FINALTERM EXAMINATION
Spring 2009
CS614- Data Warehousing

(VISIT VURANK FOR MORE)

Marks: 70

Question No: 1 (Marks: 1) - Please choose one
It is observed that every year the amount of data recorded in an organization is
Doubles (handouts page # 6)
Triples
Quartiles
Remains same as previous year

Question No: 2 (Marks: 1) - Please choose one
Multidimensional databases typically use proprietary _____ format to store pre-summarized cube structures.
File (Page # 69)
Application
Aggregate
Database

Question No: 3 (Marks: 1) - Please choose one
Pre-computed _____ can solve performance problems
Aggregates (page # 101)
Facts
Dimensions

Question No: 4 (Marks: 1) - Please choose one
_____, if fits into memory, costs only one disk I/O access to locate a record by given key.
A Dense Index (page # 211)
A Sparse Index
An Inverted Index
None of These

Question No: 5 (Marks: 1) - Please choose one
The degree of similarity between two records, often measured by a numerical value between _____, usually depends on application characteristics.
0 and 1 (page # 157)
0 and 10
0 and 100
0 and 99

Question No: 6 (Marks: 1) - Please choose one
The purpose of the House of Quality technique is to reduce _____ types of risk.
Two (page # 181)
Three
Four
All

Question No: 7 (Marks: 1) - Please choose one
NUMA stands for _____
Non-uniform Memory Access (page # 194)
Non-updateable Memory Architecture
New Universal Memory Architecture

Question No: 8 (Marks: 1) - Please choose one
Which is the least appropriate join operation for Pipeline parallelism?
Hash Join
Inner Join
Outer Join
Sort-Merge Join
Question No: 9 (Marks: 1) - Please choose one
There are many variants of the traditional nested-loop join. If the index is built as part of the query plan and subsequently dropped, it is called
Naive nested-loop join
Index nested-loop join
Temporary index nested-loop join (page # 230)
None of these
Question No: 10 (Marks: 1) - Please choose one
Data mining derives its name from the similarities between searching for valuable business information in a large database, for example, finding linked products in gigabytes of store scanner data, and mining a mountain for a _____ of valuable ore.
Furrow
Streak
Trough
Vein

Question No: 11 (Marks: 1) - Please choose one
With data mining, the best way to accomplish this is by setting aside some of your data in a _____ to isolate it from the mining process; once the mining is complete, the results can be tested against the isolated data to confirm the model's validity.
Cell
Disk
Folder
Vault

Question No: 12 (Marks: 1) - Please choose one
The Kimball s iterative data warehouse development approach drew on decades of experience to develop the _____.
Business Dimensional Lifecycle (page # 276)
Data Warehouse Dimension
Business Definition Lifecycle
OLAP Dimension

Question No: 13 (Marks: 1) - Please choose one
We must try to find the one access tool that will handle all the needs of their users.
True
False

Question No: 14 (Marks: 1) - Please choose one
For a smooth DWH implementation we must be a technologist.
True
False (page # 306)

Question No: 15 (Marks: 1) - Please choose one
During the application specification activity, we also must give consideration to the organization of the applications.
True (page # 294)
False

Question No: 16 (Marks: 1) - Please choose one
Investing years in architecture and forgetting the primary purpose of solving business problems, results in inefficient application. This is the example of _____ mistake.
Extreme Technology Design
Extreme Architecture Design
None of these (page # 303)

Question No: 17 (Marks: 1) - Please choose one
The most recent attack is the _____ attack on the cotton crop during 2003-04, resulting in a loss of nearly 0.5 million bales.
Boll Worm (VIDO LECTURE # 38)
Purple Worm
Blue Worm
Cotton Worm

Question No: 18 (Marks: 1) - Please choose one
The users of data warehouse are knowledge workers in other words they are _____ in the organization.
Decision maker (page# 10)

