CS614- Data Warehousing
SPRING 2012

1; what are the two extremes for technical architecture design? 2

Answer:- (Page 95)

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.

2;Different b/w non key or key data access?2

Answer:- (Page 231)

There are two ways to access data: non-keyed access and keyed access.

Non-keyed access uses no index. Each record of the database is accessed sequentially, beginning with the first record, then second, third and so on.

3;"Be a diplomat not a technologist"?2

Answer:- (Page 320)

The biggest problem you will face during a warehouse implementation will be people, not the technology or the development. You're going to have senior management complaining about completion dates and unclear objectives. You're going to have development people protesting that everything takes too long and why can't they do it the old way? You're going to have users with outrageously unrealistic expectations, who are used to systems that require mouse-clicking but not much intellectual investment on their part. And you're going to grow exhausted, separating out Needs from Wants at all levels. Commit from the outset to work very hard at communicating the realities, encouraging investment, and cultivating the development of new skills in your team and your users (and even your bosses).

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

4;Dirty bit?2

Answer:- (Page 438)

- It can be boolean type column
- This column will help us in keeping record of rows with errors, during data profiling

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

Answer:- (Page 12)

- 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.

6;Indexing using I/O bottelneck?3

Answer:- (Page 221)

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.

7; what is mean by the classification process? How we measure the accuracy of classifiers? 3

Answer:- (Page 259)

Classification means that based on the properties of existing data, we have made or groups i.e. we have made classification.

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

8;SQL server meta services advantages(3)

Answer:- (Page 385)

We may maintain meta data information of the databases involved in the packages and we may keep version information of each package. Furthermore package can be stored in a structured file and Microsoft visual basic file.

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

Answer:- (Page 334)

Will adopt a pilot project approach, because:

- A full-blown DWH requires extensive investment.
- Show users the value of DSS.
- Establish blue print for full-blown system.
- Identify problem areas.
- Reveal true data demographics.
- Pilot projects are supposed to work with limited data.

10: Purpose of DTS services(5)

Answer:- (Page)

Many organizations need to centralize data to improve corporate decision-making. However, their data may be stored in a variety of formats and in different locations. Data Transformation Services (DTS) address this vital business need by providing a set of tools that let you extract, transform, and consolidate data from disparate sources into single or multiple destinations supported by DTS connectivity.

DTS allows us to connect through any data source or destination that is supported by OLE DB. This wide range of connectivity that is provided by DTS allows us to extract data from wide range of legacy systems.

Heterogeneous source systems store data with their local formats and conventions. While consolidating data from variety of sources we need to transform names, addresses, dates etc into a standard format. For example consider a student record management system of a university having four campuses. A campus say ‘A’ follows convention to store city codes “LHR” for Lahore. An other campus say ‘B’ stores names of cities “Lahore”, campus ‘C’ stores city names in block letters ‘LAHORE’, and the last campus ‘D’ store city names as ‘lahore’. When the data from all the four campuses is combined as it is and query is run “How many students belong to ‘Lahore’?” We get the answer only from campus B because no other convention for Lahore matches to the one in query.

To combine data from heterogeneous sources with the purpose of some useful analysis requires transformation of data. Transformation brings data in some standard format.

11: Explain analytic dta application specification in kimbill 5 marks

Answer:- (Data Warehouse Toolkit, Page 363)

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.

Before we start designing the initial applications, it’s helpful to establish standards for the applications, such as common pull-down menus and consistent output look and feel. Using the standards, we specify each application template, capturing sufficient inklllformation 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.

12: Business rules are validated using student database in LAB 5 marks

CS614- Data Warehousing
SPRING 2012

Why an organization refuses to visit a person to understand the strategy of Data Warehouse system? 2 Marks

What are the methods used to develop DWH? 2 Marks
Answer:- (Page 283)
Development methodologies
- Waterfall model
- Spiral model
- RAD Model
- Structured Methodology
- Data Driven
- Goal Driven
- User Driven

Define Forward proxy? 2 Marks
Answer:- (Page 369)
It is outside of our control because it belongs to a networking company or an ISP.

Differentiate between data driven DSS and Knowledge driven DSS? 2 Marks
Answer:- Click here for detail
Knowledge-Driven DSS based on Expert System technologies attempts to reason about an input using its knowledge base and logical rules for problem solving.

Model-Driven DSS has a sequence of predefined instructions (a model) for responding to a change in inputs.

