

Come & Join Us at VUSTUDENTS.net

For Assignment Solution, GDB, Online Quizzes, Helping Study material,
Past Solved Papers, Solved MCQs, **Current Papers**, E-Books & more.

Go to <http://www.vustudents.net> and click **Sing up to register.**



<http://www.vustudents.net>

VUSTUENTS.NET is a community formed to overcome the disadvantages of distant learning and virtual environment, where pupils don't have any formal contact with their mentors, This community provides its members with the solution to current as well as the past Assignments, Quizzes, GDBs, and Papers. This community also facilitates its members in resolving the issues regarding subject and university matters, by providing text e-books, notes, and helpful conversations in chat room as well as study groups. Only members are privileged with the right to access all the material, so if you are not a member yet, kindly SIGN UP to get access to the resources of VUSTUDENTS.NET

» » Regards » »

VUSTUDENTS.NET TEAM.

Virtual University of Pakistan

Come and Join Us at www.vustudents.ning.com

i) K-mean weakness

Similar to other algorithm, K-mean clustering has many weaknesses:

When the numbers of data are not so many, initial grouping will determine the cluster significantly.

The number of cluster, K, must be determined before hand.

We never know the real cluster, using the same data, if it is inputted in a different order may produce different cluster if the number of data is a few.

Sensitive to initial condition. Different initial condition may produce different result of cluster. The algorithm may be trapped in the local optimum.

We never know which attribute contributes more to the grouping process since we assume that each attribute has the same weight.

weakness of arithmetic mean is not robust to outliers. Very far data from the centroid may pull the centroid away from the real one.

The result is circular cluster shape because based on distance.

One way to overcome those weaknesses is to use K-mean clustering only if there are available many data. To overcome outliers problem, we can use median instead of mean.

Some people pointed out that K means clustering cannot be used for other type of data rather than quantitative data. This is not true! See how you can use multivariate data up to n dimensions (even mixed data type) here. The key to use other type of dissimilarity is in the distance matrix.

ii) Classification process and its accuracy?

Accuracy: Accuracy is the measure of correctness of your model e.g. in classification we have two data sets, training and test sets. A classification model is built based on the data properties and relationships in training data. Once built the model is tested for accuracy in terms of % correct results as the classification of the test data is already known. So we specify the correctness or confidence level of the technique in terms % accuracy.

iii) Diff b/w data matrix & similarity/dissimilarity matrix

Similarity or dissimilarity matrix is the measure the similarity or dissimilarity obtained by pair wise comparison of rows. First of all you measure the similarity of the row1 in data matrix with itself that will be 1. So 1 is placed at index 1, 1 of the similarity matrix. Then you compare row 1 with row 2 and the measure or similarity value goes at index 1, 2 of the similarity matrix and son. In this way the similarity matrix is filled. It should be noted that the similarity between row1 and row2 will be same as between row 2 and 1. Obviously, the similarity matrix will then be a square matrix, symmetric and all values along the diagonal will be same (here 1). So if your data matrix has n rows and m columns then your similarity matrix will have n rows and n columns. What will be the time complexity of computing similarity/dissimilarity matrix? It will be $O(n^2)$ (m), where m accounts for the vector or header size of the data. Now how to measure or quantify the similarity or dissimilarity? Different techniques available like Pearson correlation and Euclidean distance etc..

iv) Name of authority to pest

v) Do you think it will create the problem of non-standardized attributes,

The second problem is non standardized attributes across campuses. While looking at the header of data from different campuses we came to know the following problems regarding attributes and is summarized in the table

in the slide.

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.

We prefer to go for later option and drop all automatic mappings because optimizer does not transform genders (from 0/1 to M/F) or names etc. As source data is distributed over different campuses therefore the issues like difference in date formats, conventions of storing gender (M/F, 0/1, 1/0), etc are obvious. Microsoft SQL Server has a good support to resolve these issues.

vi) Inverted index

Inverted index: Concept

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 instead of the view obtained from scanning documents where a document is found and then its terms are seen (think of a list of documents each pointing to a list of terms it contains), an index is built that maps terms to documents (pretty much like the index found in the back of a book that maps terms to page numbers). Instead of listing each 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 that is found in the document. Within the search engine domain, data are searched far more frequently than they are updated. This is typical for a data warehouse, where updates hardly take place. Given this situation a data structure called an inverted index commonly used by search engines is also applicable for the data warehouse environment. An inverted index is able to do many accesses in $O(1)$ time at the price of significantly longer time to do an update, in the worst case $O(n)$. Index construction time is longer as well, but query time is generally faster than with a B-tree i.e. $O(\log n)$. Since index construction is an off-line activity, so it is an appropriate tradeoff i.e. shorter query times at the expense of lengthier index construction times.