Manager
Database Administrator
DWH Analyst

Question No: 19 (Marks: 1) - Please choose one
_____ breaks a table into multiple tables based upon common column values.
Horizontal splitting (page # 46)
Vertical splitting

Question No: 20 (Marks: 1) - Please choose one
Execution can be completed successfully or it may be stopped due to some error. In case of successful completion of execution all the transactions will be _____
Committed to the database (page # 398 last line)
Rolled back

Question No: 21 (Marks: 2)
What is meant by the statement Be a diplomat NOT a technologist in the context of a data warehouse development project?
7. Be a diplomat NOT a technologist
The biggest problem you will face during a warehouse implementation will be people, not the technology or the development. You're going to have senior management complaining about completion dates and unclear objectives. You're going to have development people protesting that everything takes too long and why can't they do it the old way? You're going to have users with outrageously unrealistic expectations, who are used to systems that require mouse-clicking but not much intellectual investment on their part. And you're going to grow exhausted, separating out Needs from Wants at all levels. Commit from the outset to work very hard at communicating the realities, encouraging investment, and cultivating the development of new skills in your team and your users (and even your bosses).

Question No: 22 (Marks: 2)
Elaborate the concept of data parallelism.
Parallel execution ♣ of a single data manipulation task across multiple partitions of data.
Partitions static ♣ or dynamic
Tasks executed ♣ almost-independently across partitions.
"Query coordinator" ♣ must coordinate between the independently executing processes.
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 here is that these partitions of data may be defined statically or dynamically fine, but we are requiring the same operator across these multiple partitions concurrently. And this idea actually of data parallelism has existed for a very long time.

Question No: 23 (Marks: 2)
What will be the effect if we program a package by using DTS object model?

Question No: 24 (Marks: 3)
What is meant by the classification process? How we measure the accuracy of classifiers?
Classification means that based on the properties of existing data, we have made or groups i.e. we have made classification.

Question No: 25 (Marks: 3)
How page dimension captures the static and dynamic nature of different web pages?
Question No: 26 (Marks: 3)
Write down the limitations of pipelining parallelism?
Pipeline parallelism is a good fit for data warehousing (where we are working with lots of data), but it makes no sense for OLTP because OLTP tasks are not big enough to justify breaking them down into subtasks.

Question No: 27 (Marks: 5)
For a maximum performance of Bitmapped index, what characteristics a query should have?

Question No: 28 (Marks: 5)
How the three parallel tracks capture the user requirements in the Kimball s data warehouse life cycle Road Map?

Question No: 29 (Marks: 5)
How time contiguous log entries and HTTP secure socket layer are used for user session identification? What are the limitations of these techniques?

Question No: 30 (Marks: 10)
What are the issues regarding the record management tools at campuses where text files are used to store data?
Main issues
Data duplication
Update the data
Data deletion
We can easily elaborate these issues

Question No: 31 (Marks: 10)
Shared RDBMS architecture requires a static partitioning. How do you perform the partitioning.

FINALTERM EXAMINATION

Spring 2010

CS614- Data Warehousing (Session - 3)

Time: 90 min
Marks: 60

Question No: 1 (Marks: 1) - Please choose one
A data warehouse may include
- ► **Legacy systems**
- ► Only internal data sources
- ► Privacy restrictions
- ► Small data mart

Question No: 2 (Marks: 1) - Please choose one
De-Normalization normally speeds up
- ► **Data Retrieval**
- ► Data Modification
- ► Development Cycle
- ► Data Replication

Question No: 3 (Marks: 1) - Please choose one
In horizontal splitting, we split a relation into multiple tables on the basis of
- ► **Common Column Values**
- ► Common Row Values
- ► Different Index Values
- ► Value resulted by ad-hoc query

Question No: 4 (Marks: 1) - Please choose one
Multidimensional databases typically use proprietary _____ format to store pre-summarized cube structures.
- ► **File**
- ► Application
- ► Aggregate
- ► Database

Question No: 5 (Marks: 1) - Please choose one
A dense index, if fits into memory, costs only _____ disk I/O access to locate a record by given key.
- ► **One**
- ► Two
- ► lg (n)
- ► n

Question No: 6 (Marks: 1) - Please choose one
All data is _____ of something real.