Define Classification Process? How to measure the accuracy of Classifier? 3 Marks
Answer:- Rep

What operations are provided by MS DTS? 3 Marks
Answer:- (Page 375)
DTS Overview: Operations
- **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)

Why a pilot project strategy is highly recommended in DWH construction? 3 Marks
Answer:- Rep

Some activities are represented by different format in different data sectors. What issues can be there regarding database structure in DWH 3 Marks
Answer:- (Page 168)
Within the data warehousing field, data cleansing is applied especially when several databases are merged. Records referring to the same entity are represented in different formats in the different data sets or are represented erroneously. Thus, duplicate records will appear in the merged database. The issue is to identify and eliminate these duplicates. The problem is known as the *merge/purge problem*. Instances of this problem appearing in literature are called record linkage, semantic integration, instance identification, or object identity problem.

Fundamental strengths and weakness of k-mean clustering? 5 Marks
Answer:- (Page 281)
Strength
 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*
 Not suitable to discover clusters with *non-convex shapes*

Write a query to find out total numbers of female students registered in BS Telecom. 3 Marks
Answer:- (Page 425)

```
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‘
```

BOOK

<u>ISBN</u>	TITLE	PUBLISHER	ADDRESS

Is this table is in First and second normal form? If yes then what about third normal form? 5 Marks

What issues may occur during data acquisition and cleansing in agriculture case study? 5 Marks

Answer:- (Page 341)

- 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.
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, such as 600 dpi. Subsequently DEO’s (Data Entry Operators) were employed to digitize the scouting sheets by typing. To reduce spelling errors in pesticide names and addresses, drop down menu or combo boxes with standard and correct names were created and used.

CS614- Data Warehousing
SPRING 2012

Qno.1 How time contiguous log entries and HTTP secure socket layer are used for user session identification? What are the limitations of these techniques?

Answer:- (Page 365)

Using Time-contiguous Log Entries

□ A session can be consolidated by collecting time-contiguous log entries from the same host (Internet Protocol, or IP, address).
□ Limitations
□ The method breaks down for visitors from large ISPs
□ Different IP addresses
□ Browsers that are behind some firewalls.

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)
 Offers an opportunity to track a visitor session

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

Limitations

- The downside to using this method is that to track the session, the entire information exchange needs to be in high overhead SSL, and the visitor may be put off by security advisories that can pop up when certain browsers are used.
- Each host server must have its own unique security certificate.

Qno.2 What is the process of gathering information about columns, what is the purpose? Describe briefly
Answer:- (Page 440)
Data profiling, gathering information about columns, fulfils the following two purposes
– Identify the type and extent to which transformation is required
– Gives us a detailed view of data quality

Qno.3 what is mean by click stream? how it can be useful in a web DWH environment
Answer:- (Page 363)
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 *clickstream.* In its most elemental form, the clickstream is every page event recorded by the web server. The clickstream contains a number of new dimensions such as page, session, and referrer-that were previously unknown in conventional DWH environment.

Qno.4 Differentiate between DDS, Data mining and Data Warehouse DWH
Answer:- http://dwbi1.wordpress.com/2012/07/14/what-is-big-data-data-warehouse-data-mining/
A DDS is a database that stores the data warehouse data in a different format than OLTP.
Data mining is the process of exploring data to find the patterns and relationships that describe the data and to predict the unknown or future values of the data. The key value of data mining is the ability to understand why some things happened in the past and the ability to predict what will happen in the future.
A data warehouse is a system that retrieves and consolidates data periodically from the source systems into a dimensional or normalized data store.

Qno.5 Suppose that we collected data of agriculture from the Punjab, We are require underutilized process data to be successful and what your suggestion after decode the data?

Qno.6 What affect on the Data Warehouses if the data across an erroneous record, what should be taken measure by support decision?

Answer:- (Page 159)

Erroneous data leads to unnecessary costs and probably bad reputation when used to support business processes.

Move all such records to exception table.

Qno.7 Differentiate between Range partitioning and Expression Partitioning

Answer:- (Page 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.

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.

Qno.8 What are the two extremes for technical architecture design?

