Using Data Quality Rules
    c) Data Histograming
    d) Indexes page 189

22.2 Data Quality Validation Techniques

- Referential Integrity (RI).
- Attribute domain.
- Using Data Quality Rules.
- Data Histograming.

17. "If resources increase in proportion to increase in data size, time is constant'. The statement refers to:
    a) Scale-up page 203
    b) Speed-up
    c) Size-up
    d) Over-utilized system

Scale-Up
If resources increased in proportion to increase in data size, time is constant



18. In context of data parallelism, to get a speed-up of N with N partitions, it must be ensured that:
    a) There are enough computing resources
    b) Query coordinator is very fast as compared to query servers
    c) Work done in each position almost same
    d) All of the given option page 213

ARSL SHAHZAD & MALIK & MCS GROUP TEAM

To get a speed-up of N with N partitions, it must be ensured that:

- There are enough computing resources.
- Query-coordinator is very fast as compared to query servers.
- Work done in each partition almost same to avoid performance bottlenecks.
- Same number of records in each partition would not suffice.
- Need to have uniform distribution of records w.r.t. filter criterion across partitions.

19. If a task takes "T' time units to execute on a single data item, then execution of this task on "N" data item will take_____ time units.
    a) N*T page 215
    b) N/T
    c) N+T
    d) N-T

If we take a "big" task that takes T time units to execute on a single data item, then execution of this task on N data items will take N*T time units.

20. Assume a company with a multi-million row customer table i.e. n rows. Checking for the Referential Integrity (RI), using a smart technique with some kind of
    a) O(logn)
    b) O(n) page 190
    c) O(1)
    d) None of the given

implementations. Assume a company with a multi million row customer table i.e. *n* rows. Checking for RI using a naive approach would take *O(n)* time, using a smart technique

21. _____ lists each term in the collection only once and then shows a list of all the documents that contain the given term.
    a) Inverted index page 232
    b) Bitmap index
    c) Cluster index
    d) Join index

document once (and each term repeated for each document that contains the term), an inverted index lists each *term* in the collection only once and then shows a list of all the documents that contain the given term. Each document identifier is repeated for each term

22. An optimized structure which is built primarily for retrieval, with update being only a secondary consideration is

ARSL CHANNA MALIK & MS FAMILY GROUP

a) OLTP
b) OLAP
c) DSS
d) Inverted index page 232

An inverted index is an optimized structure that is built primarily for retrieval, with update being only a secondary consideration. The basic structure *inverts* the text so that

23. If a product meets formally defined "requirement specifications," yet fails to be quality Product from the customer' perspective, this means the requirement is
    a) Defective page 180
    b) Unclear
    c) Unrefined
    d) Un defined

If a product meets formally defined "requirement specifications," yet fails to be a quality product from the customers' perspective, this means the requirements were defective.

24. _____ incorporate the concept of product quality, process control, quality assurance, and quality improvement.
    a) Total Quality Management page 182
    b) Intrinsic Quality Management
    c) Realistic Quality Management
    d) Strong Quality Management

21.4 Total Quality Management (TQM)

- Philosophy of involving all for systematic and continuous improvement.
- It is customer oriented. Why?
- TQM incorporates the concept of product quality, process control, quality assurance, and quality improvement.

25. Which is not a/an Data Quality Validation Technique?
    a) Consistency Integrity page 189
    b) Referential Integrity
    c) Attribute domain
    d) Using Data Quality Rules

ARSLAN MALIK & MCS FAMILY GROUP TEAM

22.2 Data Quality Validation Techniques

- Referential Integrity (RI).
- Attribute domain.
- Using Data Quality Rules.
- Data Histograming.

26. Which of the following is not an "Orr's Law of Data Quality"?

a) "Data that is not used cannot be correct!"
b) "Data quality is a function of its use, not its collection!"
c) "Data will be no better than its most stringent use!"
d) "Data duplication can be harmful for the organization" page 182

21.3 Orr's Laws of Data Quality

Law #1: "Data that is not used cannot be correct!"

© Copyright Virtual University of Pakistan

Law #2: "Data quality is a function of its use, not its collection!"