I**An Abstraction**

IIA Representation
Which of the following option is true?
- ► I Only

► II Only
► **Both I & II (P# 181)**
► None of I & II

Question No: 7 (Marks: 1) - Please choose one
The key idea behind _____ is to take a big task and break it into subtasks that can be processed concurrently on a stream of data inputs in multiple, overlapping stages of execution.
► **Pipeline Parallelism**
► Overlapped Parallelism
► Massive Parallelism
► Distributed Parallelism

Question No: 8 (Marks: 1) - Please choose one
Non uniform distribution, when the data is distributed across the processors, is called _____.
► **Skew in Partition (P # 218)**
► Pipeline Distribution
► **Distributed Distribution**
► Uncontrolled Distribution

Question No: 9 (Marks: 1) - Please choose one
The goal of ideal parallel execution is to completely parallelize those parts of a computation that are not constrained by data dependencies. The smaller the portion of the program that must be executed _____, the greater the scalability of the computation.
► None of these
► **Sequentially**
► In Parallel
► Distributed

Question No: 10 (Marks: 1) - Please choose one
If 'M' rows from table-A match the conditions in the query then table-B is accessed 'M' times. Suppose table-B has an index on the join column. If 'a' I/Os are required to read the data block for each scan and 'b' I/Os for each data block then the total cost of accessing table-B is _____ logical I/Os approximately.
► **(a + b)M**
► (a - b)M
► (a + b + M)
► (a * b * M)

Question No: 11 (Marks: 1) - Please choose one
Data mining is a/an _____ approach, where browsing through data using data mining techniques may reveal something that might be of interest to the user as information that was unknown previously.
► **Exploratory**
► Non-Exploratory
► Computer Science

Question No: 12 (Marks: 1) - Please choose one
Data mining evolve as a mechanism to cater the limitations of _____ systems to deal massive data sets with high dimensionality, new data types, multiple heterogeneous data resources etc.
► **OLTP**
► OLAP
► DSS
► DWH

Question No: 13 (Marks: 1) - Please choose one

_____ is the technique in which existing heterogeneous segments are reshuffled, relocated into homogeneous segments.

► **Clustering**
► Aggregation
► Segmentation
► Partitioning

Question No: 14 (Marks: 1) - Please choose one

To measure or quantify the similarity or dissimilarity, different techniques are available. Which of the following option represent the name of available techniques?

► Pearson correlation is the only technique
► Euclidean distance is the only technique
► **Both Pearson correlation and Euclidean distance**
► None of these

Question No: 15 (Marks: 1) - Please choose one

For a given data set, to get a global view in un-supervised learning we use

► **One-way Clustering (P# 271)**
► Bi-clustering
► Pearson correlation
► Euclidean distance

Question No: 16 (Marks: 1) - Please choose one

In DWH project, it is assured that _____ environment is similar to the production environment

► Designing
► **Development**
► Analysis
► Implementation

Question No: 17 (Marks: 1) - Please choose one

For a DWH project, the key requirement are _____ and product experience.

► Tools
► **Industry (P# 320)**
► Software
► **None of these**

Question No: 18 (Marks: 1) - Please choose one

Pipeline parallelism focuses on increasing throughput of task execution, NOT on _____ sub-task execution time.

► **Increasing**

► **Decreasing (P# 215)**
► Maintaining
► None of these

Question No: 19 (Marks: 1) - Please choose one

Many data warehouse project teams waste enormous amounts of time searching in vain for a _____.

► **Silver Bullet**
► Golden Bullet
► Suitable Hardware
► Compatible Product

Question No: 20 (Marks: 1) - Please choose one
Focusing on data warehouse delivery only often end up _____.