Answer:- Rep

Qno.9 What should be done in the case where golden copy is missing dates

Answer:- (Page 457)

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

When golden copy is unavailable replace with a global value 1/1/50

CS614- Data Warehousing
SPRING 2012

1. Horizontal splitting of tables

Answer:- (Page 55)

Horizontal splitting breaks a relation into multiple record set specifications by placing different rows into different tables based upon common column values. For the multi-campus example being considered; students from Islamabad campus in the Islamabad table, Peshawar students in corresponding table etc. Each file or table created from the splitting has the same record lay out or header.

2. Nested loop join and its variants

Answer:- (Page 239)

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.

3. Hash based join and its working

Answer:- (Page 245)

Hash joins are suitable for the VLDB environment as they are useful for joining large data sets or tables. The choice about which table first gets hashed plays a pivotal role in the overall performance of the join operation, and left to the optimizer.

4. Define DWH, data mining & knowledge discovery in database

Answer:- (Page 254)

Knowledge Discovery

--Overall process of discovering useful knowledge

Data Mining (Knowledge-driven exploration**)**

-- Query formulation problem.
-- Visualize and understand of a large data set.
-- Data growth rate too high to be handled manually.

Data Warehouses (Data-driven exploration):

-- Querying summaries of transactions, etc. ***Decision support***

5. Kimball's life cycle model

Answer:- (Page)

DWH Lifecycle: Key steps

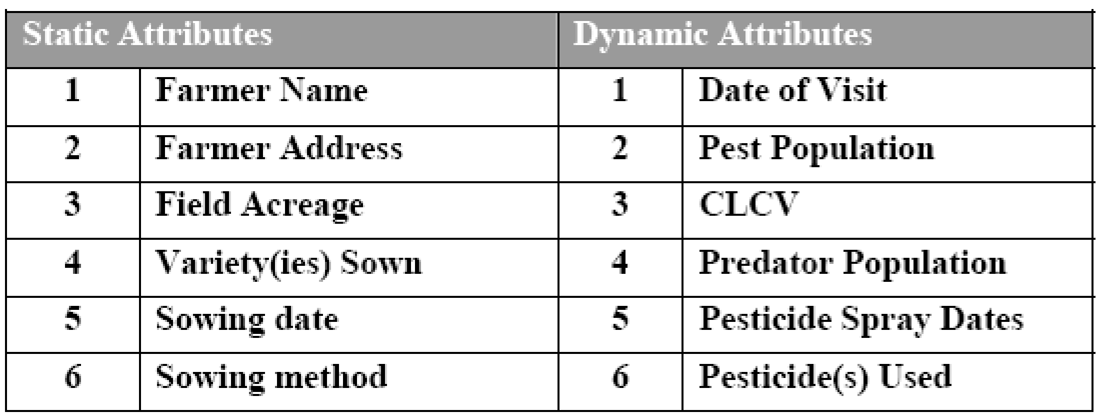
1. Project Planning
2. Business Requirements Definition
3. Parallel Tracks

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

4. Deployment
5. Maintenance

6. Why should companies entertain students to visit their company's place?
Answer:- (Page 328)
• 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.

7. Static & dynamic attributes and examples from agriDWH
Answer:- (Page 342)



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

8. Write SQL to fetch total number of female students registered in BS telcom (5 marks)
Answer:- Rep

9. Is it possible to have erroneous data in DWH? How this will impact business processes? (5 marks)
Answer:- (Page)
Any record having value other than 0 or 1 is erroneous.
Erroneous data leads to unnecessary costs and probably bad reputation when used to support business processes. Consider a company using a list of consumer addresses and buying habits and preferences to advertise a new product by direct mailing. Invalid addresses cause the letters to be returned as undeliverable. People being duplicated in the mailing list account for multiple letters sent to the same person, leading to unnecessary expenses and frustration. Inaccurate information about consumer buying habits and preferences contaminate and falsify the target group, resulting in advertisement of products that do not correspond to consumer's needs. Companies trading such data face the possibility of an additional loss of reputation in case of erroneous data.

10. How is hardware utilization different in DWH? (2 marks)
Answer:- (Page 24)
Although there are peaks and valleys in the operational processing, but ultimately there is relatively static pattern of utilization. There is an essentially different pattern of hardware utilization in the data warehouse environment i.e. a binary pattern of utilization, either the hardware is utilized fully or not at all. Calculating a mean utilization for a DWH is not a meaningful activity. Therefore, trying to mix the two environments is a recipe for disaster. You can optimize the machine for the performance of one type of application, not for both.

