Paper 1:

1. **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. 2 m**

Answer: page 405

The major problem is the inconsistent data sources at different campuses. The attributes summarizes the data sources at two genders. The problem is non-standardized attributes across Genders, Different conventions for representing Gender across the campuses e.g. Lahore campus uses 0/1 while Islamabad uses 1/0 for representing male and female respectively. Similarly, there are different conventions for representing degree attribute across different campuses.

2. **What is the "intrinsic skew" problem in hash based join? Give an example? 2m**

Answer:

Intrinsic skew can become a problem for hash, as well as sort-merge join.
The skew is in data, NOT due to hash function.
Example: Many non-CS majors registering for CS-101 instead of CS students in summer.
Intrinsic skew occurs when attributes are not distributed uniformly; it is also called attribute value skew. For example a basic Computer Science (CS) course being offered in summer, and taken by many non-CS majors who want to know about computers. The course taken by few CS-majors who missed it or got an incomplete (i.e. I) grade during the regular semester due to one reason or another. Ironically, intrinsic skew effects the performance of both hash and sort-merge joins. Sort-merge join works best when the join attributes are the primary key of both tables. This property guarantees absence of duplicates, so that a record in the left-hand-side of the relation will join with at most one record in the right-hand-side of the relation, thus avoiding the intrinsic skew.

3. **List down the two ways to simplify the ER model. 2m**

Answer:

There are actually two ways of "simplifying" the ER model i.e. (i) De-normalization and (ii) Dimensional Modeling.

4. **List down any four Dynamic attributes recorded by the scouts in agriculture data warehouse case study. 2m**

Answer:

Both static and dynamic attributes

Static Attributes		Dynamic Attributes	
1	Farmer Name	1	Date of Visit
2	Farmer Address	2	Pest Population
3	Field Acreage	3	CLCV
4	Variety(ies) Sown	4	Predator Population
5	Sowing date	5	Pesticide Spray Dates
6	Sowing method	6	Pesticide(s) Used

The problem of Referential Integrity always occurs in traditional OLTP system as well as in DWH

5. **Identify the given statement as correct or incorrect and justify your answer in either case.**
   "The problem of Referential Integrity always occurs in traditional OLTP system as well as in DWH". 3m

Answer:
While doing total quality measurement, you measure RI every week (or month) and hopefully the number of orphan records will be going down, as you will be fine tuning the processes to get rid of the RI problems. Remember, RI problem is peculiar to a DWH, this will not happen in a traditional OLTP system.

6. **Identify the given statement as correct or incorrect and justify your answer in either case.**
   "One-way Clustering gives a local view whereas Two-way Clustering gives a global view". 3m

Answer: page 271
This above statement is incorrect.
Two-way Clustering/Biclustering gives a local view of your data set while one-way clustering gives a global view. It is possible that you first take global view of your data by performing one-way clustering and if any cluster of interest is found then you perform two-way clustering to get more details. Thus both the methods complement each other.

7. **We can identify the session in World Wide Web by using "Time-contiguous log entries" however there are some limitations of this technique. Briefly explain any two limitations. 3m**

Answer: A session can be consolidated by collecting time-contiguous log entries from the same host (Internet Protocol, or IP, address). In many cases, the individual hits comprising a session can be consolidated by collating time-contiguous log entries from the same host (Internet Protocol, or IP, address). If the log contains a number of entries with the same host ID in a short period of time (for example, one hour), one can reasonably assume that the entries are for the same session.
Limitations: • This method breaks down for visitors from large ISPs because different visitors may reuse dynamically assigned IP addresses over a brief time period.
• Different IP addresses may be used within the same session for the same visitor.

• This approach also presents problems when dealing with browsers that are behind some firewalls.

8. **Give an example of the company which is using Data Warehouse, for each of the following fields in Pakistan?**
    1. **Financial service/insurance**
    2. **Telecommunications**
    3. **Transportation ……. 3m**

Answer:

Financial service/insurance
– Union Bank
– State Bank of Pakistan
• **Telecommunications.**
– UFone
– PTCL
– PAKNET
• **Transportation.**
– PIA
iv) Government
-NADR

9. **Describe how business rules are validated using student databases in LAB lectures? 5m**

Answer:

Data profiling is a process which involves gathering of information about column through execution of certain queries with intention to identify erroneous records. In this process we identify Validation of business rules. Data profiled date of registration as well therefore we can validate the following business rules
• All new registrations are done in month of August before 28th
• 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. There can be a lot of business rules that are supposed to be validated at this stage. It totally depends upon the business.

10. **Identify the given statements as correct or incorrect and justify your answer in either case.**
    1. **“Median is an example of Distributive Aggregate”.**
    2. **“with the help of “current value field” method we can track as many changes as we wish”…..5m**

Answer:

First statement is incorrect because, Distributive Aggregate is computing aggregate directly from sub-aggregates. Examples: MIN, MAX, COUNT, SUM and Holistic Aggregate is Require unbounded amount of information about each subgroup. Examples: MEDIAN, COUNT DISTINCT.

Second statement is correct because, you use the third technique when you want to track a change in a dimension value, but it is legitimate to use the old value both before and after the

change. This situation occurs most often in the infamous sales force realignments, where although you have changed the names of your sales regions, you still have a need to state today's sales in terms of yesterday's region names, just to "see how they would have done" using the old organization.

11. **What is clustering? How it can be used in Telecom Industry for targeted sale promotion? 5m**

Answer: Clustering in the computer science world is the classification of data or object into different groups. It can also be referred to as partitioning of a data set into different subsets. There are two general algorithms used in data clustering. These categories are hierarchical and partitional. Clustering is useful for exploring data. If there are many cases and no obvious groupings, clustering algorithms can be used to find natural groupings. Clustering can also serve as a useful data-preprocessing step to identify homogeneous groups on which to build supervised models.

Answer: page 265

To know the customer is crucial for any organization/company so as to satisfy the customer which is a key of any company's success in terms of profit. The company can rum targeted sale promotion and marketing effort to target customers i.e. students.

12. **A pilot project strategy is highly recommended in data warehouse construction. In your opinion, what could be the reasons of this recommendation? 5m**

Answer:

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.

Paper 2:

Q. List down any two disadvantages of MOLAP. 2m

Answer: 1. Long load time (pre-calculating the cube may take days!). Second is Very sparse cube (wastage of space) for high cardinality (sometimes in small hundreds). E.g. number of heaters sold in Jacobabad or Sibi.

- Within some MOLAP Solutions the processing step (data load) can be quite lengthy, especially on large data volumes. This is usually remedied by doing only incremental processing, i.e., processing only the data which have changed (usually new data) instead of reprocessing the entire data set.
- Some MOLAP methodologies introduce data redundancy.

Q. what is unsupervised learning in data mining? 2m

Answer: page 27

Unsupervised learning where you don't know the number of clusters and obviously no idea about their attributes too. In other words you are not guiding in any way the DM process for performing the DM, no guidance and no input. Unsupervised learning is closer to the exploratory spirit of

Data Mining as stressed in the definitions given above. In unsupervised learning situations all variables are treated in the same way, there is no distinction between explanatory and dependent variables. However, in contrast to the name undirected data mining there is still some target to achieve. This target might be as general as data reduction or more specific like clustering. For unsupervised learning typically either the target variable is unknown or has only been recorded for too small a number of cases.

Q. list down any four ways to identify the session in World Wide Web 2m
Answer: page 364
The basic protocol for the World Wide Web, HTTP, is stateless-that is, it lacks the concept of a session. There are no intrinsic login or logout actions built into the HTTP, so session identity must be established in some other way. There are several ways to do this

1. Using Time-contiguous Log Entries
2. Using Transient Cookies
3. Using HTTP's secure sockets layer (SSL)
4. Using session ID Ping-pong
5. Using Persistent Cookies

Most web-centric data warehouse applications will require every visitor session (visit) to have its own unique identity tag similar to a grocery store point-of-sale ticket ID or session ID. The rows of every individual visitor action in a session, whether derived from the click stream or from an application interaction, must contain this tag. However, it must be kept in mind that the operational application generates this session ID, not the Web server.

Q. Which authority is recording the pest scouting data in Punjab? 2m
Answer:
The data, pest scouting data, has been taken from department of pest warning Punjab for the year 2000 and 2001. Some of the attributes like farm area, cotton variety cultivated and pesticide used are selected. Using these and some other attributes a similarity matrix has been formed based on Pearson's Correlation. The pest scouting data is being constantly recorded by the Directorate of Pest Warning and Quality Control of Pesticides (DPWQCP), Punjab since 1984. However, despite pest scouting, yield losses have been occurring. The most recent being the Boll Worm attack on the cotton crop during 2003-04, resulting in a loss of nearly 0.5 million bales.

Q. identifies the given statement as correct or incorrect and justifies your answer in either case. 3m
"if defects are found in the process of Attribute Domain Validation then it is better to fix the errors in DWH and leave the data source as it as".
Answer: page 190
Once the defects are found, go and track them down back to source cause(s).
Point to be noted is that, if at all possible, fix the problem in the source system. People have the tendency of applying fixes in the DWH. This is a wrong i.e. if you are fixing the problems in the DW; you are not fixing the root cause. A crude analogy would clarify the point. If you keep cleaning the lake, and keep on flushing the toilet in the lake, you are not solving the problem. The problem is not being fixed at the source system, therefore, it will persist.

Q. identifies the given statement as correct or incorrect and justifies your answer in either case. 3m
"In prediction, Test data set is used to form the model and the associated rules". Page 261
Answer: This above statement is incorrect. Because, The existing data set is divided into two subsets, one is called the training set and the other is called test set. The training set is used to form model and the associated rules. Once model built and rules defined, the test set is used for grouping. It must be noted the test set groupings are already known but they are put in the model to test its accuracy.

Q. Briefly explain the tree major approaches for accessing the information stored on the web. 3m
Answer: page 351
Currently, users can choose from three major approaches when accessing information stored on the Web:
(i) Keyword-based search or topic-directory browsing with search engines such as Google or Yahoo, which use keyword indices or manually built directories to find documents with specified keywords or topics;
(ii) Querying deep Web sources—where information, such as amazon.com's book data and realtor.com's real-estate data, hides behind searchable database query forms—that, unlike the surface Web, cannot be accessed through static URL links; and
(iii)Random surfing that follows Web linkage pointers. The success of these techniques, especially with the more recent page ranking in Google and other search engines shows the Web's great promise to become the ultimate information system.

Q. In most of the retail banks 50% of the customers are unprofitable. By considering this situation, which particular application of DWH can help the banks to improve the business and how? 5m
Answer: page 38
If you go to an average bank, to Pakistan or anywhere in the world, the bank will know if they are profitable or not. But they don't know which customers are profitable or not, and at most banks this is true, especially true for retail banks. These banks do business with the consumers, such that more than 50 % of the customers are unprofitable. In other word the bank is losing money on 50% of their customers. But they don't know which 50%? That's the problem. So the idea behind profitability analysis is to do the analysis to figure out which customers are profitable and which customers are not profitable. And then based on that they can restructure their product offering and there pricing strategies to do more profitable business. I used a banking example but same is true for Telecommunications Company or any other consumer oriented business that requires access to detail data. It is not sufficient to just know the account balance, but also transactional behavior and so on to do profitability analysis. And once I know profitability retrospectively, I can also build predictive models to understand what it's going to be prospectively. So I can build what's called lifetime value models to using the historical data to predict what the future profitability will be for the next 3 or 5 years.

Q. we have seen many issues during data acquisition & cleansing in agriculture data warehouse case study. In your point of view why those issues have arisen? Briefly explain….5m

Answer: page 340

Major issues of data cleansing had arisen due to data processing and handling at four levels by different groups of people

1. Hand recordings by the scouts at the field level.
2. Typing hand recordings into data sheets at the DPWQCP office.
3. Photocopying of the typed sheets by DPWQCP personnel.
4. Data entry or digitization by hired data entry operators.