Finally, inverted index storage structures can exceed the storage demands of the document collection itself. However, the inverted index can be compressed for many systems, to around 10% of the original document collection. Given the alternative (of 26 minute searches), search engine developers are happy to trade index construction time and storage for query efficiency. Same is also true for the DSS environment.

vii) Diff b/w knowledge Discovery, data mining and DWH

How Data Mining is different?

- „ Knowledge Discovery
 - Overall process of discovering useful knowledge
- „ Data Mining (Knowledge-driven exploration)
 - Query formulation problem.
 - Visualize and understand of a large data set.
 - Data growth rate too high to be handled manually.
- „ Data Warehouses (Data-driven exploration):

Querying summaries of transactions, etc. Decision support DWH is the process of bringing input data in a form that can readily be used by data mining techniques to find hidden patterns. Both terms KDD and DM are sometimes used to refer to the same thing but KDD refers to the overall process from data extraction from legacy source systems, data preprocessing, DWH building, data mining and finally the output generation. So KDD is a mega process having sub processes like DWH and DM being its constituent parts.

viii) Why DASD is better than tape storage w.r.t access time

DASD (Direct Access Storage Device).

CS614 solved all current papers BY JAMIA

Disk storage was fundamentally different from magnetic tape storage in the sense that data could be accessed directly on DASD i.e. non-sequentially. There was no need to go all the way through records 1, 2, 3, . . . k so as to reach the record k + 1. Once the address of record k + 1 was known, it was a simple matter to go to record k + 1 directly.

Furthermore, the time required to go to record k + 1 was significantly less than the time required to scan a magnetic tape. Actually it took milliseconds to locate a record on a DASD i.e. orders of magnitude better performance than the magnetic tape. With DASD came a new type of system software known as a DBMS (Data Base Management System). The purpose of the DBMS was to facilitate the programmer to store and access data on DASD. In addition, the DBMS took care of such tasks as storing data on DASD, indexing data, accessing it etc. With the winning combination of DASD and DBMS came a technological solution to the problems of magnetic tape based master files. When we look back at the mess that was created by master files and the mountains of redundant data aggregated on them, it is no wonder that database is defined as a single source of data for all processing and a prelude to a data warehouse i.e. "a single source of truth".

ix) Transient cookie & persistent cookie

On the Web, a transient cookie, sometimes called a session cookie, is a small file that contains information about a user that disappears when the user's browser is closed. Unlike a persistent cookie, a transient cookie is not stored on your hard drive but is only stored in temporary memory that is erased when the browser is closed. A transient cookie is created by simply not setting a date in the Set-Cookie option when an application creates the cookie. (For a persistent cookie, an expiration date is set and the cookie is stored on the user's hard drive until the expiration date or until the user deletes it.)

Transient cookies are often used to enable a site to be able to track the pages that a user has visited during a visit so that information can be customized for the user in some way. Some sites use Secure Sockets Layer (SSL) to encrypt the information contained in a cookie.

Also called a permanent cookie, or a stored cookie, a cookie that is stored on a user's hard drive until it expires (persistent cookies are set with expiration dates) or until the user deletes the cookie. Persistent cookies are used to collect identifying information about the user, such as Web surfing behavior or user preferences for a specific Web site.

x) How Business validation rule implemented in DTS?

Two business rules are required to be validated here

- All new registrations are done in month of August before 28
- Transfer cases can also be dealt in January
- At the time of registration for BS age must be greater than 16 years and for MS age must be greater than 20 years

xi) click stream

40.2 Clickstream

Clickstream is every page event recorded by each of the company's Web servers

f Although most exciting, at the same time it can be the most difficult and most frustrating. f Not JUST another data source. Web-intensive businesses have access to a new kind of data, in some cases literally consisting of the gestures of every Web site visitor. This is called as the click stream. In its most elemental form, the clickstream is every page event recorded by the web server. The clickstream contains a number of new dimensions such as page, session, and referrer-that were previously unknown in conventional DWH environment.