11. In case of non uniform distribution, what will be the impact on performance? (5 marks)
Answer:-
http://help.sap.com/saphelp_470/helpdata/en/06/8a983b8e847725e10000000a114084/content.htm
Changes in the B* tree structure can lead to a non-uniform distribution of the data pages. If certain branches of the B* tree have more pages than others, data page distribution is not uniform. This can affect performance, because more accesses are generally required to find data.
Non-uniform distributions are detected by SAP DB when INSERT, UPDATE, or DELETE statements are performed, and the tree is rebalanced when the current operation is carried out. This means that the tree is constantly maintained for optimum operation. During this procedure, page entries are moved to new locations and page pointers are redirected. As a result, data pages are used more efficiently.
Uniform distribution of data prevents individual data regions from overflowing. The only restriction on the size of tables is the storage space available in the database system.

12. Why pilot project methodology is highly recommended in DWH?
Answer:- Rep

CS614- Data Warehousing
Fall 2011

Write the Bit Map index advantages (2)
Answer:- (Page 235)

- Very low storage space.
- Reduction in I/O, just using index.
- Counts & Joins
- Low level bit operations.

Differentiate drill down and pivoting approach (2)
Answer:- Click here for detail
Data drilling (also drilldown) refers to any of various operations and transformations on tabular, relational, and multidimensional data.
A pivot query allows multiple representations of data according to different dimensions.

Explain non standardized attributes of 1/0 and 0/1 stored in attribute library do what problem (2)
Answer:- (Page 405)
Different conventions for representing

“Keep the Competition HOT “ explain in Fradut selection and instat (2)

Answer:- (Page 305)

- Even if single winner, keep at least two in
- Use virtual competition to bargain with the winner

Skew Problem in hash Join how can be solved (2)

Answer:- (Page 247)

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

What is the task performed by the import & export data wizard to load data (3)

Answer:- (Page 411)

□ *Import and Export Data Wizard* provides the easiest method of loading data.
□ The wizard creates package which is a collection of tasks

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

What are the three major applications for accessing information stored on the web (3)

Answer:- (Page 350)

(i) Keyword-based search or topic-directory browsing with search engines such as Google or Yahoo
(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
(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**.**

Differences between one way clustering and two way clustering how they are use together (3)

Answer:- (Page 271)

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.

Briefly explain the types of constraints that we use in DTS (5)

Answer:- (Page 395)

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.

Give single example of DWH of (5)
Financial service/insurance
Tale communication
Transportation Company
Government

Answer:- (Page 323)

Example DWH Target Organizations

• Financial service/insurance.
– Union Bank
– State Bank of Pakistan
• Telecommunications.
– UFone
– PTCL
– PAKNET
• Transportation.
– PIA
• Government.
– NADRA

For example, as per *www.paksearch.com* Muslim Commercial Bank has 900+ branches all over Pakistan. With an average of 500 customers per branch, the total number of customers is in the order of half a million. It would not be surprising if the weekly ATM transactions all over Pakistan run into millions. Such banks are potential candidates for a data warehouse. Same is true for telecommunication companies. As per recent government figures, there are 10+ million mobile phone users in Pakistan, and as per *www.fdi.com* the number of mobile phone users of Mobilink is 3.7 million. Again, it would not be surprising to have literally millions of mobile phone calls made/received per day. So these businesses fall under the category which you should be looking at to select and study as part of your semester project.

NADRA (National Database and Registration Authority) probably has the largest data warehouse in Pakistan and is the repository of the 1998 census which covered the entire population of Pakistan, which at time stood at 130 millions. As part of the census (source: *www.nadra.gov.pk*) 64 million NDFs or National Data Forms were collected and scanned which are presently stored in a 4.2 Tera byte DWH in NADRA.

CS614- Data Warehousing
Fall 2011

Q.1-Difference between one-to-one scalar transformation and one-to-many transformation on data warehouse? (2)
Answer:- (Page 144)
Simple scalar transformation is a one-to-one mapping from one set of values to another set of values using straightforward rules.
A one-to-many transformation is more complex than scalar transformation. As a data element form the source system results in several columns in the DW.

Q.2-What will be effect if we program Package by using DTS object Model?(2)
Answer:- (Page 144)
Package can also be programmed by using DTS object model instead of using graphical tools but DTS programming is rather complicated.