► **Rebuilding**
► Success
► Good Stable Product
► None of these

Question No: 21 (Marks: 1) - Please choose one
Pakistan is one of the five major _____ countries in the world.

► **Cotton-growing**
► Rice-growing
► Weapon Producing

Question No: 22 (Marks: 1) - Please choose one
_____ is a process which involves gathering of information about column through execution of certain queries with intention to identify erroneous records.

► **Data profiling (P# 439)**

► **Data Anomaly Detection**

► Record Duplicate Detection
► None of these

Question No: 23 (Marks: 1) - Please choose one
Relational databases allow you to navigate the data in _____ that is appropriate using the primary, foreign key structure within the data model.

► Only One Direction
► **Any Direction**
► Two Direction
► None of these

Question No: 24 (Marks: 1) - Please choose one
DSS queries do not involve a primary key

► **True**
► False

Question No: 25 (Marks: 1) - Please choose one
_____ contributes to an under-utilization of valuable and expensive historical data, and inevitably results in a limited capability to provide decision support and analysis.

► **The lack of data integration and standardization (P# 330)**
► Missing Data
► **Data Stored in Heterogeneous Sources**

Question No: 26 (Marks: 1) - Please choose one

DTS allows us to connect through any data source or destination that is supported by _____

► **OLE DB**
► OLAP
► OLTP
► Data Warehouse

Question No: 27 (Marks: 1) - Please choose one

Data Transformation Services (DTS) provide a set of _____ that lets you extract, transform, and consolidate data from disparate sources into single or multiple destinations supported by DTS connectivity.

► **Tools**
► Documentations
► Guidelines

Question No: 28 (Marks: 1) - Please choose one

Execution can be completed successfully or it may be stopped due to some error. In case of successful completion of execution all the transactions will be _____

► **Committed to the database**
► Rolled back

Question No: 29 (Marks: 1) - Please choose one

If some error occurs, execution will be terminated abnormally and all transactions will be rolled back. In this case when we will access the database we will find it in the state that was before the _____.

► **Execution of package**
► Creation of package
► Connection of package

Question No: 30 (Marks: 1) - Please choose one

To judge effectiveness we perform data profiling twice.

► One before Extraction and the other after Extraction
► **One before Transformation and the other after Transformation**
► One before Loading and the other after Loading

Question No: 31 (Marks: 2)

What are the two extremes for technical architecture design? Which one is better?

Theoretically there can be two extremes i.e. free space and free performance. If storage is not an issue, then just pre-compute every cube at every unique combination of dimensions at every level as it does not cost anything. This will result in maximum query performance. But in reality, this implies huge cost in disk space and the time for constructing the pre-aggregates. In the other case where performance is free i.e. infinitely fast machines and infinite number of them, then there is not need to build any summaries. Meaning zero cube space and zero pre-calculations, and in reality this would result in minimum performance boost, in the presence of infinite performance.

Question No: 32 (Marks: 2)

What is value validation process?

Value validation is the process of ensuring that each value that is sent to the data warehouse is accurate.

Question No: 33 (Marks: 2)

What is the difference between training data and test data?

Question No: 34 (Marks: 2)
Do you think it will create the problem of non-standardized attributes, if one source uses 0/1 and second source uses 1/0 to store male/female attribute respectively? Give a reason to support your answer.

Question No: 35 (Marks: 3)
Why building a data warehouse is a challenging activity? What are the three broad categories of data warehouse development methods?

1. Waterfall model
2. RAD model
3. Spiral Model

Question No: 36 (Marks: 3)
What are three fundamental reasons for warehousing Web data?
1. Web data is unstructured and dynamic, Keyword search is insufficient.
2. Web log contain wealth of information as it is a key touch point.
3. Shift from distribution platform to a general communication platform.

Question No: 37 (Marks: 3)
What types of operations are provided by MS DTS?