The clickstream is a stream of data, easily being the largest text and number data set we have ever considered for a data warehouse. Although the clickstream is the most exciting new development in data warehousing, at the same time it can be the most difficult and most frustrating to handle and process. The clickstream is not just another data source that is extracted, cleaned, and dumped into the data warehouse. It is an evolving collection of data sources having more than a dozen Web server log file formats for capturing clickstream data. These formats have optional data components that, if used, can be very helpful in identifying visitors, sessions, and the true meaning of behavior.

xii) Data profiling is a process of gathering information about columns,

CS614 solved all current papers BY JAMIA

Data profiling is a process of gathering information about columns, It must fulfil the following purposes

- Identify the type and extent to which the transformation is required
- The number of columns which are required to be transformed and which transformation is required, meaning date format or gender convention.
- It should provide us a detailed view about the quality of data. The number of Erroneous values and the number of values out of domain. To judge effectiveness of transformation we perform data profiling twice. One before transformation and the other after transformation.

what are the purpose that it must fulfill? Describe briefly

The server returns the requested page, webpage.html. Once the document is entirely retrieved, the visitor's browser scans for references to other Web documents that it must fulfill before its work is completed. In order to speed up the response time, most browsers will execute these consequential actions in parallel, typically with up to 4 or more HTTP requests being serviced concurrently.

1. Difference b/w MOLAP and DOLAP implementation 2marks

OLAP Implementations

1. MOLAP: OLAP implemented with a multi-dimensional data structure.
2. ROLAP: OLAP implemented with a relational database.
3. HOLAP: OLAP implemented as a hybrid of MOLAP and ROLAP.
4. DOLAP: OLAP implemented for desktop decision support environments.

MOLAP physically builds "cubes" for direct access - usually in the proprietary file format of a multi-dimensional database (MDD) or a user defined data structure. therefore ANSI SQL is not supported. ROLAP or a Relational OLAP provides access to information via a relational database using ANSI standard SQL.

HOLAP provides a combination of relational database access and "cube" data structures within a single framework. The goal is to get the best of both MOLAP and ROLAP: scalability (via relational structures) and high performance (via pre-built cubes).

DOLAP allows download of "cube" structures to a desktop platform without the need for shared relational or cube server. This model facilitates a mobile computing paradigm.

DOLAP is particularly useful for sales force automation types of applications by supporting extensive slide and dice. A DOLAP implementation needs to be much more efficient in disk and memory utilization than typical server implementations because computing power is often limited on a laptop computer.

2. What are three methods for creating a DTS package? 2marks

A DTS package is an organized collection of connections, DTS tasks, DTS transformations, and workflow constraints assembled either with a DTS tool or programmatically and saved to Microsoft® SQL Server™, SQL Server 2000 Meta Data Services, a structured storage file, or a Microsoft Visual Basic® file. Each package contains one or more steps that are executed sequentially or in parallel when the package is run. When executed, the package connects to the correct data sources, copies data and database objects, transforms data, and notifies other users

4. Difference b/w classification and clustering 2 marks

<http://www.vustudents.net>

30.1 CLASSIFICATION

f Classification consists of examining the properties of a newly presented observation and assigning it to a predefined class.

- f Assigning customers to predefined customer segments (good vs. bad)
- f Assigning keywords to articles
- f Classifying credit applicants as low, medium, or high risk
- f Classifying instructor rating as excellent, very good, good, fair, or poor

Classification means that based on the properties of existing data, we have made or groups i.e. we have made classification. The concept can be well understood by a very simple example of student grouping. A student can be grouped either as good or bad depending on his previous record. Similarly an employee can be grouped as excellent,

CS614 solved all current papers BY JAMIA

good, fair etc based on his track record in the organization. So how students or employees were classified? Answer is using the historical data. Yes history is the best predictor of the future. When an organization conducts test and interviews from candidate employees, their performance is compared with those of the existing employees. The knowledge can be used to predict how good you can perform if employed. So we are doing classification, here absolute classification i.e. either good or bad or in other words we are doing binary classification. Either you are in this group or this. Each entity is assigned one of the groups or classes. An example where classification can prove to be beneficial is in customer segmentation. The businesses can classify their customers as either good or bad; the knowledge thus can be utilized for executing targeted marketing plans.

30.5 CLUSTERING

Task of segmenting a heterogeneous population into a number of more homogenous sub-groups or clusters.

Unlike classification, it does NOT depend on predefined classes. It is up to you to determine what meaning, if any, to attached to resulting clusters. Rules:

{Milk} $\not\in$ {Cola}

{Diaper, Milk} \in {Juice}

It could be the first step to the market segmentation effort.

What else data mining can do? We can do clustering with DM. Clustering is the technique of reshuffling, relocating exiting segments in given data which is mostly heterogeneous so that the new segments have more homogeneous data items. This can be very easily understood by a simple example. Suppose some items have been segmented on the basis of color in the given data. Suppose the items are fruits, then the green segment may contain all green fruits like apple, grapes etc. thus a heterogeneous mixture of items.