Q.3-Name of the authority who is recording Pest scouting data? Also given year this authority is working?(2)
Answer:- (Page 333)
DPWQCP… 1984

Q.4-What is meant by the statement ”keep the competition hot” ?(2)
Answer:- Rep

Q.5-What is the basic concept of Inverted Index?(2)
Answer:- (Page 232)
An inverted index is an optimized structure that is built primarily for retrieval, with update being only a secondary consideration.

Q.6-Give the possible reason to use enrichment during data transformation?(3)
Answer:- (Page 136)
This task is the rearrangement and simplification of individual fields to make them more useful for the data warehouse environment. You may use one or more fields from the same input record to create a better view of the data for the data warehouse. This principle is extended when one or more fields originate from multiple records, resulting in a single field for the data warehouse.

Q.7-Document Architecture Requirement of Kimball’s model with example? (3)
Answer:- (Page 302)
Document Architecture Requirements
Once the business requirements definition process is leveraged and supplemental IT interviews conducted, the findings need to be documented. A simplistic MATRIX can be used for this purpose. The rows of the matrix list each business requirement that has an impact on the architecture, while matrix columns contain the list of architectural implications.
As an example supposes that a business is spread globally and there is a need to deliver global sales performance data on a nightly basis. The technical implications might include 24/7 worldwide availability, data mirroring for loads, robust metadata for support global access, adequate network bandwidth, and sufficient staging horsepower to handle the complex integration of operational data and so on.

Q.8-Tasks perform through import / export wizard to load data?(3)
Answer:- Rep

Q.9-What is meant by click stream? How can be useful in Web DWH environment?(3)
Answer:- Rep

Q.10- List and explain fundamental advantages of Bit map index?(3)
Answer:- Rep

Q.11-Explain the seven steps for the extracting data using SQL and DTS wizard? (5)
Answer:- (Page 413)
1. Launch the Wizard
2. Choose a Data Source
3. Choose a Database
 Specification of file format incase of Text files
4. Specify the Destination
5. Choose Destination Database
 Selection of existing database or creation of a new database
6. Select a table
 Selection of existing table or creation of a new table
7. Finalizing and Scheduling the package

Q.12-What is difference between the data matrix and similarity/ dissimilarity matrix in term of rows and columns? Which one is the symmetric? (5)
Answer:- (Page 270)
Data matrix means the table or database used as the input to the DM algorithm. What will be the dimensions or size of that table normally? The size of records (rows) is much greater than the number of columns. The attributes may be 10, 15 or 25 but the number of rows far exceeds the number of columns e.g. a customer table may have 20-25 attributes but the total records may be in millions. As I said previously that the mobile users in Pakistan are about 10 million. If a company even has 1/3 of the customers then 3.3 *lakh* customer records in the customer table. Thus greater number of rows than columns and there will be indices *i* and *j* in the table and you can pick the particular contents of a cell by considering the intersection of the two indices.

Similarity or dissimilarity matrix is the measure the similarity or dissimilarity obtained by pair wise comparison of rows. First of all you measure the similarity of the row1 in data matrix with itself that will be 1. So 1 is placed at index 1, 1 of the similarity matrix. Then you compare row 1 with row 2 and the measure or similarity value goes at index 1, 2 of the similarity matrix and son. In this way the similarity matrix is filled. It should be noted that the similarity between row1 and row2 will be same as between row 2 and 1. Obviously, the similarity matrix will then be a square matrix, symmetric and all values along the diagonal will be same (here 1). So if your data matrix has *n* rows and *m* columns then your similarity matrix will have *n* rows and *n* columns. What will be the time complexity of computing similarity/dissimilarity matrix? It will be O (n2) (m), where m accounts for the vector or header size of the data. Now how to measure or quantify the similarity or dissimilarity? Different techniques available like Pearson correlation and Euclidean distance etc. but in this lecture we have used Pearson correlation which you might have studied in your statistics course.

Give the simple example of company name that using the DWH
Financial Services/insurance

Telecommunication

Transportations

Government

Answer:- Rep

CS614- Data Warehousing
Fall 2011

Q1: What sorts of objectives metric are use by companies what are possible issues in formulation these metric? 2 marks

Answer:- (Page 186)

When performing objective assessments, companies follow a set of principles to develop metrics specific to their needs, there is hard to have “one size fits all” approach. Three pervasive functional forms are (i) simple ratio, (ii) min or max operation, and (iii) weighted average.