After achieving acceptable level of data quality, the data was loaded into Tera_data data warehouse; subsequently each column was probed using SQL for erroneous entries. Some of the errors found were correct data in wrong columns, nonstandard or invalid variety names etc.

Another hand: 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.

Q. Briefly explain any two types of precedence constraints that we can use in DTS….5m
Answer: page 395

Precedence constraints sequentially link tasks in a package. In DTS, you can use three types of precedence constraints, which can be accessed either through DTS Designer or programmatically:

Unconditional: If you want Task 2 to wait until Task 1 completes, regardless of the outcome, link Task 1 to Task 2 with an unconditional precedence constraint.

On Success: If you want Task 2 to wait until Task 1 has successfully completed, link Task 1 to Task 2 with an On Success precedence constraint.

On Failure: If you want Task 2 to begin execution only if Task 1 fails to execute successfully, link Task 1 to Task 2 with an On Failure precedence constraint. If you want to run an alternative branch of the workflow when an error is encountered, use this constraint.

1. **Explain sequence clustering algorithm.**

Answer

Sequence clustering algorithm collects similar or related paths, sequences of data containing events. E.g. Sequence clustering algorithm may help finding the path to store a product of "similar" nature in a retail ware house.

2. **What is data modeling and data mining?**

Designing a model for data or database is called data modeling. Data is reposited in fact table and dimension table. Fact table consists of data about transaction and dimensional table consists of master data. Data model is used to design abstract model of database.

The process of obtaining the hidden trends is called as data mining. Data mining is used to transform the hidden into information. Data mining is also used in a wide range of practicing profiles such as marketing, surveillance, fraud detection.

3. **What is data modeling and data mining? What is this used for?**

Data modeling aims to identify all entities that have data. It then defines a relationship between these entities. Data models can be conceptual, logical or Physical data models. Conceptual models

are typically used to explore high level business concepts in case of stakeholders. Logical models are used to explore domain concepts. While Physical models are used to explore database design, Data mining is used to examine or explore the data using queries. These queries can be fired on the data warehouse. Data mining helps in reporting, planning strategies, finding meaningful patterns etc. it can be used to convert a large amount of data into a sensible form.

4. What are Critical Success Factors?

The key areas of activity in which favorable results are necessary for a company to reach its goal, **there are four basic types of CSFs which are:** Industry CSFs, Strategy CSFs, Environmental CSFs, and Temporal CSFs.

A few CSFs are: Money, Your future, Customer satisfaction, Quality, Product or service development, Intellectual capital, Strategic relationships, Employee attraction and retention, Sustainability.

The advantages of identifying CSFs are: they are simple to understand; they help focus attention on major concerns; they are easy to communicate to coworkers; they are easy to monitor; and they can be used in concert with strategic planning methodologies.

5. What is Chained Data Replication?

In Chain Data Replication, the non-official data set distributed among many disks provides for load balancing among the servers within the data warehouse. Blocks of data are spread across clusters and each cluster can contain a complete set of replicated data. Every data block in every cluster is a unique permutation of the data in other clusters. When a disk fails then all the calls made to the data in that disk are redirected to the other disks when the data has been replicated. At times replicas and disks are added online without having to move around the data in the existing copy or affect the arm movement of the disk. In load balancing, Chain Data Replication has multiple servers within the data warehouse share data request processing since data already have replicas in each server disk.

6. What is the purpose of cluster analysis in Data Warehousing?

Cluster analysis is used to define the object without giving the class label. It analyzes all the data that is present in the data warehouse and compare the cluster with the cluster that is already running. It performs the task of assigning some set of objects into the groups are also known as clusters. It is used to perform the data mining job using the technique like statistical data analysis. It includes all the information and knowledge around many fields like machine learning, pattern recognition, image analysis and bio-informatics. Cluster analysis performs the iterative process of knowledge discovery and includes trials and failures. It is used with the pre-processing and other parameters as a result to achieve the properties that are desired to be used.

7. What are the different models used in cluster analysis?

There are many algorithms that can be used to analyze the database to check the maintenance of all the data sets that are already present. The different types of cluster models include as follows:

• **Connectivity models:** these are the models that connect one cluster to another cluster. This includes the example of hierarchical clustering that is based on the distance connectivity of one model to another model.

• **Centroid models:** these are the models that are used to find the clusters using the single mean vector. It includes the example of k-means algorithm.

• **Distribution models:** it includes the specification of the models that are statistically distributed for example multivariate normal distribution model.

• **Density models:** deals with the clusters that are densely connected with one another in the regions having the data space.
• **Group models:** specifies the model that doesn’t provide the refined model for the output and just gives the grouping information.

8. What is the difference between agglomerative and divisive Hierarchical Clustering?

• Agglomerative Hierarchical clustering method allows the clusters to be read from bottom to top and it follows this approach so that the program always reads from the sub-component first then moves to the parent. Whereas, divisive uses top-bottom approach in which the parent is visited first then the child.
• Agglomerative hierarchical method consists of objects in which each object creates its own clusters and these clusters are grouped together to create a large cluster. It defines a process of merging that carries on till all the single clusters are merged together into a complete big cluster that will consists of all the objects of child clusters. Whereas, in divisive the parent cluster is divided into smaller cluster and it keeps on dividing till each cluster has a single object to represent.

9. Why is chameleon method used in data warehouse?

Chameleon is a hierarchical clustering algorithm that overcomes the limitations of the existing models and the methods present in the data warehousing. This method operates on the sparse graph having nodes that represent the data items and edges represent the weights of the data items. The representation of it allows large data set to be created and operated on successfully. The method finds the clusters that are used in the data set using the two phase algorithm. The first phase consists of the graph partitioning that allows the clustering of the data items into large number of sub-clusters. Second phases use an agglomerative hierarchical clustering algorithm to search for the clusters that are genuine and can be combined together with the sub-clusters that are produced.

10. What are the key features of chameleon that separates it from other algorithms?

The key features that are in the chameleon are:
•The chameleon method determines the pair of similar sub-clusters that can be connected with other clusters. It also finds the closeness of the clusters from one another.
• The chameleon with the above property overcomes the limitation that is present in agglomerative hierarchical model.
• It uses different methods to take the internal characteristics of the clusters and matches with those which are already present.
• It doesn't depend on static model that is supplied by the user and uses automated functions to perform the merging of the clusters that are already associated in the cluster.

1. **Write 2 Quality improvement attributes?**

   Ans.
   The four categories of Data Quality Improvement
   - Process
   - System
   - Policy & Procedure

- Data Design

2. **Data Cleansing and acquisition in Agri Data warehouse??**

   Ans.
   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 issues
   1. The pest scouting sheets are larger than A4 size (8.5” x 11”), hence the right end was cropped when scanned on a flat-bed A4 size scanner.

   2. The right part of the scouting sheet is also the most troublesome, because of pesticide names for a single record typed on multiple lines i.e. for multiple farmers.
   3. As a first step, OCR (Optical Character Reader) based image to text transformation of the pest scouting sheets was attempted. But it did not work even for relatively clean sheets with very high scanning resolutions.
   4. Subsequently DEO’s (Data Entry Operators) were employed to digitize the scouting sheets by typing.

3. Why RAD is better than other development methodologies

Ans.
Assemble a small, very bright team of database programmers, hardware technicians, designers, quality assurance technicians, documentation and decision support specialists, and a single manager.
2. Define and involve a small "focus group" consisting of users (both novice and experienced) and managers (both line and upper). These are the people who will provide the feedback necessary to drive the prototyping cycle. Listen to them carefully.
3. Generate a user's manual and user interface first. These will prove to be amazing in terms of user feedback and requirements specification

5. 5 statements were given or poocha tha k classification, estimation or prediction konsi hain in mai?
Ans.
Identify the following examples corresponds to which Data mining Technique:
Note: Review examples in Chapter no : 30
Assigning customers to predefined customer segments (good vs. bad)
Answer:
Classification
Classifying instructor rating as excellent, very good, good, fair, or poor
Answer:
Classification

Building a model and assigning a value from 0 to 1 to each member of the set
Answer:
Estimation
Predicting how much customers will spend during next 6 months.

Answer:
Prediction
6. One statement was given or batana tha k correct or incorrect??
7. Query likhni thi dirty bit say related thi??
I think k ye wali ho gi
Two tables were given one was employee table and other exception table with attribute IsAgeValid. Sql query was required to find the outliers not having age between 26 and 75 and set the dirtyBit in exception table to 0?5 marks
Answer:
Select * From (Select * From Employee Where DirtyBit= 0)IsAgeValid Where (Age <26 And >75))

8. If organization has not a subject expert in data modeling then what will the implications of it on organization
Ans.
Without this person, it becomes difficult to get a definitive answer on many of the questions, and the entire project gets dragged out, as the end users may not always be available. It is essential to have a subject-matter expert as part of the data modeling team. This person can be an outside consultant or can be someone in-house with extensive industry experience.
10. Data profiling purposes?
Ans.
Data profiling is a process which involves gathering of information about column through execution of certain queries with intention to identify erroneous records. In this process we identify the following:

- Total number of values in a column
- Number of distinct values in a column
- Domain of a column
- Values out of domain of a column
- Validation of business rule

11. Define Justification??
Ans. Page 292
Justification requires an estimation of the benefits and costs associated with a data warehouse. The anticipated benefits grossly outweigh the costs. IT usually is responsible for deriving the expenses. You need to determine approximate costs for the requisite hardware and software. Data warehouses tend to expand rapidly, so be sure the estimates allow some room for short-term growth.
Q.1 Forward Proxy (2
Ans.
The type of proxy we are referring to in this discussion is called a forward proxy. It is outside of our control because it belongs to a networking company or an ISP.

Drawbacks of waterfall model for DHW (3.
In which scenario we can use waterfall (2
Ans. there is 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.
1 .Write two unsupervised learning page no. 270
ans.
1.one way clustering
2.two way clustering

2. Output of run length encoding (2 marks)
Run length used in bitmap indexing
Output 1 may be
15#02# 18# (mean 1 come 5 time and 0 come 2 times and 1 come 1 8 times (111110011111111))
Output 2 may be
11#01#11#
Output may 3 be
112#012#

Q.2 Bitmap index: run length encoding ka ek question tha input di hoi output find out kerni thi Page no.234
B-tree vs. hash indexes men se ye query di hoi thi
SELECT*FROM R WHERE A= 5 page no.228
Btana tha k is men dense index sparse index B-tree index and bitmap index men se konsi technique use ho gi aur explain kerna tha ise
This given statement I think answer is
Indexing (using B-trees) good for range searches, e.g.:
SELECT * FROM R WHERE A > 5
Hashing good for match based searches, e.g.:
SELECT * FROM R WHERE A = 5
A B-tree index is more functional and flexible than a hash scan. It allows partial values to be specified as the retrieval criteria, and it brings back the rows in sorted order. It is most useful for fixed-structure or hierarchical columns such as dates or financial account numbers. Hash based indexing is fast for specific value searches as it takes O(1)for uniform hashing as opposed to O(log n) time for a B-Tree based index

Identify kerna tha k ye statement correct he ya incorrect aur reason btana tha
no idea

Bayesian modeling is an example of unsupervised learning" page no 270
Page no. 278 pe
IF(ITEMS/TIME)>6
Then

Gender= female
Else
No idea but this statement is in page no 278
Gender= male esa ek question tha aur baki yad nahi but past men se the paper start wale lectures men se tha

Q.1 SSL and SMP stand for? 2 mark
Ans. secure sockets layer (SSL
Symmetric multi-processors (SMP)
Q.2 Type of Parallelism?

