

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 bottleneck?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 data 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 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.

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

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, $k, t \ll n$.*
- ☐ 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

ISBN	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.

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

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 layout 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 telecom (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