Refinements of these functional forms, such as addition of sensitivity parameters, can be easily incorporated. Often, the most difficult task is precisely defining a dimension, or the aspect of a dimension that relates to the company's specific application. Formulating the metric is straightforward once this task is complete.

Q2: which script language is used to perform complex transformation in DTS package? 2 marks

Answer:- (Page 373)

Complex transformations are achieved through VB Script or Java Script that is loaded in DTS package.

Q3: Cleansing can be breaking down in Whom many steps, write their names? 2 marks
Answer:- (Page 168)
break down the cleansing into six steps:
elementizing, standardizing, verifying, matching, house holding, and documenting.

Q4: What do you mean by “keep competition hot in context of production selection and transformation while designing a data warehouse “. 2 marks
Answer:- (Page 305)
□ Make private not public commitment.
□ Don’t let the vendor you are completely sold.
□ During *trial period*, put to real use.
□ Near the end of trial, negotiate.

Q5: Who murge column are selected in case of sort merge? 2 marks
Answer:- (Page 243)
There may be multiple equalities in the WHERE clause, in such a case, the merge columns are taken from only some of the given equality clauses.

Q6: purposes dta data profiling 3 marks
Answer:- (Page 264)
Data profiling, gathering information about columns, fulfils the following two purposes
– Identify the type and extent to which transformation is required
– Gives us a detailed view of data quality

Q7: What issues may Accor during data acquisition and cleansing in agriculture case Study? 3marks
Answer:- Rep

Q8: Meant of classification process, How measure accuracy of classification? 3marks
Answer:- Rep

Q9: Data parallelism explain with example 3 marks
parallel execution is to completely
parallelize those parts of a computation that are not constrained by data dependencies.
The smaller the portion of the program that must be executed sequentially (s), the greater
the scalability of the computation.
Q10: Under what condition an operation can be execute in parallel? 3 marks
Answer:- (Page 201)
Parallel execution dramatically reduces response time for data-intensive operations on large databases typically associated with Decision Support Systems (DSS) and data warehouses. You can also implement parallel execution on certain types of online transaction processing (OLTP) and hybrid systems.

Q.11: Explain anayletic dta application specification in kimbill 5 marks

Answer:- Rep

Q12: 2 real life examples of clustering. 5 marks

Answer:- (Page 264)

- 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.

CS614- Data Warehousing
Fall 2011

Q1: IN ROUND ROBIN THE DISTRIBUTION IS Pre DEFINED. DO YOU AGREE OR NOT SUPPORT YOU'RE ANSWER WITH REASON? 2

Answer:- (Page 66)

In round robin distribution is Not pre-defined. Round-robin spreads data evenly across the partitions, but does not facilitate partition elimination. Round-robin is typically used only for temporary tables where partition elimination is not important and co-location of the table with other tables is not expected to yield performance benefits.

Q2: what are major operations of data mining? 2

Answer:- Click here for detail

Data Mining methods may be classified by the function they perform or according to the class of application they can be used in

- Predictive modeling
- Segmentation (Clustering)
- Dependency Modeling
- Summarization
- Change and deviation detection

Q3: which scripting language is used to perform complex transformation in DST package? 2
Answer:- Rep

Q4: a person wanted to visit and understand the data warehouse implementation strategies adopted in that organization has refused to allow. What may be the carrier of this refusal?

Q4: how the applications of parallelism differ for OLTP and DSS environment? 2
Answer:- (Page 205)
There is a big difference.
In DSS Parallelization of a SINGLE query
In OLTP Parallelization of MULTIPLE queries Or Batch updates in parallel

Q5: keeping view the uniform distribution in hash based partition .if the partitions are not uniformly distributed across the process? 3
Answer:- (Page 218)
There can be two types of skews i.e. non uniform distribution when the data is distributed across the processors. One type of skew is dependent in the properties of the data, consider the example of data about cancellation of reservations. It is obvious that most cancellations in the history of airline travel occurred during the last quarter of 2001. Therefore, whenever the data is distributed based on date for year 2001 it will be always skewed. This can also be looked at from the perspective of partition skew, as date is typically seen to result in non-uniform istribution of data.