- Data Parallelism
- Spatial Parallelism

Q.7 Limitations of web data ware housing
No idea
Q.8 Name of three DWH development methodologies?
Answer:
• Waterfall model
• Spiral model
• RAD Model
• Structured Methodology
• Data Driven
• Goal Driven
• User Driven
Q.9 what issues may occur during data acquisition and cleansing in agriculture study? 5 marks
Ans: The pest scouting sheets are larger than A4 size (8.5” x 11”), hence the right end was cropped when scanned on a flat-bed A4 size scanner. The right part of the scouting sheet is also the most troublesome, because of pesticide names for a single record typed on multiple lines i.e. for multiple farmers.
As a first step, OCR (Optical Character Reader) based image to text transformation of the pest scouting sheets was attempted. But it did not work even for relatively clean sheets with very high scanning resolutions. Subsequently DEO’s (Data Entry Operators) were employed to digitize the scouting sheets by typing.
The pest scouting sheets are larger than A4 size (8.5” x 11”), hence the right end was cropped when scanned on a flat-bed A4 size scanner. The right part of the scouting sheet is also the most troublesome, because of pesticide names for a single record typed on multiple lines i.e. for multiple farmers.

Q.10 do you think it will create the problem of non-standardized attributes if
Uses 0/1 second source uses 0/1 to store male/female respectively?
Ans. Page no 405
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.
Each of the campuses uses different attribute name for the identification or primary keys e.g. Lahore uses SID while Peshawar uses Reg# and so on.

Different conventions for representing Gender across the campuses e.g. Lahore campus uses 0/1 while Islamabad uses 1/0 for representing male and female respectively.
Similarly, there are different conventions for representing degree attribute across different campuses.

Q.11 ik table diyahoatha bitmap index table
No idea how draw bitmap index
Q.12 Why pilot strategy is recommended for construction of DWH? Explains with reasons
Ans.
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.

CS614 VU Today's Final Term Paper for fall 2013/2014

Q1: Identify the statements correct or incorrect justify in either case: (5)
1. “Hash based indexing keeps the index entries in B-tree structure”.
2. “Just like primary key primary index has to be unique always”.
Answer:
First statement is incorrect as the correct one is: page 227
Index entries kept in hash organized tables rather than B-tree structures.

Second statement is also incorrect the correct one is: page 229
Primary Key (PK) & Primary Index (PI):
PK is ALWAYS unique.
PI can be unique, but does not have to be.

Q3: How gender guide is used for large no of records if gender is missing? (5) Page 457
Answer: Gender guide contains only two columns name and gender. Populate Gender guide table by a query for selecting all distinct first names from student table. Then manually placing their gender, this table can serve us as guide by telling what can be the gender against this particular name. For example if we have hundred students in our database with first name equal to ‘Muhammad’. Then in our Gender guide table we will have just one entry ‘Muhammad’ and we will manually set the gender as ‘Male’ against ‘Muhammad’.
Now to fill missing genders in exception table we will just do an inner join on Error table and Gender guide table.

Q5: Data profiling is a process which involves gathering of information about column. What is Data profiling purpose? (3) Page 439
Answer: To identify the degree of transformation required we will perform data profiling
Data profiling is a process which involves gathering of information about column through execution of certain queries with intention to identify erroneous records. In this process we

identify the following:
Total number of values in a column
Number of distinct values in a column
Domain of a column
Values out of domain of a column
Validation of business rules

Q6: Write down three cotton pest scouting Dynamic attributes? (3) page 342
Answer:

Static Attributes		Dynamic Attributes	
1	Farmer Name	1	Date of Visit
2	Farmer Address	2	Pest Population
3	Field Acreage	3	CLCV
4	Variety(ies) Sown	4	Predator Population
5	Sowing date	5	Pesticide Spray Dates
6	Sowing method	6	Pesticide(s) Used

Table-38.1: Cotton pest scouting attributes recorded by DPWQCP surveyors

Q7: What is the ranking in DSS? (3)
Answer: Page no : 143 ch:17
Ranking is all about selecting the “right” source system. Rank establishment has to be based on which source system is known to have the cleanest data for a particular attribute. Obviously you take the data element from the source system with the highest rank where the element exists. However, you have to be clever about how you use the rank.

Q9: What are problem you will face if low priority is given to cube construction? (2)
Answer: page 313
Low priority for OLAP Cube Construction
Make sure your OLAP cube-building or pre-calculation process is optimized and given the right priority. It is common for the data warehouse to be on the bottom of the nightly batch loads, and after the loading the DWH, usually there isn't much time left for the OLAP cube to be refreshed. As a result, it is worthwhile to experiment with the OLAP cube generation paths to ensure optimal performance.

Q10: Is there any fixed strategy to standardize the column? (2) page 480
Answer: There are no fixed strategies to standardize the columns.

Q12: Which DML operation is used in OLAP? (2) page 76
Answer: In OLAP applications the typical user is an analyst who is interested in selecting data needed for decision support. He/She is primarily not interested in detailed data, but usually in aggregated data over large sets of data as it gives the big picture. A typical OLAP query is to find the average amount of money drawn from ATM by those customers who are male, and of age

between 15 and 25 years from (say) Jinnah Super Market Islamabad after 8 pm. For this kind of query there are no DML operations and the DBMS contents do not change.

Why an organization refuses to visit a person to understand the strategy of Data Warehouse system?
Answer:

Define Forward proxy?
Answer: Ch#40 Page no : 369
The type of proxy we are referring to in this discussion is called a forward proxy. It is outside of our control because it belongs to a networking company or an ISP. When people talk about a proxy server (often simply known as a "proxy"), more often than not they are referring to a forward proxy. Let me explain what this particular server does. When one of these clients makes a connection attempt to that file transfer server on the Internet, its requests have to pass through the forward proxy first. A forward proxy is typically used in tandem with a firewall to enhance an internal network's security by controlling traffic originating from clients in the internal network that are directed at hosts on the Internet.

Define Reverse proxy?
Answer: Ch#40 Page no : 369
Another type of proxy server, called a reverse proxy, can be placed in front of our enterprise's Web servers to help them offload requests for frequently accessed content. This kind of proxy is entirely within our control and usually presents no impediment to Web warehouse data collection. It should be able to supply the same kind of log information as that produced by a Web server. A reverse proxy does the exact opposite of what a forward proxy does. While a forward proxy proxy in behalf of clients (or requesting hosts), a reverse proxy proxies in behalf of servers. To the client in our example, it is the reverse proxy that is providing file transfer services. The client is oblivious to the file transfer servers behind the proxy, which are actually providing those services. In effect, whereas a forward proxy hides the identities of clients, a reverse proxy hides the identities of servers.

What is mean by click stream? How it can be useful in a web DWH environment
Answer: Ch#40 Page no : 363
Click stream is every page event recorded by each of the company's Web servers**.** The click stream 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 click stream data. These formats have optional data components that, if used, can be very helpful in identifying visitors, sessions, and the true meaning of behavior. 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 click stream is every page event recorded by the web server. The click stream contains a number of new dimensions such as page, session, and referrer-that were previously unknown in conventional DWH environment.

Define Classification Process? How to measure the accuracy of Classifier?
Answer: Ch#31 Page no : 276

The classification process will know the box/class of the document that it belongs to. Thus in this way classification is performed. The classification can be used for customer segmentation, to detect fraudulent transactions and issues related to money laundering and the list goes on. It works by creating a model based on known facts and historical data by dividing into training and test set. Accuracy or confidence level = matches/ total number of matches. In simple words, accuracy is obtained by dividing number of correct assignments by total number of assignments by the classification model. It should be noted that you know the class for each record in test set and this fact is used to measure the accuracy or confidence level of the classification model.

What operations are provided by MS DTS?

Answer: lab Page no : 373

• A set of tools for
– Providing connectivity to different databases
– Building query graphically
– Extracting data from disparate databases
– Transforming data
– Copying database objects
– Providing support of different scripting languages (by default VB-Script and J-Script)

Differentiate between Range partitioning and Expression Partitioning?

Answer: ch#9 Page no : 66

The most common use of range partitioning is on date. This is especially true in data warehouse deployments where large amounts of historical data are often retained. Hot spots typically surface when using date range partitioning because the most recent data tends to be accessed most frequently.

Expression partitioning is usually deployed when expressions can be used to group data together in such a way that access can be targeted to a small set of partitions for a significant portion of the DW workload.

Kimball's life cycle model?

Answer: ch#33 Page no : 289

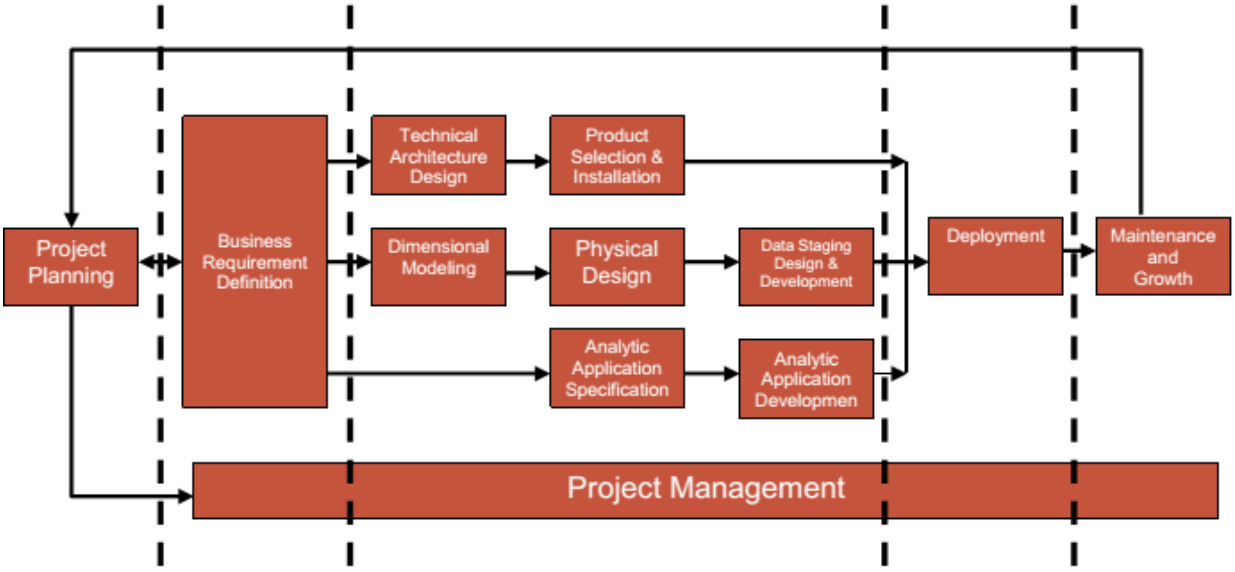


Figure -33.1: Business Dimensional Lifecycle (Kimball's Approach)

1. Project Planning
2. Business Requirements Definition
3. Parallel Tracks

3.1 Lifecycle Technology Track
3.1.1 Technical Architecture
3.1.2Product Selection
3.2 Lifecycle Data Track
3.2.1 Dimensional Modeling
3.2.2Physical Design
3.2.3Data Staging design and development
3.3 Lifecycle Analytic Applications Track
3.3.1Analytic application specification
3.3.2Analytic application development

4. Deployment
5. Maintenance

How time contiguous log entries and HTTP secure socket layer are used for user session identification? What are the limitations of these techniques?

Answer: CH#40 page 365

A session can be consolidated by collecting time-contiguous log entries from the same host (Internet Protocol, or IP, address). In many cases, the individual hits comprising a session can be consolidated by collating time-contiguous log entries from the same host (Internet Protocol, or IP, address). If the log contains a number of entries with the same host ID in a short period of time (for example, one hour), one can reasonably assume that the entries are for the same session.