Law #3: "Data will be no better than its most stringent use!"

Law #4: "Data quality problems increase with the age of the system!"

Law #5: "The less likely something is to occur, the more traumatic it will be when it happens!"

27. One of the main reasons for the failure of DWH deployments is_____.

a) Data quality page 179
b) Data integrity
c) Data duplication
d) Data anomaly

Data is the "fuel" on which a Data Warehouse (DWH) runs. One of the main reasons for the failure of DWH deployments is data quality or lack of it. Unfortunately the problems

28. Most DWH implementations today do not use _____ enforced by the database , but as TQM methods improve overall data quality and data

a) Consistency Integrity
b) Referential Integrity page 190
c) Attribute domain
d) Using Data Quality Rules

The bottom line is, most DWH implementations today do not use RI constraints enforced by the database, but as TQM methods improve overall data quality and database optimizers become more sophisticated in the use of constraints, this will become a more attractive option.

29. _____ refers to "parallelism execution of single data operation across multiple partitions of data"

a) Hardware parallelism
b) Software parallelism
c) Data Parallelism page 212
d) Operational parallelism

So data parallelism is I think the simplest form of parallelization. The idea is that we have parallel execution of single data operation across multiple partitions of data. So the idea

30. In context of bitmap index, the length of the bit vector is:

a) The possible number of domain values in corresponding filed(column)
b) The number of records in the base table page 233
c) The possible number of bitmap tables formed for corresponding field(column)
d) None of the given options

27.2 Bitmap Indexes: Concept

- Index on a particular column
- Index consists of a number of bit vectors or bitmaps
- Each value in the indexed column has a corresponding bit vector (bitmaps)
- The length of the bit vector is the number of records in the base table

31. Which is not a/an characteristics of data quality?

a) Reliability
b) Uniqueness
c) Accessibility
d) Consistency page 186

More on Characteristics of Data Quality

Data Quality Dimensions
Believability
Appropriate Amount of Data
Timeliness
Accessibility
Objectivity
Interpretability
Uniqueness

32. The_____ operator is conservative in that it assigns to the dimension aggregate value no higher than the value of its weakest data quality indicator.
    a) Min page 188
    b) Max
    c) Min and Max
    d) None of the given

quality indicators. The min operator is conservative in that it assigns to the dimension an aggregate value no higher than the value of its weakest data quality indicator (evaluated and normalized to between 0 and 1). The maximum operation is used if a liberal

33. A _____ index , if fits in the memory, costs only one disk I/O access to locate a record given a key.
    a) Dense page 223
    b) Sparse
    c) B-Tree
    d) Hash

A dense index, if fits in the memory, costs only one disk I/O access to locate a record given a key

34. In nested-loop join case, if there are 'M' rows in outer table and 'N ' rows in inner table, time complexity is
    a) O (M log N)
    b) O (log MN)
    c) O (MN)
    d) O (M+N)
35. The _____ operator proves useful in more complex metrics applicable to the dimensions of timeliness and accessibility.
    a) Max page 188
    b) Min

ARSLSHA NINA MALIK CS FAMILY GROUP TEAM

c) Min and Max
d) None of the given

believability and appropriate amount of data. The max operator proves useful in more complex metrics applicable to the dimensions of timeliness and accessibility.

36. The exact formula for speed-up is:
a) (Time on Serial Processor) / (Time on parallel Processor) page 204
b) (Time on Serial Processor) * (Time on parallel Processor)
c) (Time on Serial Processor) + (Time on parallel Processor)
d) (Time on Serial Processor) - (Time on parallel Processor)

$$Speedup = \frac{T_s}{T_m}$$

T_s: Time on serial processor

T_m: Time on multiple processors

37. A join is identified by multiple tables in the _____ clause
a) FROM page 239
b) SELECT
c) GROUP BY
d) SORT BY

Join commands are statements that retrieve data from multiple tables. A join is identified by multiple tables in the FROM clause, and the relationship between the tables to be joined

38. _____ Index uses even less space than _____ index, but the block has to be searched, even for unsuccessful searches.
a) Dense, Sparse
b) Sparse, dense page 224
c) Dense , inverted
d) Sparse, inverted

Sparse Index: Adv & Dis Adv

• Store first value in each block in the sequential file and a pointer to the block.