Clustering segregates such items and brings all apples in one segment or cluster although it may contain apples of different colors red, green, yellow etc. thus a more homogeneous cluster than the previous cluster.

Clustering is a difficult task, why? In case of classification we already know the number of classes, either good or bad or yes or no or any number of classes. We also have the knowledge of classes properties so its easy to segment data into known classes. However, in case of clustering we don't know the number of clusters a priori. Once clusters are found in the data business intelligence, domain knowledge is needed to analyze the found clusters. Clustering can be the first step towards market segmentation i.e. we can use countermining to know the possible clusters in the data. Once clusters found and analyzed classification can be applied thus gaining more accuracy than any standalone technique. Thus clustering is at higher level than classification not only because of its complexity but also because it leads to classification.

5. Waterfal method can be used for dwh? 2 marks

Waterfall Model: The model is a linear sequence of activities like requirements definition, system design, detailed design, integration and testing, and finally operations and maintenance. The model is used when the system requirements and objectives are known and clearly specified. While one can use the traditional waterfall approach to developing a data warehouse, there are several drawbacks. First and foremost, the project is likely to occur over an extended period of time, during which the users may not have had an opportunity to review what will be delivered. Second, in today's demanding competitive environment there is a need to produce results in a much shorter timeframe.

Q.1: Explain analytic data application specification in Kimball 5 marks

Kimball also proposes a four-step approach where he starts to choose a business process, takes the grain of the process, and chooses dimensions and facts. He defines a business process as a major operational process in the organization that is supported by some kind of legacy system (or systems).

Q2: Bisinuss rules are validated using student database in LAB 5 marks

Q3: 2 real life examples of clustering. 5 marks

Examples of Clustering Applications

□ Marketing: Discovering distinct groups in customer databases, such as customers who make lot of long-distance calls and don't have a job. Who are they? Students.

Marketers use this knowledge to develop targeted marketing programs.

□ Insurance: Identifying groups of crop insurance policy holders with a high average

CS614 solved all current papers BY JAMIA

claim rate. Farmers crash crops, when it is “profitable”.

□ Land use: Identification of areas of similar land use in a GIS database.

□ Seismic studies: Identifying probable areas for oil/gas exploration based on seismic data.

Q5: What issues may accord during data acquisition and cleansing in agriculture case study?

Data Acquisition & Cleansing

Trained scouts from DPWQCP periodically visit randomly selected points and manually

note 35 attributes, with some given in Table 2. These hand-written sheets are

subsequently filed. For the last 10 years, the data collected was recorded by typing the hand-filled pest scouting sheets

Q7: Data parallelism explain with example 3 marks

Data Parallelism: Concept

□ 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. So the idea is that you are getting parallelization because we are getting semi-independent execution, data manipulation across the partitions. And as long as we keep the coordination required, we can get very good speedups. Well again this query coordinator, the thing that keeps the query distributed but still working and then collects its results.. Now that query coordinator can potentially be a bottleneck, because if it does too much work, that is serial execution. So the query coordination has to be very small amount of work. Otherwise the overhead gets higher and the serialization of the workload gets higher.

Q8: Under what condition an operation can be execute in parallel? 3 marks

Q10: which script language are used to perform complex transformation in DTS package? 2 marks

DTS designer is an application that uses graphical objects to help you build packages containing complex workflows. DTS Designer includes a set of model **DTS Package Templates**, each designed for a specific solution that you can copy and customize for your own installation. It is recommended for sophisticated data transformation solutions requiring multiple connections, complex workflows, and event-driven logic. DTS package templates are geared toward new users who are learning about DTS Designer or more experienced users who want assistance setting up specific DTS functionalities

Q11: Cleansing can be break down in Who many steps, write their names? 2 marks

Automatic Data Cleansing...