Limitations

- This method breaks down for visitors from large ISPs because different visitors may reuse dynamically assigned IP addresses over a brief time period.
- Different IP addresses may be used within the same session for the same visitor.

- This approach also presents problems when dealing with browsers that are behind some firewalls.

Using HTTP's secure sockets layer (SSL):
This offers an opportunity to track a visitor session because it may include a login action by the visitor and the exchange of encryption keys.

Limitations

- ƒ To track the session, the entire information exchange needs to be in high
- overhead SSL
- Ƒ each host server must have its own unique security certificate.
- Ƒ Visitors are put-off by pop-up certificate boxes.

Explain analytic data application specification in Kimball?
Answer: CH#34 page 306

- ƒ Starter set of 10-15 applications.
- ƒ Prioritize and narrow to critical capabilities.
- ƒ Single template use to get 15 applications.
- ƒ Set standards: Menu, O/P, look feel.
- ƒ From standard: Template, layout, I/P variables, calculations.
- ƒ 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, the number of specific analyses that can be created from a single template merely by changing variables will surprise you.

Microsoft® SQL Server™ 2000 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.

Answer: Data Transformation Services
• DTS Overview
• SQL Server Enterprise Manager
• DTS Basics
–DTS Packages
–DTS Tasks
–DTS Transformations

–DTS Connections
–Package Workflow
The purpose of this lecture is to get an understanding of DTS basics, which is necessary to learn the use of DTS tools. These DTS basics describe the capabilities of DTS and summarize the business problems it addresses.

Ralph Kimball approach is business dimensional lifecycle.
Answer:
Kimball is considered as an authority in the DWH field, and his goal driven approach is a result of decades of practical experience. This presentation is a an overview of a data warehouse project lifecycle, based on this approach, from inception through ongoing maintenance, identifying best practices at each step, as well as potential vulnerabilities. It is believed that everyone on the project team, including the business analyst, architect, database designer, data stager, and analytic application developer, needs a high-level understanding of the complete lifecycle of a data warehouse. The Kimball's iterative data warehouse development approach drew on decades of experience to develop the business dimensional lifecycle. The name was because it reinforced several of key tenets for successful data warehousing.

Which is one of the five largest production countries in world?
2Marks
Answer: CH#37 page 330
Pakistan is one of the five major cotton-growing countries in the world. Almost 70% of world cotton is produced in China (Mainland), India, Pakistan, USA and Uzbekistan**.**

What do you say about Waterfall Model for DWH development?
3Marks
Answer: CH#32 page 284
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.
Define Data Profiling.
Answer: lab:4page 439
To identify the degree of transformation required we will perform data profiling
Data profiling is a process which involves gathering of information about column through execution of certain queries with intention to identify erroneous records. In this process we identify the following:
Total number of values in a column
Number of distinct values in a column
Domain of a column
Values out of domain of a column

Validation of business rules

Q. Issues of Click stream. Under which category it is lie?
Answer: ch:40 page 363
Click stream data has many issues.
1. Identifying the Visitor Origin
2. Identifying the Session
3. Identifying the Visitor
4. Proxy Servers
5. Browser Caches

Unlike data from OLTP system, where there were nice user identifications such as unique IDs that were the primary keys, in the context of a web log, this is one of the most issues i.e. identification of the visitor, so is where the visitor actually came from. In OLTP system there was a clean session beginning and session ending, but web is session less. It is very difficult and challenging to identify the session of a visitor, and the list goes on. Click stream data contains many ambiguities. Identifying visitor origins, visitor sessions, and visitor identities is something of an interpretive art. Browser caches and proxy servers make these identifications even more challenging.

Prepare Shaku Atretopic: CH: 37 Page no : 335

Phase_1: Planning & Design	Phase_2: Building & Testing	Phase_3: Roll-Out & Maintenance
1. Determine Users' Needs	6. Data Acquisition & Cleansing	12. Deployment & System Management
2. Determine DBMS Server Platform	7. Data Transform, Transport & Populate	
3. Determine Hardware Platform	8. Determine Middleware Connectivity	
4. Information & Data Modeling	9. Prototyping, Querying & Reporting	
5. Construct Metadata Repository	10. Data Mining	
	11. On Line Analytical Processing	

Table-37.1: The 12-step implementation approach of a data warehouse of Shaku Atre

Kimball Process. Four step approach. (Business process-->Grains-->Facts-->Dimensions sees assignment to clear this concept). (Read "Business Development Lifecycle" see page#290

Drawbacks of traditional web searches
Answer: ch: 39 page 351
1. Limited to keyword based matching.
2. Cannot distinguish between the contexts in which a link is used.
3. Coupling of files has to be done manually.

Data warehousing concepts are being applied over the Web today. Traditionally, simple search engines have been used to retrieve information from the Web. These serve the basic purpose of data recovery, but have several drawbacks.

There are some sign of trouble which serve as key indicator that the data ware house project is under threat list only five.

Answer: Page no : 311 Chapter: 35

1. Project proceeded for two months and nobody has touched the data.
2. End users are not involved hands-on from day one throughout the program.
3. IT team members doing data design (modelers and DBAs) have never used the access tools.
4. Summary tables defined before raw atomic data is acquired and base tables have been built.
5. Data design finished before participants have experimented with tools and live data.

What is Web Data Warehouse?

Answer: Page no: 350 Chapter: 39

Web Warehousing can be used to mine the huge web content for searching information of interest. It's like searching the golden needle from the haystack. Second reason of Web warehousing is to analyze the huge web traffic. This can be of interest to Web Site owners, for e-commerce, for e-advertisement and so on. Last but not least reason of Web warehousing is to archive the huge web content because of its dynamic nature.

We use static algorithm in data mining yes or no.

Answer: Page no:251 Chapter: 29

NO, Data mining consists of algorithms for extracting useful patterns from huge data. Their goal is to make prediction or/and give description. Prediction involves using some variables to predict unknown values (e.g. future values) of other variables while description focuses on finding interpretable patterns describing the data

Which department in Punjab monitors agriculture?

Answer: Page no : 333 Chapter: 37

Directorate of Pest Warning and Quality Control of Pesticides (DPWQCP), Punjab since 1984

Nested loop efficient way of accessing the inner table:

Answer: Ch#28 Page no : 239

Typically used in OLTP environment.

Limited application for DSS and VLDB

In DSS environment we deal with VLDB and large sets of data. Traditionally Nested-Loop join has been and is used in OLTP environments, but for many reasons, such a join mechanism is not suitable for VLDB and DSS environments. Nested loop joins are useful when small subsets of data are joined and if the join condition is an efficient way to access inner table

Why students allowed by company to visit company data ware house Bit map join k table banana thaw

Answer: Ch: 27 Page:234

- **The index consists of bitmaps, with a column for each unique value:**

Index on City (larger table):

SID	Islamabad	Lahore	Karachi	Peshawar
1	0	1	0	0
2	1	0	0	0
3	0	1	0	0
4	0	0	0	1
5	0	0	1	0
6	0	0	1	0
7	0	0	0	1
8	0	0	0	1
9	0	1	0	0

Index on Tech (smaller table):

SID	CS	Elect	Telecom
1	1	0	0
2	0	1	0
3	0	1	0
4	1	0	0
5	0	0	1
6	0	1	0
7	0	0	1
8	1	0	0
9	1	0	0

Page dimension:
Answer:
The page dimension describes the page context for a Web page event. The grain of this dimension is the individual page. Our definition of page must be flexible enough to handle the evolution of Web pages from the current, mostly static page delivery to highly dynamic page delivery in which the exact page the customer sees is unique at that instant in time.

How can we come to know our e_mailetc campaign is successful?
Answer:
The success of a specific e-mail, marketing or adcampaign can be directly measured and quantified by integrating the Web log with other operational systems such as sales force automation (SFA), customer relationship management (CRM) and enterprise resource planning (ERP) applications.

Why web dwh?
Answer:
1. Searching the web (web mining).
2. Analyzing web traffic.
3. Archiving the web.

The three reasons for warehousing web data are as listed in the slide. First, web warehousing can be used to mine the huge web content for searching information of interest. It's like searching the golden needle from the haystack. Second reason of Web warehousing is to analyze the huge web traffic. This can be of interest to Web Site owners, for e-commerce, for e-advertisement and so on. Last but not least reason of Web warehousing is to archive the huge web content because of its dynamic nature.

Shared architecture?
Answer:
All Processors have equal access to the data stored on disk. Architecture can both dictate, and facilitate, our behaviors. The sharing Architecture enables better matching of workload to micro architecture resources by replacing static cores with virtual cores which can be

dynamically reconfigured to have different numbers of ALUs and amount of cache. The sharing Architecture across a benchmark suite of Apache, APECint and parts of PARSEC, The sharing Architecture is a fine-grain composable architecture which provides the flexibility to dynamically “synthesize” a flexible core out of the correct amount of ALUs, fetch bandwidth and cache, based on an applications demand and the need to optimize cost.

Which authority records pest population?
Answer:
Directorate of Pest Warning and Quality control of Pesticides (DPWQCP).

Where can we use nested loop?
We can use nested loop in OLTP systems it isn’t suitable to use in DSS environment. Well it can be used when small datasets is joined and if condition efficiently accessed the inner table. Another hand: A *nested loop* is a loop within a loop, an inner loop within the body of an outer one. How this works is that the first pass of the outer loop triggers the inner loop, which executes to completion. Then the second pass of the outer loop triggers the inner loop again. This repeats until the outer loop finishes. Of course, a *break* within either the inner or outer loop would interrupt this process.

DTS package?
Answer:
DTS contains a set of tools that provides a very easy approach to build a package and execute it. Writing or building a package through programming is a complex task but DTS tools like DTS Designer and Import/Export Wizard do this entire complex task for user just through a single click of button.
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.

Orr's law?
Answer:
Law #1: “Data that is not used cannot be correct!”
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!”

What is pest scouting?
Answer:
The term 'Pest scouting ' as it applies to the area of agriculture can be defined as ' Inspecting a field for pests, including insects, weeds, and pathogens. Pest scouting is a basic component of integrated pest management programs. It is used to determine whether pest populations are at levels that warrant control intervention and also may help to determine the most appropriate method of control'. Pest scouting is a systematic field sampling process that provide field specific

information on pest pressure and crop injury. The pest scouting data is being constantly recorded by the Directorate of Pest Warning and Quality Control of Pesticides (DPWQCP), Punjab since 1984. However, despite pest scouting, yield losses have been occurring. The most recent being the Boll Worm attack on the cotton crop during 2003-04, resulting in a loss of nearly 0.5 million bales.

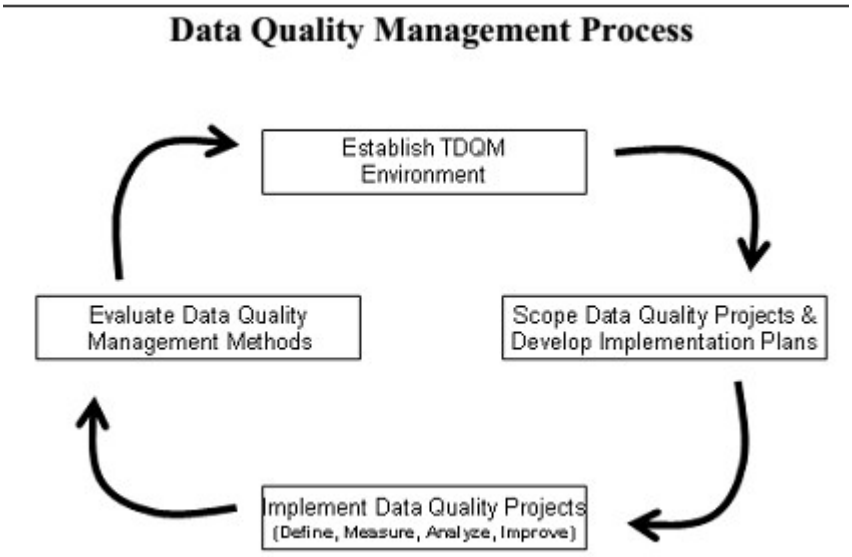
Amdahl's formula chose karnatha

Answer:

S≤1/ f+(1-f)/N f is the fraction of the problem that must be computed sequentially
N is the number of processors, As f approaches 0, S approaches N.