• Uses even less space than dense index, but the block has to be searched, even for unsuccessful searches.

39. _____ in a data base or data warehouse has no actual value ; it only has potential value.
a) Data page 181
b) Entity
c) Flat tables
d) Data marts

MCS FAMILY GROUP TEAM

equivalent reproduction of something real. Figure 21.1 shows an employee data. Data in a database or data warehouse has no actual value; it only has potential value. Data has a

40. _____ index stores first value in each block in the sequential file and a pointer to the block.
    a) Dense
    b) Sparse page 224
    c) B-Tree
    d) Hash

Sparse Index: Adv & Dis Adv

- Store first value in each block in the sequential file and a pointer to the block.

41. The _____ dimension represents data correctness.
    a) Free-of-error page 187
    b) Completeness
    c) Consistency
    d) Correctness

The free-of-error dimension represents data correctness. If one is counting the data units

42. Which of the following is NOT a variants of Nested-loop join?
    - Nave nested loop join
    - Indexed nested loop join
    - Temporary indexed nested loop join
    - Binary index nested loop join page 243

Nested-Loop Join: Variants

1. Naive nested-loop join
2. Index nested-loop join
3. Temporary index nested-loop join

43. Data is the"----------"on which DWH runs:
    - Fuel page 179
    - Component
    - Entity
    - Element

Data is the "fuel" on which a Data Warehouse (DWH) runs. One of the main reasons for

44. Which of the following is not an activity of Data Quality Analysis project?
    - "Define"
    - "Measure"
    - "Analyze"
    - "Compression" page 194

3. **Implement Data Quality Projects (Define, Measure, Analyze, Improve)**

45. Two interesting examples of Quality Dimensions that can make use of the min operator----------
    - Believability and appropriate amount of data page 188
    - Believability and Consistency
    - Believability and Redundancy
    - Reliability and appropriate amount of data

ratio. Two interesting examples of dimensions that can make use of the min operator are believability and appropriate amount of data. The max operator proves useful in more

46. The extent to which data is appropriate languages, symbols, and units and the definitions are clear is known as----------.
    - Interpretability page 185
    - Uniqueness
    - Accessibility
    - Consistency

Interpretability	The extent to which data is in appropriate languages, symbols, and units, and the definitions are clear.

47. In context of data parallelism, the work done by query processor should be:
    - Almost zero page 213
    - Maximum
    - Pipelined
    - Filtered across partitions

The first thing is that the amount of work done by the query coordinator should be very small. It should be small because that work is serial. And if it is serial it will kill the performance. So it should be almost zero.

48. Bitmap index is appropriate for:
    - Low cardinality data page 236
    - High cardinality data
    - Clustered data
    - Aggregated data

ARSL SHANAWAZ MALIK & MUSTAFA GROUP TEAM

Bitmap Index: Dis. Adv.

- Locking of many rows
- Low cardinality

49. In context of nested-loop join. actual number of matching rows returned as a result of the join would be ____ of the order of tables.

- Dependent
- Independent page#241
- Superset
- Subset

It is true that the actual number of matching rows returned as a result of the join would be independent of the order of tables i.e. inner or outer; actually it is going to be even independent of the type of join used. Now assume that depending on the filter criterion

50. If every key in the data file is represented in the index file then index is

- Dense Index page#223
- Sparse Index
- Inverted Index
- None

Dense Index: Every key in the data file is represented in the index file

51. The goal of ____is to look at as few block as possible to find the matching records.

- Indexing page#222
- ETL
- Joining
- Dimensional Modeling

Indexing Goal

Look at as few blocks as possible to find the matching record(s)

52. The goal of ___ is to look at as few blocks as possible to find matching records(s).

- Indexing page 222

Indexing Goal

Look at as few blocks as possible to find the matching record(s)

ASIF SHAHZAD MALIK & CS FAMILY GROUP TEAM