1) Statistical

2) Pattern Based

3) Clustering

4) Association Rules

Q12: What do u mean by “ keep competition hot in context of production selection and transformation while designing a data warehouse “. 2 marks

Keep the competition “hot”

□ Even if single winner, keep at least two in

□ Use virtual competition to bargain with the winner

Q13: Who merge column are selected in case of sort merge? 2 marks

Sort-Merge Join

Joined tables to be sorted as per WHERE clause of the join predicate. Query optimizer scans for (cluster) index, if exists performs join. In the absence of index, tables are sorted on the columns as per WHERE clause.If multiple

CS614 solved all current papers BY JAMIA

equalities in WHERE clause, some merge columns used. The Sort-Merge join requires that both tables to be joined are sorted on those columns that are identified by the equality in the WHERE clause of the join predicate. Subsequently the tables are merged based on the join columns. The query optimizer typically scans an index on the columns which are part of the join, if one exists on the proper set of columns, fine, else the tables are sorted on the columns to be joined, resulting in what is called a cluster index. However, in rare cases, there may be multiple equalities in the WHERE clause, in such a case, the merge columns are taken from only some of the given equality clauses. Because each table is sorted, the Sort-Merge Join operator gets a row from each table and compares it one at a time with the rows of the other table. For example, for equi-join operations, the rows are returned if they match/equal on the join predicate. If they are not equal or don't match, whichever row has the lower value is discarded, and next row is obtained from that table. This process is repeated until all the rows have been exhausted.

2) what are the two extremes for technical architecture design? 2

Pitfall: Extremes of Tech. Arch. Design

Common mistake: Attacking the problem from two extremes, neither is correct.

□ Focusing on data warehouse delivery, architecture feels like a distraction and impediment to progress and often end up rebuilding.

□ Investing years in architecture, forgetting primary purpose is to solve business problems, not to address any plausible (and not so plausible) technical challenge.

3) Different b/w non key or key data access?2

non-keyed access and keyed access

Non-keyed access uses no index. Each record of the database is accessed sequentially, beginning with the first record, then second, third and so on. This access is good when you wish to access a large portion of the database (greater than 85%).

Keyed access provides direct addressing of records. A unique number or character(s) is used to locate and access records. In this case, when specified records are required (say, record 120, 130, 200 and 500), indexing is much more efficient than reading all the records in between.

4) "Be a diplomat not a technologist"?2

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). Most of all, keep smiling. When all is said and done, you'll have a resource in place that will do magic, and your grief will be long past. Eventually, your smile will be effortless and real.

5) Dirty bit?2

To keep record of the rows that have been inserted into error tables due to certain errors we need an additional column in student table that will serve as dirty bit. Dirty bit of those records is set to true that are inserted in the error table

6) What are the problem face industry when the growth in usage of master table file increase?3

This growth in usage of master files, resulted in huge amounts of redundant data. The spreading of master files and massive redundancy of data presented some very serious problems, such as:

- Data coherency i.e. the need to synchronize data upon update.
- Program maintenance complexity.
- Program development complexity.
- Requirement of additional hardware to support many tapes.

In a nut-shell, the inherent problems of master files because of the limitations of the

medium used started to become a bottleneck. If we had continued to use only the magnetic tapes, we may not have had an Information revolution! Consequently, there would have never been large, fast MIS (Management Information Systems) systems, ATM systems, Airline Flight reservation systems, maybe not even Internet as we know it. As one of my teachers very rightly said, "every problem is an opportunity" therefore, the ability to store and manage data on diverse media (other than magnetic tapes) opened up the way for a very different and more powerful type of processing i.e. bringing the IT and the business user together as never before.

7) Indexing using I/O bottleneck? 3

Need For Indexing: I/O Bottleneck

Throwing more hardware at the problem doesn't really help, either. Expensive and multiprocessing servers can certainly accelerate the CPU-intensive parts of the process, but the bottom line of database access is disk access, so the process is I/O bound and I/O doesn't scale as fast as CPU power. You can get around this by putting the entire database into main memory, but the cost of RAM for a multi-gigabyte database is likely to be higher than the server itself! Therefore we index.

Although DBAs can overcome any given set of query problems by tuning, creating indexes, summary tables, and multiple data marts, or forbidding certain kinds of queries, they must know in advance what queries users want to make and would be useful, which requires domain-specific knowledge they often don't have. While 80% of database queries are repetitive and can be optimized, 80% of the ROI from database information comes from the 20% of queries that are not repetitive. The result is a loss of business or competitive advantage because of the inability to access the data in corporate databases in a timely fashion.

Implementation strategies 3

32.2 Implementation Strategies Top Down & Bottom Up approach: A Top Down approach is generally useful for projects where the technology is mature and well understood, as well as where the business problems that must be solved are clear and well understood. A Bottom Up approach is useful, on the other hand, in making technology assessments and is a good technique for organizations that are not leading edge technology implementers. This approach is used when the business objectives that are to be met by the data warehouse are unclear, or when the current or proposed business process will be affected by the data warehouse.