Q6: what is the task performed through import export data wizard to load data? 3
Answer:- Rep

Q7: what is mean by click stream? How it can be useful in a web DWH environment? 3
Answer:- Rep

Q8: what is mean by the classification process? How we measure the accuracy of classifiers? 3
Answer:- Rep

Q9: discuss need for indexing with reference to i/o speed? 3
Answer:- (Page 220)
Consider the "Find" operation in Windows; a user search is initiated and a search starts through each file on the hard disk. When a directory is encountered, the search continues through each directory. With only a few thousand files on a typical laptop, a typical "find" operation takes a minute or longer.

Q10: why a pilot project strategy is highly recommended in DWH construction? 5
Answer:- Rep

Issues of cluster index (2)

Answer:- (Page 237)

Cluster Index: Issues

 Works well when a single index can be used for the majority of table accesses.
 Selectivity requirements for making use of a cluster index are much less stringent than for a non-clustered index.
 High maintenance cost to keep sorted order or frequent reorganizations to recover clustering factor.

Fixed strategy of standardizing column(2)

Answer:- (Page 480)

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

Be a diplomat and not technologist(2)

Answer:- Rep

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

Answer:- (Page 223)

Dense Index: Adv. & Dis. Adv.

For each record store the key and a pointer to the record in the sequential file. Why?
It uses less space, hence less time to search. Time (I/Os) logarithmic in number of blocks used by the index.
Can also be used as secondary index, i.e. with another order of records.
Dense Index*:* Every key in the data file is represented in the index file
ProA 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: 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.

Weaknesses of K-mean clustering(3)

Answer:- Rep

SQL server meta services advantages(3)
Answer:- Rep

Issued faced in data cleansing of AgriDWH(3)
Answer:- Rep

Problems which may face in construction of AgriDWH(3)
Answer:- (Page 339)
At the pilot level of Agri-DWH the most important aspects were "what mean what". The issues of who can access what and how, was not there, as there were only one or two users. The business metadata issues in the context of business rule will be briefly discussed in this section.

Why pilot strategy is recommended for construction of DWH(5)
Answer:- Rep

Strength and weaknesses of k-mean clustering(5)
Answer:- Rep

CS614- Data Warehousing
Fall 2011

Q #1: Why "Justification" is required in project planning?
Answer:- (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.

Q #2: Define forward proxy?
Answer:- Rep

Q #3: Is there any fixed strategy to standardize a column?
Answer:- Rep

Q #4: Major operation of data mining?
Answer:- Rep

Q #5: In Round-Robin the distribution is pre-defined, you agree or not? Support your answer.
Answer:- Rep

Q #6: Differentiate in dense index and spars index? What are advantages and disadvantages of these?
Answer:- Rep

Q #7: what is classification process? How we measure the accuracy of classifiers?
Answer:- Rep

Q #8: For what reason trivial quires give wrong result in Agri-DWH? Explain with Example.

Q #9: Purposes of data profiling?
Answer:- Rep

Q #10: TQM benefits? Why organizations prefer this technique on other techniques?

Q #11: How mistakes should avoid in data warehouse process?

Q #12: How time contiguous log entries and HTTP secure socket layer are used for user session identification? Limitation of this.
Answer:- (Page 365)
Using Time-contiguous Log Entries
□ A session can be consolidated by collecting time-contiguous log entries from the same host (Internet Protocol, or IP, address).
□ Limitations
□ The method breaks down for visitors from large ISPs
□ Different IP addresses
□ Browsers that are behind some firewalls.
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)
□ Offers an opportunity to track a visitor session
□ **Limitations**
□ To track the session, the entire information exchange needs to be in high overhead SSL
□ Each host server must have its own unique security certificate.
□ Visitors are put-off by pop-up certificate boxes.
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

• The downside to using this method is that to track the session, the entire information exchange needs to be in high overhead SSL, and the visitor may be put off by security advisories that can pop up when certain browsers are used.

• Each host server must have its own unique security certificate.

CS614- Data Warehousing
Fall 2011

i) K-mean weakness

Answer:- Rep

ii) Classification process and its accuracy?

Answer:- Rep

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

Answer:- Rep

iv) Name of authority to pest

v) 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.