1. Providing connectivity to different databases
2. Building query graphically
3. Extraction data from disparate databases
4. Transforming data
5. Copying database objects
6. Providing support of different scripting languages (by default VB-script and Java –

Question No: 38 (Marks: 3)
What problems may be faced during Change Data Capture (CDC) while reading a log/journal tape?
Problems with reading a log/journal tape are many:

1. Contains lot of extraneous data
2. Format is often arcane
3. Often contains addresses instead of data values and keys
4. Sequencing of data in the log tape often has deep and complex

5. implications

6. Log tape varies widely from one DBMS to another.

Question No: 39 (Marks: 5)

What are seven steps for extracting data using the SQL server DTS wizard?

SQL Server Data Transformation Services (DTS) is a set of graphical tools and programmable objects that allow you extract, transform, and consolidate data from disparate sources into single or multiple destinations. SQL Server Enterprise .Manager provides an easy access to the tools of DTS.

Question No: 40 (Marks: 5)

Explain Analytic Applications Development Phase of Analytic Applications Track of Kimball's Model?

Ans:

The DWH development lifecycle (Kimball's Approach) has three parallel tracks emanating from requirements definition. These are

1. technology track,
2. data track and
3. Analytic applications track.

Analytic Applications Track:

Analytic applications also serve to encapsulate the analytic expertise of the organization, providing a jump-start for the less analytically inclined. It consists of two phases.

1. Analytic applications specification
2. Analytic applications development

Analytic applications specification:

The main features of Analytic applications specification are:

- Starter set of 10-15 applications.
- Prioritize and narrow to critical capabilities.
- Single template use to get 15 applications.
- Set standards: Menu, O/P, look feel.
- From standard: Template, layout, I/P variables, calculations.
- Common understanding between business & IT users.

Following the business requirements definition, we need to review the findings and collected sample reports to identify a starter set of approximately 10 to 15 analytic applications. We want to narrow our initial focus to the most critical capabilities so that we can manage expectations and ensure on-time delivery. Business community input will be critical to this prioritization process. While 15 applications may not sound like much,
Before designing the initial applications, it's important to establish standards for the applications, such as

- common pull-down menus and
- Consistent output look and feel.

Using the standards, we specify each application

- template,
- capturing sufficient Information about the layout,
- input variables,
- calculations, and
- breaks

so that both the application developer and business representatives share a common understanding. During the application specification activity, we also must give consideration to the organization of the applications. We need to identify structured navigational paths to access the applications, reflecting the way users think about their business. Leveraging the Web and customizable information portals are the dominant strategies for disseminating application access.
Analytic applications development:
The main features of Analytic applications development consisits of:

1. Standards: naming, coding, libraries etc.
2. Coding begins AFTER DB design complete, data access tools installed,

subset of historical data loaded.

3. Tools: Product specific high performance tricks, invest in tool-specific

education.

4. Benefits: Quality problems will be found with tool usage => staging.
5. Actual performance and time gauged.

When we do work into the development phase for the analytic applications, we again need to focus on standards. Standards for

- naming conventions,

- calculations,
- libraries, and
- coding

should be established to minimize future rework. The application development activity can begin once the database design is complete, the data access tools and metadata are installed, and a subset of historical data has been loaded. The application template specifications should be revisited to account for the inevitable changes to the data model since the specifications were completed.
We should take approperiate-specific education or supplemental resources for the development team.
While the applications are being developed, several ancillary benefits result. Application developers, should have a robust data access tool, quickly will find needling problems in the data haystack despite the quality assurance performed by the staging application. we need to allow time in the schedule to address any flaws identified by the analytic applications.
After realistically test query response times developers now reviewing performance-tuning strategies. The application development quality-assurance activities cannot be completed until the data is stabilized. We need to make sure that there is adequate time in the schedule beyond the final data staging cutoff to allow for an orderly wrap-up of the application development tasks.

(GOOD LUCK FOR FINAL EXAM)

(KEEP VISITING VURANK FOR MORE)