CS614 VU Today's Final Term Paper for Fall 2013

Total Questions = 52 of Total 80 Marks

Total MCQ = 40 Each of 1 marks
Total Short Questions = 4 Each of 2 marks
Total Short Questions = 4 Each of 3 marks
Total Long Questions = 4 Each of 5 marks

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.

Q2: What are different issues during data acquisition and cleansing in agricultural data warehouse? (5) page 340

Solution:

Step-6: Why the issues?

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.

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 'Muhammed'. Then in our Gender_guide table we will have just one entry 'Muhammed' and we will manually set the gender as 'Male' against 'Muhammed'.
Now to fill missing genders in exception table we will just do a 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.

Q8: Following statement is correct or incorrect? If incorrect then justify your answer? (3)
Answer:

"One way clustering gives local view and two way clustering gives global view".
The above statement is incorrect: **page 271**
Bi-clustering (Two way clustering) gives a local view of your data set while one-way clustering gives a global view.

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.

Q11: What is unsupervised learning in Data Mining? (2) page 271
Answer:

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.

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.

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.

What is mean by click stream? How it can be useful in a web DWH environment

Answer: Ch#40 Page no : 363

Clickstream is every page event recorded by each of the company's Web servers**.** The clickstream is not just another data source that is extracted, cleaned, and dumped into the data warehouse. It is an evolving collection of data sources having more than a dozen Web server log file formats for capturing clickstream data. These formats have optional data components that, if used, can be very helpful in identifying visitors, sessions, and the true meaning of behavior.

Define Classification Process? How to measure the accuracy of Classifier?

Answer: Ch#31 Page no : 276

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.

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**

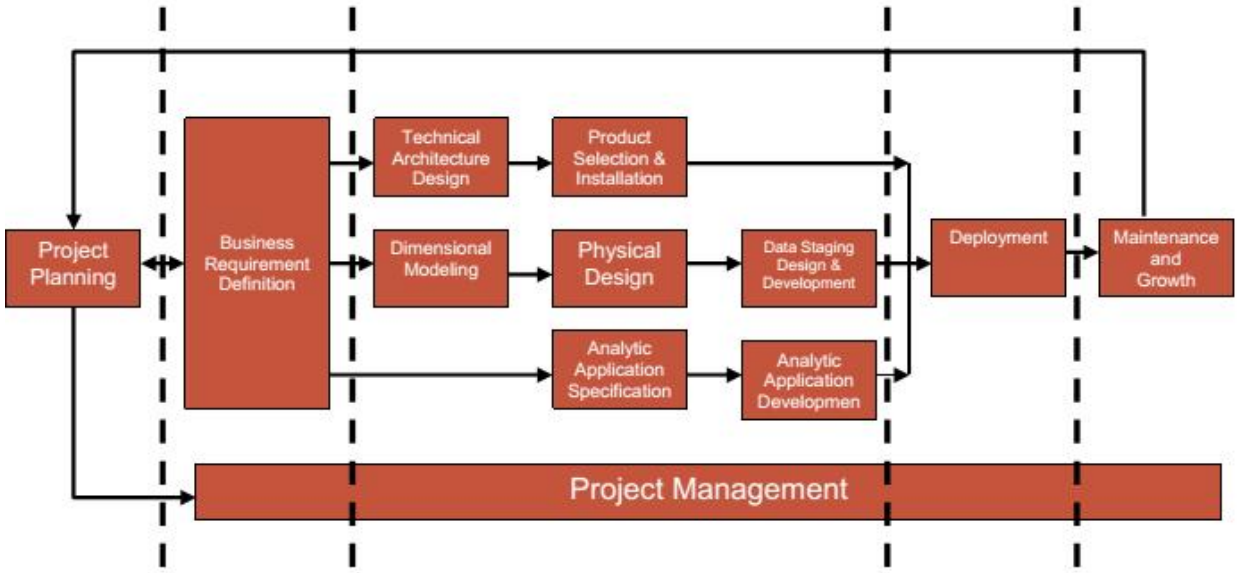


Figure -33.1: Business Dimensional Lifecycle (Kimball's Approach)

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

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

Answer: CH#38 page 340

Step-6: Why the issues?

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.

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 whendealing 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.

Ralph Kimball approach is business dimensional lifecycle.

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:4 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

Issues of Click stream. Under which category it is lie?

Answer: ch:40 page 363

Clickstream 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. Clickstream 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 Atre topic : 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.

List any five functions 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)

Two conditions are given and asked about which type of transformation it is? see Page#153

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. Its 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

Forward and backward 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.

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.

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

Objective mooz ki file sa nai the. har handout of lec sa aik 2 thay espacillay lect 27,28,29 and 33

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_mail etc 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.

Shared architecture?

Answer:

All Processors have equal access to the data stored on disk.

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.

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

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:

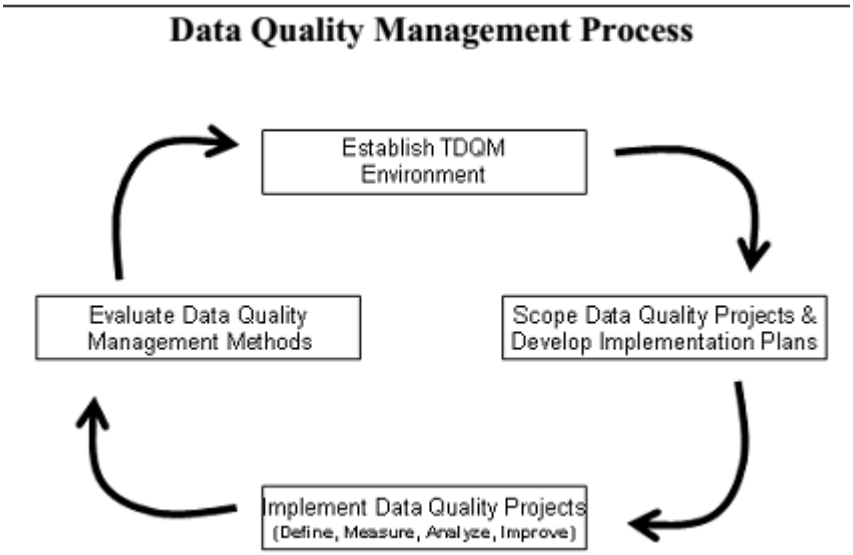
Amdahl's formula chose karna tha

Answer:

$S <= 1/ f+(1-f)/N$

Quality management ka tha

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:

Limitations

 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

3. Ways of accessing information on web? 2 marks

Answer:

(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.

(iii) Random surfing that follows Web linkage pointers.

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. Draw bitmap index of given table? 5 marks

Answer:

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

Answer:

A&B = No of blocks Access by A + No of qualified blocks by A* No of Blocs Accessed by B

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))

My Paper:

Data ware house development methodologies:

Answer:

- Waterfall model
- Spiral model
- RAD Model
- Structured Methodology
- Data Driven
- Goal Driven
- User Driven

Data Quality improvement categories:

Answer:

The four categories of Data Quality Improvement

- Process
- System
- Policy & Procedure
- Data Design

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 to locate the rows.

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

Why to save package in SQL Server 2000 Meta Data Services ?

Answer: Page no : 387 Ch# lab lecture 1

To maintain version information packages are saved in "SQL Server 2000 Meta Data Services".

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

Define Automatic Data Cleansing Techniques.

Answer: Page no : 164 Ch#19

1) Statistical
2) Pattern Based
3) Clustering
4) Association Rules

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