Answer:- Rep

vi) Inverted index

Answer:- Rep

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

Answer:- Rep

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

Answer:- (Page 13)

it took milliseconds to locate a record on a DASD i.e. orders of magnitude better performance than the magnetic tape. With DASD came a new type of system software known as a DBMS (Data Base Management System). The purpose of the DBMS was to facilitate the programmer to store and access data on DASD. In addition, the DBMS took care of such tasks as storing data on DASD, indexing data, accessing it etc.

ix) Transient cookie & persistent cookie
Answer:- (Page 367)
Using 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

Using Persistent Cookies
□ Establish a persistent cookie in the visitor's PC
□ **Limitations**
□ No absolute guarantee that even a persistent cookie will survive.
□ Certain groups of Web sites can agree to store a common ID tag

xi) clickstream
Answer:- Rep

xii) Data profiling is a process of gathering information about columns, what are the purpose that it must fulfill? Describe briefly
Answer:- Rep

CS614- Data Warehousing
Fall 2011

1) How Business validation rule implemented in DTS? 5

2) Clickstream? How it is useful in a web dwh environment. 3
Answer:- Rep

3) Data profiling is a process of gathering information about columns, what are the purpose that it must fulfill? Describe briefly 5
Answer:- Rep

4) Why a Pilot project strategy Is highly recommended in dwh construction. 5
Answer:- Rep

5) Need for indexing with reference to i/o speed 3
Answer:- Rep

Question No: 6 (Marks: 5)
Explain Analytic Applications Development Phase of Analytic Applications Track of Kimball's Model?
Answer:- (Data Warehouse Toolkit, Page 362)
When we move into the development phase for the analytic applications, we again need to focus on standards. Standards for naming conventions, calculations, libraries, and coding should be established to minimize future rework. The application development activity can begin once the database design is complete, the data access tools and metadata are installed, and a subset of historical data has been loaded. The application template specifications should be revisited to account for the inevitable changes to the data model since the specifications were completed.

Each tool on the market has product-specific tricks that can cause it to jump through hoops backwards. Rather than trying to learn the techniques via trial and error, you should invest in appropriate tool-specific education or supplemental resources for the development team.

While the applications are being developed, several ancillary benefits result. Application developers, armed with a robust data access tool, quickly will find needling problems in the data haystack despite the quality assurance performed by the staging application. This is one reason why we prefer to get started on the application development activity prior to the supposed completion of staging. Of course, we need to allow time in the schedule to address any flaws identified by the analytic applications. The developers also will be the first to realistically test query response times. Now is the time to begin reviewing our performance-tuning strategies.

The application development quality-assurance activities cannot be completed until the data is stabilized. We need to make sure that there is adequate time in the schedule beyond the final data staging cutoff to allow for an orderly wrap-up of the application development tasks.

8) Which script language is used to perform complex transformation in dts package 2
Answer:- Rep

9) What types of operations are performed by MS DTS. 3
Answer:- Rep

10) Name of the pest scouting org and the year of its starting. 2 marks
Answer:- Rep

CS614- Data Warehousing
Spring 2011

1- Define parallelism explain it with example, under what condition an operation can be executed in parallel.

Answer:- (Page 204)

Parallel execution is to completely parallelize those parts of a computation that are not constrained by data dependencies. The smaller the portion of the program that must be executed sequentially (s), the greater the scalability of the computation.

Parallel execution is to completely parallelize those parts of a computation that are not constrained by data dependencies. The smaller the portion of the program that must be executed sequentially (s), the greater the scalability of the computation.

2- What is value validation process?

Answer:- (Page 159)

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

3- Explain analytic application specification phase of analytic application track of Kimball's Model?

Answer:- Rep

4- How gender guide is used if large number of records, gender is missing?

Answer:- (Page 458)

If for very large number of records gender is missing, it would become impossible for us to manually check each and every individual's name and identify the gender. In such cases we can formulate a mechanism to correct gender. We can either use a standard gender guide or create a new table Gender_guide. 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.

5- How page dimensions is useful to check the static and dynamic nature of the web pages.

Answer:- (Page 362)

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. We will assume even in the case of the dynamic page that there is a well defined function that characterizes the page, and we will use this to describe the page. We will not create a page row for every instance of a dynamic page, because that would yield a dimension with an astronomical number of rows, yet the rows would not differ in interesting ways. What we want is a row in this dimension for each interesting, distinguishable type of page. Static pages probably get their own row, but dynamic pages would be grouped by similar function and type.

6- Give name of activities to be performed in planning and design phase as discussed in agri-DWH case study

Answer:- (Page 335)

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

7- What are the three methods of creating a DTS package

Answer:- (Page 380)

Package can be created by one of the following three methods:

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