Quality management katha

Answer:



Quality Management Maturity Grid k stages guess karni thin.

Answer:

Quality Management Maturity Grid

	Management understanding	Quality organ. status	Problem handling	Cost of quality % of sales	Company attitude
Stage-1 Uncertainty	No comprehension of quality.	Quality Dept. part of manufacturing or engineering.	Fire fighting approach.	Reported unknown, actual high.	No organized activity.
Stage-2 Awakening	Recognize quality management may be of value.	Quality Dept. still part of manufacturing or engineering.	Short term solutions, no long term approach.	Reported as low, actually high.	All talk no real action.
Stage-3 Enlightenment	Become supportive and helpful.	Quality Dept. reports to top management.	Problems faced and solved orderly.	Reported as medium actually on the higher side.	Identifying and resolving problems
Stage-4 Wisdom	Understand absolutes of quality management.	Senior quality manager position.	Identified at an early stage.	Reported as about medium actually about medium.	Defect prevention is a routine.
Stage-5 Certainty	Quality management essential part of company policy.	Quality manager on board of directors.	Identified and resolved at an early stage.	Reported as low, actually is low.	Know why there are problems.

Table 23.1: Quality Management Maturity Grid

2. Limitation of Session ID ping pong? 3 marks

Answer: page 366: Requires a great deal of control over the Web site's page-generation methods. Approach breaks down if multiple vendors are supplying content in a single session. This method of session tracking requires a great deal of control over the Web site's page-generation methods to ensure that the thread of session ID is not broken. If the visitor clicks on links that don't support this method a single session will appear to be multiple sessions. This approach also breaks down if multiple vendors are supplying content in a single session:

4. Types of partitioning used in shared nothing environment? 3 marks

Answer:

1. Range Partitioning
2. Hash Partitioning
3. List Partitioning
4. Round Robin
5. Combination

5. Calculate Nested loop join cost, same example from lecture notes.5 marks

Answer:

Join cost = Cost of accessing Table_A+
# of qualifying rows in Table_A ×Blocks of Table_B to be scanned for each qualifying row
OR
Join cost = Blocks accessed for Table_A+

Blocks accessed for Table_A ×Blocks accessed for Table_B
A&B = No of blocks Access by A + No of qualified blocks by A* No of Blocs Accessed by B.
For example, if qualifying blocks for Table_A QB (A) = 50 and qualifying blocks for Table_B QB (B) = 100 and size of Table_A is 500 blocks and size of Table_B is 700 blocks then Join cost A&B = 500 + 50×700 = 35,500 I/Os and using the other order i.e. Table_B outer table and Table_A as inner table, the join cost B&A = 700 + 100×500 = 50,700 I/Os i.e. an increase in I/O of about 43%.

6. If an organization does not freeze requirements during development phase e.g it is too much accommodating then what are the implications? 5 marks
Answer:
You need to think like a software developer and manage three very visible stages of developing each data mart: (1) the business requirements gathering stage, where every suggestion is considered seriously, (2) the implementation stage, where changes can be accommodated~ but must be negotiated and generally will cause the schedule to slip, and (3) the rollout stage, where project features are frozen. In the second and third stages, you must avoid insidious scope creep (and stop being such an accommodating person).

7. Two tables were given one was employee table and other exception table with attribute IsAgeValid. Sql query was required to find the outliers not having age between 26 and 75 and set the dirtyBit in exception table to 0?5 marks
Answer:
Select * From (Select * From Employee Where DirtyBit= 0) IsAgeValid Where (Age <26 And >75))
Q. Outer Table:
Answer:
The outer table is usually the one that has:
• The smallest number of qualifying rows, and/or
• The largest numbers of I/Os required locating the rows.
In a left join, the outer table, the outer table and inner table are also referred to as the row-preserving and null-supplying tables, respectively.
In a right join, the outer table and inner table are the right and left tables, respectively.
For example, in the queries below, T1 is the outer table and T2 is the inner table:
T1 left join T2 or T2 right join T1 Or, using Transact-SQL syntax: T1 *= T2 or T2 =* T1.

Identify the following examples corresponds to which Data mining Technique:
Note: Review examples in Chapter no : 30

a. **Assigning customers to predefined customer segments (good vs. bad)**

Answer:
Classification

b. **Classifying instructor rating as excellent, very good, good, fair, or poor**

Answer:
Classification

c. **Building a model and assigning a value from 0 to 1 to each member of the set**

Answer:
Estimation

d. Predicting how much customers will spend during next 6 months.

Answer:
Prediction

Q. Why to save package in SQL Server 2000 Meta Data Services?
Answer: Page no: 387 Ch# lab lecture 1
Packages that are required for very complicated tasks are not trivial to build. To develop such packages, DTS Designer or programming tools are used. Once such packages are built they are saved for further use. We may edit these packages later on. For editing purposes either DTS designer or programming is used. After edit a package we may keep both packages that is package before editing and package after editing as two different versions of same package. To maintain version information packages are saved in "SQL Server 2000 Meta Data Services".

Q. Attributes of Page Dimension:
Answer: Page no : 362 Ch# 40

Details of W DWH Page Dimension

ATTRIBUTE	SAMPLE VALUES
Page Key	Surrogate values, 1-*N*
Page Source	Static, Dynamic, Unknown, Corrupted, Inapplicable
Page Function	Portal, Search, Product Description, Corporate Information
Page Template	Sparse, Dense
Item Type	Product SKU, Book ISBN Number, Telco Rate Type
Graphics Type	GIF, JPG, Progressive Disclosure, Size Pre-Declared, Combination
Animation Type	Similar to Graphics Type
Sound Type	Similar to Graphics Type
Page File Name	File Name

Q. Define Automatic Data Cleansing Techniques.
Answer: Page no: 164 Ch#19
1) **Statistical:** Identifying outlier fields and records using the values of mean, standard deviation, range, etc., based on Chebyshev's theorem, considering the confidence intervals for each field.
2) **Pattern Based:** Identify outlier fields and records that do not conform to existing patterns in the data. Combined techniques (partitioning, classification, and clustering) are used to identify patterns that apply to most records.
3) **Clustering:** Identify outlier records using clustering based on Euclidian (or other) distance. Existing clustering algorithms provide little support for identifying outliers. However, in some cases clustering the entire record space can reveal outliers that are not identified at the field level inspection.
4) **Association Rules:** Association rules with high confidence and support define a different kind of pattern. As before, records that do not follow these rules are considered outliers. The power of

association rules is that they can deal with data of different types. However, Boolean association rules do not provide enough quantitative and qualitative information.

Q. Identify the following Statements corresponds to which Data Quality Analysis Project activity:
Answer: Page no: 194 Ch#23

• **Identify** functional user data quality requirements and establish data quality metrics.
Answer: Define
• Measure conformance to current business rules and develop exception reports.
Answer: Measure

July 24, 2013 at 12:20pm
CS614 Final term latest papers
1. Identify give statement?
1. According to Bill Inmon, requirements are the first thing to be considered in the decision support development. Identify?
2. A top down approach is useful for project where problems must be solved are unclear.
Answer= First statement is false because, According to Inmon, requirements are the last thing to be considered in the decision support development lifecycle.
Second statement is also false because. 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.

2. Identify give statement? Correct and incorrect
1. Dimensional modeling is a physical design technique?

2. The concept of facts and dimensions are used in ERD?
Answer= first statement is false because, DM is a logical design technique that seeks to present the data in a standard, instinctive structure that supports high-performance and ease of understanding.
Second statement is also false because ROLAP tools will query the relational database using SQL generated to conform to a framework using the facts and dimensions paradigm using the star schema.

3. Roll up is a cube operation. Define this?
Answer= Rollup: summarize data
e.g. Given sales data, summarize sales for last year by product category and region**.**

Q. Reverse proxy?
Answer= Reverse Proxy
Another type of proxy server, called a reverse proxy, can be placed in front of our enterprise's Web servers to help them offload requests for frequently accessed content. This kind of proxy is entirely within our control and usually presents no impediment to Web warehouse data collection. It should be able to supply the same kind of log information as that produced by a Web server.

2 marks question

1. Importance of multidimensional analysis?
Answer= MULTIDIMENSIONAL is the key requirement to the letter and at the heart of the cube concept of OLAP. The system must provide a multidimensional logical view of the aggregated data, including full support for hierarchies and multiple hierarchies, as this is certainly the most logical way to analyze organizations and businesses. There is no “magical” minimum number of dimensions that must be handled as it is too application specific.

2. Intrinsic skew?
Answer= Intrinsic skew occurs when attributes are not distributed uniformly; it is also called attribute value skew. For example a basic Computer Science (CS) course being offered in summer, and taken by many non-CS majors who want to know about computers. The course taken by few CS-majors who missed it or got an incomplete (i.e. I) grade during the regular semester due to one reason or another. Ironically, intrinsic skew effects the performance of both hash and sort-merge joins**.**

3. Why analytic track is called as fun part?
Answer= As a well-respected application developer once told, "Remember, this is the fun part!"

5 marks

Q. Claude Shannon's theory statement is given and asked about correct or incorrect.
Answer= Claude Shannon’s theory states that as the volume increases the information content decreases and vice versa.

Q. Statistics is assumption driven? Correct or incorrect
Answer= 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.

Q. reasons for underutilization of DWH in agriculture
answer= Thus the lack of data integration and standardization contributes to an under-utilization of historical data, and inevitably results in an inability to perform any scientific predictive analysis for effective decision support and policy making. The implementation of a Pilot Agriculture Data Warehouse (Agri DWH) is discussed. Such a data warehouse can support decision making using Data Mining and Online Analytical Processing (OLAP). Based on literature review, no such work was found to have been undertaken in the agriculture sector of Pakistan and elsewhere.

Q. Explain two precedence constraints? Or DTS precedence rule? (5 marks)
answer= Precedence constraints sequentially link tasks in a package. In DTS, you can use three types of precedence constraints, which can be accessed either through DTS Designer or programmatically:
Unconditional: If you want Task 2 to wait until Task 1 completes, regardless of the outcome, link Task 1 to Task 2 with an unconditional precedence constraint.
On Success: If you want Task 2 to wait until Task 1 has successfully completed, link Task 1 to Task 2 with an On Success precedence constraint.
On Failure: If you want Task 2 to begin execution only if Task 1 fails to execute successfully, link Task 1 to Task 2 with an On Failure precedence constraint. If you want to run an alternative branch of the workflow when an error is encountered, use this constraint.

Q. Explain Prediction?
Answer= Same as classification or estimation except records are classified according to some predicted future behavior or estimated value. Using classification or estimation on a training example with known predicted values and historical data a model is built. Then explain the known values, and use the model to predict future. **Example:** Predicting how much customers will spend during next 6 months.

Q. It is believed that the agricultural data0collectd in Punjab has underutilized for decades. In your opinion, what could be the reasons of this under utilization? Also suggest how can we utilize this data in a proper way? (5 marks)
Answer= Each year different government departments and agencies in Pakistan create tens of thousands of digital and non digital files from thousands of pest-scouting surveys, yield surveys, metrological data collection, river flows etc. This data collection has been going on for decades. The data collected has never been compiled, standardized and

integrated to give a complete picture. Thus the lack of data integration and standardization contributes to an under-utilization of historical data, and inevitably results in an inability to perform any scientific predictive analysis for effective decision support and policy making. An Agriculture data warehouse is the answer, as processing 100+ variables by experts for large historical data is not possible. This has repeatedly resulted in tragic outcomes.