10) Import/export wizard tasks?3

Import/Export Wizard and DTS Designer both are the graphical methods of building a package. Both tools provide support to run the package also. Building a package means putting all the tasks that are supposed to be performed in a particular package together and setting their order of execution or defining workflow. Whereas when we actually run a package all the tasks are actually performed

11) Problem using SQL to fill up tables of ROLAP cube?3

Problem with simple approach

□ Number of required queries increases exponentially with the increase in number of dimensions.

□ It's wasteful to compute all queries.

□ In the example, the first query can do most of the work of the other two queries

□ If we could save that result and aggregate over Month_Id and Product_Id, we could compute the other queries more efficiently

Using typical SQL to fill-up the tables quickly runs into a problem, as the number of dimensions increases, the number of aggregates also increases, and the number of queries required to calculate those aggregates also increases. Actually it becomes extremely wasteful to compute all queries, wasteful, because if we are smart, we can use the results of the queries already computed to get the answers to new queries. How to do this? it is

CS614 solved all current papers BY JAMIA

not very difficult. For example for the column total queries, we could just add the aggregates over the results of the months. So the moral of the story is “Work smart not hard”.

Cube clause

- The CUBE clause is part of SQL:1999
- GROUP BY CUBE (v1, v2, ..., vn)
- Equivalent to a collection of GROUP BYs, one for each of the subsets of v1, v2, ..., vn

The other problem with using standard SQL is that one has to write too many statements and that could lead to mistakes. Therefore, back in 1999 a CUBE clause was made part of SQL, and that clause is equivalent to a collection of GROUP BY clauses.

Some students who did a BS final year project with me of an HOLAP implementation, used dynamic web page generation to dynamically generate SQL instead of hard-coding the queries to generate the aggregates. Meaning, they used SQL to generate aggregates to fill a MOLAP cube. The project was a success, all got jobs based on this work; the first prize in the 13th Annual National Software Competition along with a cash prize of Rs. 30,000 was a bonus.

12) How data mining is different from statistics? Which one is better? 5

How Data Mining is different...

Data Mining Vs. Statistics

- Formal statistical inference is assumption driven i.e. a hypothesis is formed and validated against the data.
- Data mining is discovery driven i.e. patterns and hypothesis are automatically extracted from data.
- Said another way, data mining is knowledge driven, while statistics is human driven.

Although both of the two are for data analysis and none is good or bad, some of the difference between statistics and Data mining are; Statistic is assumption driven. A hypothesis is formed using the historical data and is then validated against current known data. If true the hypothesis becomes a model else the process is repeated with different parameters. DM, on the other hand, is discovery driven i.e. based on the data hypothesis is automatically extracted from the data. The purpose is to find patterns which are implicit and hidden in the data sea otherwise. Thus data mining is knowledge driven while statistics is human driven.

Data Mining Vs. Statistics

- Both resemble in exploratory data analysis, but statistics focuses on data sets far smaller than used by data mining researchers.
- Statistics is useful for verifying relationships among few parameters when the relationships are linear.
- Data mining builds much complex, predictive, nonlinear models which are used for predicting behavior impacted by many factors.

Q1 why a pilot project strategy is highly recommended in DWH construction? 5

A pilot project strategy is highly recommended in data warehouse construction, as a full blown data warehouse construction requires significant capital investment, effort and resources. Therefore, the same must be attempted only after a thorough analysis, and a valid proof of concept. A small scale project in this regard serves many purposes such as (i) show users the value of DSS information, (ii) establish blue print processes for later full-blown project, (iii) identify problem areas and, (iv) reveal true data demographics. Hence doing a pilot project on a small scale seemed to be the best strategy.

As hardly any pilot projects are thrown away, therefore, Agri-DWH was treated with all due respect as a regular project. The other objectives of Agri-DWH were (i) demonstrate the utility of such an undertaking (ii) do a dry run of the entire cycle of an AE_DWH to get a feel of the “real-thing”.

Q2 define nested loop join list and describe its variants? 5

Nested loop join works like a nested loop used in a high level programming language, where each value of the index of the outer loop is taken as a limit

Nested-Loop Join: Variants

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

<http://www.vustudents.net>

Working of Query optimizer

There are many variants of the traditional nested-loop join. The simplest case is when an entire table is scanned; this is called a **naive nested-loop join**. If there is an index, and that index is exploited, then it is called an **index nested-loop join**. If the index is built as part of the query plan and subsequently dropped, it is called as a **temporary index nested-loop join**. All these variants are considered by the query optimizer before selecting the most appropriate join algorithm/technique.

Q4keeping view the uniform distribution in hash based partition .if the partitions are not uniformly distributed across the process? 3

Hash-Based Join: Partition Skew

Partition skew can become a problem.

Hashing works under the assumption of uniformity of data distribution, may not be always true.

Consequently hash-based join degrades into nested-loop join.

Solution: Make available other hash functions to be chosen by the optimizer; that better distribute the input.

Partition skew can become a problem in hash-join. In the first step of hash join, records are hashed into the main memory into their corresponding bucket. This being done based on the hash function used. However, an attribute being hashed may not be uniformly distributed within the relation, and some buckets may then contain more records than other buckets. When this difference becomes large, the corresponding bucket may no longer fit in the main memory. As a consequence, hash-based join degrades into performance of a nested-loop join. The only possible solution is to make available other hash functions to be chosen by the optimizer; that better distribute the input.

Q13how the application of parallelism differ for OLTP and DSS environment? 2

24.3 Parallelization OLTP Vs. DSS

There is a big difference.

DSS

Parallelization of a SINGLE query

OLTP

Parallelization of MULTIPLE queries

Or Batch updates in parallel

During business hours, most OLTP systems should probably not use parallel execution. During off-hours, however, parallel execution can effectively process high-volume batch operations. For example, a bank can use parallelized batch programs to perform the millions of updates required to apply interest to accounts. The most common example of using parallel execution is for DSS. Complex queries, such as those involving joins or searches of very large tables, are often best run in parallel.

Define Dense and Sparse index, adv and disadvantages (3)

Dense Index: Adv. & Dis. Adv.

For each record store the key and a pointer to the record in the sequential file. Why?

It uses less space, hence less time to search. Time (I/Os) logarithmic in number of blocks used by the index. Can also be used as secondary index, i.e. with another order of records.

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

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

CS614 solved all current papers BY JAMIA

Con: A dense index, if too big and doesn't fit into the memory, will be expensive when used to find a record given its key

Fixed strategy of standardizing column(2)

There are no fixed strategies to standardize the columns. Again it depends on the project designer what methodology he/she devises. We can devise a simple methodology that can later be used for other columns as well.

SQL server meta services advantages(3)

Meta Data Services. The advantage which we get when we store our package to SQL Server 2000 Meta Data Services is that we may maintain meta data information of the databases involved in the packages and we may keep version information of each package. Furthermore package can be stored in a structured file and Microsoft visual basic file.

Issues faced in data cleansing of Agri DWH(3)

Data cleansing and standardization is probably the largest part in an ETL exercise. For Agri-DWH major issues of data cleansing had arisen due to data processing and handling at four levels by different groups of people i.e. (i) Hand recordings by the scouts at the field level (ii) typing hand recordings into data sheets at the DPWQCP office (iii) photocopying of the scouting sheets by DPWQCP personnel and finally (iv) data entry or digitization by hired data entry operators

Why pilot strategy is recommended for construction of DWH(5)

A pilot project strategy is highly recommended in data warehouse construction, as a full blown data warehouse construction requires significant capital investment, effort and resources. Therefore, the same must be attempted only after a thorough analysis, and a valid proof of concept. A small scale project in this regard serves many purposes such as (i) show users the value of DSS information, (ii) establish blue print processes for later full-blown project, (iii) identify problem areas and, (iv) reveal true data demographics. Hence doing a pilot project on a small scale seemed to be the best strategy.

Purpose of DTS services(5)

DTS allows us to connect through any data source or destination that is supported by OLE DB. This wide range of connectivity that is provided by DTS allows us to extract data from wide range of legacy systems. Heterogeneous source systems store data with their local formats and conventions. While consolidating data from variety of sources we need to transform names, addresses, dates etc into a standard format.

2. How the three parallel tracks capture the user requirements in the Kimball's data

Parallel Tracks

3.1 Lifecycle Technology Track

3.1.1 Technical Architecture

3.1.2 Product Selection

3.2 Lifecycle Data Track

3.2.1 Dimensional Modeling

3.2.2 Physical Design

3.2.3 Data Staging design and development

3.3 Lifecycle Analytic Applications Track

3.3.1 Analytic application specification

3.3.2 Analytic application development

warehouse life cycle Road Map / three parallel techniques. 5 marks

the DWH life

cycle road-map was divided into three parts, you only have to cover these parts i.e. (i) project planning (ii) user requirement definition and (iii) three parallel tracks.