Q. What is classification? How it can be used for a News website? Explain with and example? (5 Marks)

Answer= Classification means that based on the properties of existing data, we have made or
groups i.e. we have made classification.
Example is of a news site, where there are number of visitors and also many content developers. Now where to place a specific news item on the web site? What should be the hierarchical position of the news item, what should be the news chapter, category? Either it should be in the sports or weather section and so on. What is the problem in doing all this? The problem is that it's not a matter of placing a single news item. The site as already mentioned contains a number of content developers and also many categories. If sorting is performed humanly, then it is time consuming. That is why classification techniques can scan and process the document to decide its category or class. It is not possible and there are flaws in assigning category to any news document just based on the keyword. Frequent occurrence of the word keyword cricket in a document doesn't necessary means that the document be placed in the sports category. The document may be actually political in nature.

Q. List down the three drawbacks of traditional web search? (3 Marks)

answer= Drawbacks of traditional web searches
1. Limited to keyword based matching.
2. Cannot distinguish between the contexts in which a link is used.
3. Coupling of files has to be done manually.

1. Difference between knowledge and intelligence? (2marks)

answer= "Knowledge is power, Intelligence is absolute power!"
Knowledge, on the other hand is an application of information and data, and gives an insight by answering the "how" questions. Knowledge is also the understanding gained through experience or study. Intelligence is appreciation of "why", Remember knowledge is power.

2. What is non-trivial information? (2 marks)

answer= by nontrivial, we mean that some search or inference is involved; that is, it is not a
straightforward computation of predefined quantities like computing the average value of a set of numbers. If the sale of some items boosts up, even when no Eid was around, in some region of the country, this is non-trivial information

3. List down any four ways to identify the session in World Wide Web??? (2marks)

answer= Identifying the Session

?Web-centric data warehouse applications require every visitor session (visit) to have its own unique identity

?The basic protocol for the World Wide Web, HTTP, stateless so session identity must be established in some other way.

?There are several ways to do this

?Using Time-contiguous Log Entries

?Using Transient Cookies

?Using HTTP's secure sockets layer (SSL)

?Using session ID Ping-pong

?Using Persistent Cookies

5. Given statement is correct or incorrect explain in either case(3 marks)

“in the process of normalization number of tables is always reduced to minimum”

Answer= this statement is incorrect because
De-normalization specifically improves performance by :

?Reducing the number of tables and hence the reliance on joins, which consequently speeds up performance. The higher the level of normalization, the greater will be the number of tables

6. Given statement is correct or incorrect explain in either case (3marks)

“in prediction test data set is used to form the model and the associated rules”

Answer= the above statement is false because The training set is used to form model and the associated rules. Once model built and rules defined, the test set is used for grouping.

7. Why web data warehousing is required? Briefly explain three reason (3marks)

Answer= Reasons for web warehousing

1. Searching the web (web mining).
2. Analyzing web traffic.
3. Archiving the web.

The three reasons for warehousing web data are as listed in the slide. First, web warehousing can be used to mine the huge web content for searching information of interest. It's like searching the golden needle from the haystack. Second reason of Web warehousing is to analyze the huge web traffic. This can be of interest to Web Site owners, for e-commerce, for e-advertisement and so on. Last but not least reason of Web warehousing is to archive the huge web content because of its dynamic nature.

8. can we identify the session in world wide web by using "persistent cookies" however there are some limitation of this technique .explain two limitation (3marks)

answer= Limitations

• It's possible that the visitor will have his or her browser set to refuse cookies or may clean out his or her cookie file manually, so there is no absolute guarantee that even a persistent cookie will survive.

• Although any given cookie can be read only by the Web site that caused it to be created, certain groups of Web sites can agree to store a common ID tag that would let these sites combine their separate notions of a visitor session into a super session

12. Briefly explain the lesson which we learn concluded in agriculture data warehouse case study. (5 m

Answer= in this case study, the implementation of a Pilot Agriculture Data Warehouse (Agri DWH) is discussed. Such a data warehouse can support decision making using Data Mining and Online Analytical Processing (OLAP). Based on literature review, no such work was found to have been undertaken in the agriculture sector of Pakistan and elsewhere. Data warehouses are quite popular in telecommunications, travel industry, government etc. but an application in agriculture extension is a novel idea. The strength of this novel idea is demonstrated through a pilot implementation and discussion of interesting findings using real data.

Q. first two phases of Kimball DWH process name them

answer= 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).

<u>April 2012:</u>

Q. writes a quarry to extract total number of female students registered in BS Telecom. 5

Answer: Total Number of Female students in BS Telecom

```
SELECT COUNT (DISTINCT r.SID) AS Expr1
FROM Registration r INNER JOIN
Student s ON r.SID = s.SID AND
s.[Last Degree] IN ('F.Sc.', 'FSc',
'HSSC', 'A-Level', 'A level') AND
r.Discipline = 'TC' AND s.Gender = '1‘
```

Or another solution:

```
SELECT COUNT(SID)
FROM   REGISTRATION,STUDENT
WHERE  REGISTRATION.SID = STUDENT.SID
AND    DISCIPLINE = 'TC'
AND    GENDER = '1'
```

Q. Describe the lessons learn at during Agri-data ware house case study? Page 347

• Extract Transform Load (ETL) of agricultural extension data is a big issue. There are no digitized operational databases so one has to resort to data available in typed (or hand written) pest scouting sheets. Data entry of these sheets is very expensive, slow and prone to errors.

• Particular to the pest scouting data, each farmer is repeatedly visited by agriculture extension people. This results in repetition of information, about land, sowing date, variety etc (Table-2). Hence, farmer and land individualization are critical, so that repetition may not impair aggregate queries. Such an individualization task is hard to implement for multiple reasons.

• There is skewness in the scouting data. Public extension personnel (scouts) are more likely to visit educated or progressive farmers, as it makes their job of data collection easy. Furthermore, large land owners and influential farmers are also more frequently visited by the scouts. Thus the data does not give a true statistical picture of the farmer demographics.

• Unlike traditional data warehouse where the end users are decision makers, here the decision-making goes all the way “down” to the extension level. This presents a challenge to the analytical operations’ designer, as the findings must be fairly simple to understand and communicate.

Q. What are the fundamental strengths and weakness of k means clustering?

• **Relatively efficient:** O (tkn), where n is # objects, k is # clusters, and t is # iterations. Normally, k, t << n.

• Often terminates at a local optimum. The global optimum may be found using techniques such as: deterministic annealing and genetic algorithms

Weakness:

• Applicable only when mean is defined, then what about categorical data?

• Need to specify k, the number of clusters, in advance

• Unable to handle noisy data and outliers

Q. Data profiling is a process of gathering information about columns, what are the purpose that it must fulfill?

Data profiling is a process which involves gathering of information about column

through execution of certain queries with intention to identify erroneous records. In this process we identify the following:
• Total number of values in a column
• Number of distinct values in a column
• Domain of a column
• Values out of domain of a column
• Validation of business rules
we run different SQL queries to get the answers of above questions. During this process we can identify the erroneous records. Whenever we will come across an erroneous record, we will just copy it in error or exception table and set the dirty bit of record in the actual student table. Then we will correct the exception table. After this profiling process we will transform the records and load them into a new table Student_Info
Ref: Handout Page No. 354

Q. Define additive and non additive facts
Additive facts are those facts which give the correct result by an addition operation. Examples of such facts could be number of items sold, sales amount etc. Non-additive facts can also be added, but the addition gives incorrect results. Some examples of non-additive facts are average, discount, ratios etc. **Ref: Handout Page No. 119**

Q. What are three fundamental reasons for warehousing web data?
1. Searching the web (web mining).
2. Analyzing web traffic.
3. Archiving the web.
First, web warehousing can be used to mine the huge web content for searching information of interest. It's like searching the golden needle from the haystack. Second reason of Web warehousing is to analyze the huge web traffic. This can be of interest to Web Site owners, for e-commerce, for e-advertisement and so on. Last but not least reason of Web warehousing is to archive the huge web content because of its dynamic nature. **Ref: Handout Page No. 348**

Q. What are the two basic data warehousing implementation strategies and their suitability conditions?
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. **Ref: Handout Page No. 283**

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 ith bit is set to 1 if the ith row of the base table has the value for the indexed column. **Ref: Handout Page No. 233**

Q. List and explain fundamental advantages of bit map indexing. Page 235
• Very low storage space.
• Reduction in I/O, just using index.
• Counts & Joins
• Low level bit operations.
An obvious advantage of this technique is the potential for dramatic reductions in storage overhead. Consider a table with a million rows and four distinct values with column header of 4 bytes resulting in 4 MB. A bitmap indicating which of these rows are for these values requires about 500KB.

Q. List and explain fundamental disadvantages of bit map indexing. Page 236
• Locking of many rows
• Low cardinality
• Keyword parsing
• Difficult to maintain - need reorganization when relation sizes change (new bitmaps)
Therefore, a bitmap is practical only for low- cardinality columns that divide the data into a small number of categories, such as "M/F", "T/F", or "Y/N" values.
Keyword parsing: Bitmap indexes can parse multiple values in a column into separate keywords. For example, the title “Marry had a little lamb" could be retrieved by entering the word “Marry" or “lamb" or a combination. Although this keyword parsing and lookup capability is extremely useful, textual fields tend to contain high-cardinality data (a large number of values) and therefore are not a good choice for bitmap indexes.

Q. What are major operations of data mining? Page 259
• Classification
• Estimation
• Prediction
• Clustering
• Description

Q. What will be the effect if we program a package by using DTS object model?

DTS package is exactly like a computer program. Like a computer program DTS package is also prepared to achieve some goal. Computer program contains set of instructions whereas DTS package contains set of tasks. Tasks are logically related to each other. When a computer program is run, some instructions are executed in sequence and some in parallel. Likewise when a DTS package is run some tasks are performed in sequence and some in parallel. The intended goal of a computer program is achieved when all instructions are successfully executed. Similarly the intended goal of a package is achieved when all tasks are successfully accomplished
Package can also be programmed by using DTS object model instead of using graphical tools but DTS programming is rather complicated.

Another solution: First you will need to determine the source of the data. Depending on your selection, you might have to provide additional authentication information. For example, when importing data from another SQL Server database, you might be able to use Windows domain accounts, instead of SQL Server logins, while selecting Access or Oracle will force you to deal with different authentication choices. The choice of data source will also affect the available Advanced Connection Properties (which are OLE DB provider specific) displayed after clicking the advanced button on the Choose a Data Source page of the wizard. On the next page of the wizard, you will be prompted for equivalent configuration options for destination of data transfer (including provider type and advanced connection properties).After you specify the source and destination, We will be asked to select one of three types of data that will be imported/exported:
Ref: Handout Page No. 381

Q. Write down the steps of handling skew in range partitioning? Page 218

• Sort
• Construct the partition vector
• Duplicate entries or imbalances

There are number of ways to handle the skew in the data when it is partitioned based on the range; here date is a good example with data distributed based in quarters across four processors. One solution is to sort the data this would identify the “clusters” within the data, then bases on them more or less equal partitions could be created resulted in elimination or reduction of skew.

Q. what type of anomalies exists if a table is in 2NF not in 3NF? [2] Page 46

The table is in 2NF but NOT in 3NF Tables in 2NF but not in 3NF contain modification anomalies

Q. What are three methods for creating a DTS package? Page 381

–Import/Export wizard
–DTS Designer
–Programming DTS applications

Q. Write two extremes of Tech. Arch Design? Page 229

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

Q1: Explain analytic data application specification in Kimball 5 marks page 219

• 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

Analytic applications development

• **Standards:** naming, coding, libraries etc.

• Coding begins AFTER DB design complete, data access tools installed, subset of historical data loaded.

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

• Benefits: Quality problems will be found with tool usage => staging.

• Actual performance and time gauged.

Q3: 2 real life examples of clustering 5 marks page 264

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 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 occur during data acquisition and cleansing in agriculture case study? 3marks page 340

• The pest scouting sheets are larger than A4 size (8.5" x 11"), hence the right end was cropped when scanned on a flat-bed A4 size scanner.

• The right part of the scouting sheet is also the most troublesome, because of pesticide names for a single record typed on multiple lines i.e. for multiple farmers.

• As a first step, OCR (Optical Character Reader) based image to text transformation

of the pest scouting sheets was attempted. But it did not work even for relatively clean sheets with very high scanning resolutions.
• Subsequently DEO's (Data Entry Operators) were employed to digitize the scouting sheets by typing.

Q6: Meant of classification process, how measure accuracy of classification? 3marks page 276
First of the available data set is divided into two parts, one is called test set and the other is called the training set. We pick the training set and a model is constructed based on known facts, historical data and class properties as we already know the number of classes. After building the classification model, every record of the test set is posed to the classification model which decides the class of the input record. It should be noted that you know the class for each record in test set and this fact is used to measure the accuracy or confidence level of the classification model. You can find accuracy by Accuracy or confidence level = matches/ total number of matches
In simple words, accuracy is obtained by dividing number of correct assignments by total number of assignments by the classification model

Q7: Data parallelism explain with example 3 marks page 212
• 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.

Q8: Under what condition an operation can be execute in parallel? 3 marks
under the things which can be divided into two such as with reference to size and with reference to divide and conquer an operation can be execute in parallel. Divide and conquer means that we should be able to divide the problem and then solve it and then compile the results i.e. conquer. For example in case of scanning a large table every row has to be checked, in such a case this can be done in parallel thus reducing the overall time. There can be and are many examples too.

Q9: What sorts of objectives metric are use by companies what are possible issues in formulation these metric? 2 marks
Metrics developed for internal corporate use generally provide information on operations or

management issues, but many also provide information useful to external stakeholders. Energy use, expressed either in absolute terms or on a per-unit-of-product basis, is an example of a measure of interest to both corporate managers and external stakeholders. The latter group (e.g., customers, regulators, investors, environmental groups) generally is interested in many internal corporate metrics and is concerned about the impacts of industry activities on the environment at the local, regional, and global levels. These concerns frequently relate to such issues as air quality, water quality, product recyclability, and regulatory compliance.

Q10: Which script languages are used to perform complex transformation in DTS package? 2 marks

Microsoft SQL Server provides graphical tools to build DTS packages. These tools provide good support for transformations. Complex transformations are achieved through VB Script or Java Script that is loaded in DTS package. Package can also be programmed by using DTS object model instead of using graphical tools but DTS programming is rather complicated.

Q11: Cleansing can be breaking down in whom many steps, write their names? 2

One can break down the cleansing into six steps: element zing, standardizing, verifying, matching, house holding, and documenting.

Q12: What does 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 a single winner is left, it is a good piece of advice that always keeps at least two. What if you keep one? The sole vendor may take benefit of the situation that he is the only player and create a situation favorable for him. He might get an upper hand in the bargaining process, and mold things according to his facility and benefit. To avoid such a situation enlist a competitor too, even if a single vendor is the winner. This will create a competitive environment which may ultimately turn into your favor.

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

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.

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

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

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).

5) Dirty bit?2

• Add a new column to each student table
• This new column is named as “Dirty bit”
• It can be Boolean type column
• This column will help us in keeping record of rows with errors, during data profiling

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

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.
•

7) Indexing using I/0 bottelneck?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.

9) W8 is Click stream? Limitations?3

• Click stream is every page event recorded by each of the company's Web servers
• Web-intensive businesses
• Although most exciting, at the same time it can be the most difficult and most frustrating.
• Not JUST another data source.

Click stream data has many issues.

1. Identifying the Visitor Origin
2. Identifying the Session
3. Identifying the Visitor
4. Proxy Servers
5. Browser Caches

10) Import/export wizard tasks?3

• First of all load data

1. Connect to source Text files
2. Connect to Destination SQL Server
3. Create new database ‘Lahore_Campus’
4. Create two tables Student & Registration
5. Load data from the text files containing student information into Student table
6. Load data from the text files containing registration records into Registration table

• Import/Export Wizard is sufficient to perform all above mentioned tasks easily

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

12) How data mining is different from statics? which one is better?5

Statistics is the study of the collection, organization, analysis, interpretation and presentation of data. While interpreting and analyzing large sets of data, for prediction and pattern recognition purposes, Data Mining is used. Data Mining is just a computational method for statistical analysis. Since Data Mining and Statics are conceptually different from each other and cannot be compared to each other.

• 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. It includes everything from planning for the collection of data and subsequent data management to end of the line activities such as drawing inferences from numerical facts called data and presentation of results. Statistics is concerned with one of the most basic of human needs.
• Data mining builds many complex, predictive, nonlinear models which are used for predicting behavior impacted by many factors. Data mining is used to discover patterns and relationships in your data in order to help you make better business decisions. Data mining cannot be ignored the data are there, the methods are numerous.

13) Persistent cookies limitations? 5

Using Persistent Cookies: Establish a persistent cookie in the visitor's PC. The Web site may establish a persistent cookie in the visitor's PC that is not deleted by the browser when the session ends.
Limitations: • No absolute guarantee that even a persistent cookie will survive.
• Certain groups of Web sites can agree to store a common ID tag

Q. Misconception about data quality

1) You Can Fix Data
2) Data Quality is an IT Problem
3. All Problems is in the Data Sources or Data Entry
4. The Data Warehouse will provide a single source of truth
5. Compare with the master copy will fix the problem

Q. Classification and estimation

• Classification consists of examining the properties of a newly presented observation and assigning it to a predefined class.
• Assigning customers to predefined customer segments (good vs. bad)
• Assigning keywords to articles
• Classifying credit applicants as low, medium, or high risk
• Classifying instructor rating as excellent, very good, good, fair, or poor

ESTIMATION: As opposed to discrete outcome of classification i.e. YES or NO, deals with continuous valued outcomes

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

Traditionally Nested-Loop join has been and is used in OLTP environments, but for many reasons, such a join mechanism is not suitable for VLDB and DSS environments. Nested loop joins are useful when small subsets of data are joined and if the join condition is an efficient way of accessing the inner table.

Nested-Loop Join: Variants

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

Q. Define Dense and Sparse index, adv and disadvantage (3)

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

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

Sparse Index: normally only one key per data block is kept. A sparse index uses less space at the expense of somewhat more time to find a record given its key.

What happens when record 35 is inserted?

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.
• Time (I/Os) logarithmic in the number of blocks used by the index.

Sparse Index: Multi level

Q. What should b dune in the case where golden copy is missing dates?

If the dates are missing we must need to consult golden copy. If gender is missing we are not required to consult golden copy. In many cases name can help us in identifying the gender of the person.

Q. Tasks performed through import/export data wizard

Tasks can be as follows:
• Establish connection through source / destination systems
• Creates similar table in SQL Server
• Extracts data from text files
• Apply very limited basic transformations if required
• Loads data into SQL Server table

Q. Transient cookies
• Let the Web browser place a session-level cookie into the visitor's Web browser.
• Cookie value can serve as a temporary session ID
Limitations: You can't tell when the visitor returns to the site at a later time in a new session.

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

Q. What is the difference between training data and test data?
The existing data set is divided into two subsets, one is called the training set and the other is called test set. The training set is used to form model and the associated rules. Once model built and rules defined, the test set is used for grouping. It must be noted the test set groupings are already known but they are put in the model to test its accuracy.

Q. Why building a data warehouse is a challenging activity? What are the three broad categories of data warehouse development methods?
Building a data warehouse is a very challenging job because unlike software engineering it is quite a young discipline, and therefore, does not yet has well-established strategies and techniques for the development process. Majority of projects fail due to the complexity of the development process. To date there is no common strategy for the development of data warehouses; they are more of an art than science. Current data warehouse development methods can fall within three basic groups: data-driven, goal driven and user-driven.

1. **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.
2. **RAD:**
   Rapid Application Development (RAD) is an iterative model consisting of stages like scope, analyze, design, construct, test, implement, and review. It is much better suited to the development of a data warehouse because of its iterative nature and fast iterations.

3. **Spiral Model:**
   The model is a sequence of waterfall models which corresponds to a risk oriented iterative enhancement, and recognizes that requirements are not always available
   and clear when the system is first implemented

Ref: Handout Page No. 283

Q. 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)

Q. 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.

Q. 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.

Q. Explain Analytic Applications Development Phase of Analytic Applications Track of Kimball's Model? Page 299

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

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.

Q. 5 tech for de normalization (names)
Areas for Applying De-Normalization Techniques: Dealing with the abundance of star schemas. The fast access of time series data for analysis. Fast aggregate (sum, average etc.) results and complicated calculations. Multidimensional analysis (e.g. geography) in a complex hierarchy. Dealing with few updates but many joins queries. De-normalization will ultimately affect the database size and query performance.

Q. what is the difference between data matrix and similarity/dissimilarity in terms of rows and columns, which one is symmetric?
Data matrix
- We can measure the similarity of the row1 in data matrix with itself that will be 1.
- 1 is placed at index 1, 1 of the similarity matrix.
- We 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.
- Your data matrix has n rows and m columns then your similarity matrix will have n rows and n columns.

Similarity/dissimilarity
- Similarity or dissimilarity matrix is the measure the similarity
- time complexity of computing similarity/dissimilarity matrix
- M accounts for the vector or header size of the data.
- measure or quantify the similarity or dissimilarity

- Pearson correlation and Euclidean distance

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
Ref: Handout Page No. 270

Q. Write ten common mistakes which occur during the development of the data warehouse
We will avoid common mistakes that may halt data warehousing process.

1. Project proceeded for two months and nobody has touched the data.
2. End users are not involved hands-on from day one throughout the program.
3. IT team members doing data design (modelers and DBAs) have never used the access tools.
4. Summary tables defined before raw atomic data is acquired and base tables have been built.
5. Data design finished before participants have experimented with tools and live data.