6. Write down the steps which are performed in clustering process. 3 marks

CS614 solved all current papers BY JAMIA

7. Give name of activities to be performed in planning and design phase as discussed in agri-DWH case study. 3 marks

8. problem in Partition Skew based Hash join. 3 marks

Suitable for the VLDB environment.

The choice which table first gets hashed plays a pivotal role in the overall performance of the join operation, this decided by the optimizer. The joined rows are identified by collisions i.e. collisions are "good" in case of hash join

9. total quality management. How total quality management technique is different better from old management techniques.

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.

☐ Quality assurance is NOT Quality improvement

TQM approach is advocating the involvement of all employees in the continuous improvement process, the ultimate goal being the customer satisfaction. The TQM philosophy of management is customer-oriented. All members of a total quality management (control) organization strive to systematically manage the improvement of the organization through the ongoing participation of all employees in problem solving efforts across functional and hierarchical boundaries. Quality assurance is a system of activities that assures conformance of product to preestablished requirements. Quality improvement is making all efforts to increase effectiveness and efficiency in meeting accepted customer expectations.

10. b-tree indexing limitations. 2 marks

B-tree Indexing: Limitations

☐ If a table is large and there are fewer unique values.

Capitalization is not programmatically enforced (meaning case-sensitivity does matter and "FLASHMAN" is different from "Flashman").

☐ Outcome varies with inter-character spaces.

☐ A noun spelled differently will result in different results.

☐ Insertion can be very expensive.

There are certain instances when a B-tree index is not appropriate and will not improve performance of queries. In many of these instances, such as a column in a data warehouse with relatively few distinct values, a bitmapped index can be created to dramatically improve performance. B-tree index is a poor choice for name and text searches because it is case-sensitive and requires a left-to-right match

11. Roll out and maintenance phase of agri DWH. 3 marks

the maintenance phase is usually boring. So, if there is another phase of the data warehouse planned, start on that as soon as possible

12. Data Transformation Services (DTS) provide a set of tools, Packages, tasks and connections that lets you extract, transform, and consolidate data from disparate sources into single or multiple destinations supported by DTS connectivity 5 marks

ix) What types of operations are performed by MS DTS. 3

x) Name of the pest scouting org and the year of its starting.

Answer DPWQCP 1984, 2 marks

10. WHAT are the method of developing DHW? 3marks

ANSWER: DATA DRIVEN, USER DRIVEN, GOAL driven

11. Why a pilot project strategy is highly recommended in DWH construction? 5marks

CS614 solved all current papers BY JAMIA

- 2) why analytic track is called the "funpart" while designing a data warehouse? 2
 - 3) why you need to analyze the web traffic at lowest level? 2
 - 4) what will be the effect if we program a package by using DTS object model? 2
 - 5) how grain is related with expressiveness? 2
 - 6) differentiate between knowledge discovery in data base, data mining and data warehouse? 3
 - 7) why building a data warehouse is challenging activity what are three broad categories of data warehouse development method? 3
 - 8) data profiling process? 3
 - 9) Give name of activities of to be performed in building and testing phase as discussed in agri- DWH case study? 3
 - 10) suppose you want to enhance performance of data warehouse which strategy throwing more hardware or aggregation will be used? 3
 - 11) what are the fundamental strengths and weakness of k mean clustering? 5
 - 12) write a query to extract total number of female students in BS telecom? 5
 - 13) describe the lessons learnt during agri- datawarehouse case study? 5
 - 1) single clustering double clustering
 - 2) one to one transformation, one to many transformation
 - 3) DTS benefits and usage
 - 4) K technique benefits and drawbacks.
 - 5) Business laws in students labs
 - 6) be a technologist is necessary in the DWH
 - 7) clicking stream in web DWH
 - in how many ways a user can access web data.
 - difference between MOLAP and DOLAP
 - Write SQL Query to find all Female student in BS telecom.
 - viii) Why DASD is better than tape storage w.r.t access time
 - viii) Which script language is used to perform complex transformation in dts package 2
 - ix) What types of operations are performed by MS DTS. 3
-

CS614 solved all current papers BY JAMIA

x) Name of the pest scouting org and the year of its starting. Answer DPWQCP 1984, 2 marks

12. Data Transformation Services (DTS) provide a set of tools, Packages, tasks and connections that lets you extract, transform, and consolidate data from disparate sources into single or multiple destinations supported by DTS connectivity 5 marks

Q11 which scripting language are used to perform complex transformation in DST package? 2

Q12a person wanted to visit and understand the data warehouse implementation strategies adopted in that organization has refused to allow . what may be the carrier of this refusal?

<http://www.vustudents.net>