Ref: Handout Page No. 311

Q. Write formula of pipeline speed up?
Speedup = Pipeline depth Clock Cycle$_{unpipelined}$
1 + Pipeline stall CPI Clock Cycle$_{pipelined}$
Link : http://www.google.com.pk/#hl=en&q=write+formula+of+pipeline+speed+up%3F....&oq=write+formula+of+pipeline+speed+up%3F....&gs_l=serp.3...1099734.1099734.7.1099972.1.1.0.0.0.0.0.0..0.0...0.0...1c.nvYnltI75y8&bav=on.2,or.r_gc.r_pw.,cf.osb&fp=6ed243b376015671&biw=1152&bih=720

Q. how data can be load in DTS
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.
Ref: Handout Page No. 373

Q. Define nested loop join? What are types of nested loop join? Explain them
We define a join of two tables to be a process in which we combine rows of one table with rows of another. There are three types of nested loop join
1. Naive nested-loop joins
2. Index nested-loop joins
3. Temporary index nested-loop joins

Ref: Handout Page No. 243

Q. Write the steps to create packages in DTS?
Package can be created by one of the following three methods:
–Import/Export wizard
–DTS Designer
–Programming DTS applications

Ref: Handout Page No. 380

Q. What is Inverted Index in simple words?
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).
Ref: Handout Page No. 232

Q. DTS Operation? Explain
A set of tools for
– Providing connectivity to different databases
– Building query graphically
– Extracting data from disparate databases
– Transforming data
– Copying database objects
– Providing support of different scripting languages(by default VB-Script and J-Script)
Ref: Handout Page No. 375

Q. Difference in between 1 way and 2 way clustering
1. One-way Clustering-means that when you clustered a data matrix, you used all the attributes. In this technique a similarity matrix is constructed, and then clustering is performed on rows. A cluster also exists in the data matrix for each corresponding cluster in the similarity matrix.
2. Two-way Clustering/Biclustering-here rows and columns are simultaneously clustered. No any sort of similarity or dissimilarity matrix is constructed. Biclustering gives a local view of your data set while one-way clustering gives a global view. It is possible that you first take global view of your data by performing one-way clustering and if any cluster of interest is found then you perform two-way clustering to get more details. Thus both the methods complement each other.
Ref: Handout Page No. 271

Q. Why should companies entertain students to visit their company's place?

- You are students, and whom you meet were also once students.
- You can do an assessment of the company for DWH potential at no cost.
- Since you are only interested in your project, so your analysis will be neutral.
- Your report can form a basis for a professional detailed assessment at a later stage.
- If a DWH already exists, you can do an independent audit
- The first and the foremost reason to get help is, whom you are talking to was once a student too as you are currently, so there is a common bond. Since you have studied well DWH (hopefully), therefore, you can do a requirement assessment (in the form of lifecycle study) of the company at no cost; the company has nothing to lose.

Ref: Handout Page No. 328

Q. In case of non-uniform distribution, what will be the impact on performance?

Parallelization is based on the premise that there is a full utilization of the processors and all of them are bust most or all of the time. However, if there is a skew in the partitioning of data i.e. a non-uniform distribution, then some of the processors will be working while other will be idle. And the processor that takes the most time (which has the most data too) will become the bottleneck.
Ref: Handout Page No. 219

Q. 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 no 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.
Ref: Handout Page No. 95

Q. 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

Q. What is fully de-normalized and highly de-normalized in DWH
'De-normalization' does not mean that anything and everything goes. De-normalization does not mean chaos or disorder or indiscipline. The development of properly de-normalized data structures follows software engineering principles, which insure that information will not be lost. De-normalization is the process of selectively transforming normalized relations into un-normalized physical record specifications, with the aim of reducing query processing time. Another fundamental purpose of de-normalization is to reduce the number of physical tables that must be accessed to retrieve the desired data by reducing the number of joins required to answer a query.
Ref: Handout Page No. 50

Kimball approach k knsy steps hain handouts ma page 285
Answer: 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.

8. If organization has not a subject expert in data modeling then what will the implications of it on organization?
Answer: Without this person, it becomes difficult to get a definitive answer on many of the questions, and the entire project gets dragged out, as the end users may not always be available.

Because it is essential to have a subject matter expert as part of the data modeling team. This person can be an outside consultant or can be someone in-house with extensive industry experience.

Q. What is Data Cardinality?

Cardinality is the term used in database relations to denote the occurrences of data on either side of the relation.

There are 3 basic types of cardinality:

High data cardinality:

Values of a data column are very uncommon.
e.g.: email ids and the user names

Normal data cardinality:

Values of a data column are somewhat uncommon but never unique.
e.g.: A data column containing LAST_NAME (there may be several entries of the same last name)

Low data cardinality:

Values of a data column are very usual.
e.g.: flag statuses: 0/1

Determining data cardinality is a substantial aspect used in data modeling. This is used to determine the relationships

Types of cardinalities:

The Link Cardinality - 0:0 relationships
The Sub-type Cardinality - 1:0 relationships
The Physical Segment Cardinality - 1:1 relationship
The Possession Cardinality - 0: M relation
The Child Cardinality - 1: M mandatory relationship
The Characteristic Cardinality - 0: M relationship
The Paradox Cardinality - 1: M relationship.

Q. What is a bitmap index?

A bitmap index is a specialized variation of a B-tree index. If the degree of cardinality is high for the attribute, means that there is more unique number of values for a particular attribute. Low cardinality attribute is not suitable for bitmap index because more number of records are locked which result in the locking of a whole table, leading to the lock on a whole database. For eg. A gender column, which has only two distinct values (male and female), is optimal for a bitmap index. However, data warehouse administrators also build bitmap indexes on columns with higher cardinalities.

You can use a bitmap index when both of the following conditions are true:
The key values in the index contain many duplicates.
More than one column in the table has an index that the optimizer can use to improve performance on a table scan.

Each bit in the bitmap corresponds to a possible rowid, and if the bit is set, it means that the row with the corresponding rowid contains the key value. A mapping function converts the bit position to an actual rowid, so that the bitmap index provides the same functionality as a regular index. Bitmap indexes store the bitmaps in a compressed way. If the number of distinct key values is

small, bitmap indexes compress better and the space saving benefit compared to a B-tree index becomes even better
Note: When creating bitmap indexes, you should use NOLOGGING and
COMPUTE STATISTICS. In addition, you should keep in mind that bitmap indexes are usually easier to destroy and re-create than to maintain.

ADVANTAGES

The Advantages of using bitmap indexes are greatest for columns in which the ratio of the number of distinct values to the number of rows in the table is small
Space requirements for indexes in a warehouse are often significantly larger than the space needed to store the data, especially for the fact table and particularly if the indexes are B*trees. Hence, you may want to keep indexing on the fact table to a minimum. Typically, you may have one or two concatenated B*tree indexes on the fact table; however, most of your indexes should be bitmap indexes. Bitmap indexes also take up much less space than B*tree indexes and so should be preferred

Q. What is B-tree index?

B-tree indexes are most commonly used in a data warehouse to enforce unique keys. In many cases, it may not even be necessary to index these columns in a data warehouse, because the uniqueness was enforced as part of the preceding ETL processing, and because typical data warehouse queries may not work better with such indexes. B-tree indexes are more common in environments using third normal form schemas. In general, bitmap indexes should be more common than B-tree indexes in most data warehouse environments.

Q2: Time complexity of K-means algorithm is O(tkn) what does t, k, and n represents here? Page 281

Answer: Relatively efficient: O(tkn), where n is # objects, k is # clusters, and t is # iterations. Normally, k, t<< n.

Q4: List down any two parallel software Architectures?

Answer: Shared Memory, Shard Disk and Shared Nothing.

Q6: which scripting language is used to perform complex transformations in Data packages?

Answer: Microsoft SQL Server provides graphical tools to build DTS packages. These tools provide good support for transformations. Complex transformations are achieved through VB Script or Java Script that is loaded in DTS package. Package can also be programmed by using DTS object model instead of using graphical tools but DTS programming is rather complicated.

Q7: "Dense index consist of a number of bit vector" justify it.

Answer: Dense Index: Every key in the data file is represented in the index file. Bitmap index record (Value, Bit Vector): Bit Vector has one bit for every record in the file, ith bit of Bit Vector

is set iff record i has Value in the given column. Bit vectors typically compressed. Converted to sets of rids during query evaluation.

Q8: It is essential to have a sub-matter expert as part of data modeling team. What will be the implication if such expert is not present in organization?

Answer: It is essential to have a subject-matter expert as part of the data modeling team. This person can be an outside consultant or can be someone in-house with extensive industry experience. Without this person, it becomes difficult to get a definitive answer on many of the questions, and the entire project gets dragged out, as the end users may not always be available.

Q. RI problem is peculiar to a DWH, this will not happen in a traditional OLTP system. Describe reason?

While doing total quality measurement, you measure RI every week (or month) and hopefully the number of orphan records will be going down, as you will be fine tuning the processes to get rid of the RI problems. Remember, **RI problem is peculiar to a DWH, this will not happen in a traditional OLTP system.**

1. **Be a diplomat not technologist.**

Answer: 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).

2. **List down activities of the planning design phase.**

Answer: 1. Determine Users' Needs
2. Determine DBMS Server Platform
3. Determine Hardware Platform
4. Information & Data Modeling
5. Construct Metadata Repository

3. **A data ware house is more like scientific research than anything traditional information system agree?**

Answer: I am not agreeing with this statement. Because, **a data ware house project is more like scientific research than anything traditional information system.** The normal Information System (IS) approach emphasizes on knowing what the expected results are

before committing to action. In scientific research, the results are unknown up front, and emphasis is placed on developing rigorous, step-by-step process to uncover the truth. The activities involve regular interactions between the scientist and the subject and also among the project participants. It is advised to adopt an exploratory, hands-on process involving cross-disciplinary participation.

4. **Query for Nested loop to select the outer table**

Answer: SELECT C.CustomerID, c.TerritoryID
FROM Sales.SalesOrderHeader oh
JOIN Sales.Customer c
ON c.CustomerID = oh.CustomerID
WHERE c.CustomerID IN (10, 12)
GROUP BY C.CustomerID, c.TerritoryID

5. **Before sitting down with business community together it suggestions to set you a productive session. List at least three activities that you will consider a part of request preplanned phase.**

Answer: Before sitting down with the business community to gather requirements, it is suggested to set you up for a productive session by considering the following:
Choose the Forum: There are two primary techniques for gathering requirements i.e. interviews or facilitated sessions. Both have their advantages and disadvantages. Interviews encourage lots of individual participation. They are also easier to schedule. Facilitated sessions may reduce the elapsed time to gather requirements, although they require more time commitment from each participant.
Identify and Prepare the Requirements Team: Regardless of the approach, you need to identify and prepare the project team members who are involved. If you're doing interviews, you need to identify a lead interviewer whose primary responsibility is to ask the great open-ended questions. Meanwhile, the interview scribe takes copious notes. Before you sit down with users, you need to make sure you're approaching the sessions with the right mindset.
Business Representatives Soliciting: If this is first foray into data warehousing, talk to business people that represent horizontal breadth across the organization. This coverage is critical to formulating the data warehouse bus matrix blueprint. Within the target user community, one should cover the organization vertically.
Scheduling: Schedule representatives nicely. The scheduler needs to allow ½ hour between meetings for debriefing and other necessities. Interviewing is extremely taxing because you must be completely focused for the duration of the session. Consequently, it is recommended to schedule three to four sessions in a day because the interviewers get very tired after that, and productivity goes down.
Preparing: The optimal approach is to conduct a project launch meeting with the users. The launch meeting disseminates a consistent message about the project. The interview team must prepare the interviewees by highlighting the topics to be covered in the upcoming session. It is advised that do not include a copy of the questionnaire, which is not intended for public dissemination.

Q11) Tell me what are the steps involved in Application Development?
A11) In Application Development, we usually follow following steps:
ADDTIP.

a) A - Analysis or User Requirement Gathering
b) D - Designing and Architecture
c) D - Development
d) T - Testing (which involves Unit Testing, System Integration Testing,
UAT - User Acceptance Testing)
e) I - Implementation (also called deployment to production)
f) P - Production Support / Warranty

Q. When waterfall strategy is used in data warehousing development? (2)
Answer: Waterfall strategy is preferable because it allows the company to take time to understand a market and make appropriate adjustment to its marketing mix in order to satisfy the specific needs of each market. Managers can maximize the use of available resources; they can leverage their experience from the first market and make necessary improvements or changes to enter the next market. The waterfall strategy allows a company to transfer managerial and technological skills from one market to another.

Q. Why RAD is best technique for data warehousing construction. (5)
Answer: Rapid Application Development (RAD) is an iterative model consisting of stages like scope, analyze, design, construct, test, implement, and review. It is much better suited to the development of a data warehouse because of its iterative nature and fast iterations.
User requirements are sometimes difficult to establish because business analysts are too close to the existing infra-structure to easily envision the larger empowerment that data warehousing can offer. Development and delivery of early prototypes will drive future requirements as business users are given direct access to information and the ability to manipulate it.