CS614 Short Notes-mid-term/ Short Questions Answers

VUAnswer.com

Write a quarry to extract total number of female students registered in BS Telecom. 5 marks

Total Number of Female students in BS Telecom

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

Describe the lessons learn at during agri-data ware house case study?

- Extract Transform Load (ETL) of agricultural extension data is a big issue. There are no digitized operational databases so one has to resort to data available in typed (or hand written) pest scouting sheets. Data entry of these sheets is very expensive, slow and prone to errors.
- Particular to the pest scouting data, each farmer is repeatedly visited by agriculture extension people. This results in repetition of information, about land, sowing date, variety etc (Table-2). Hence, farmer and land individualization are critical, so that repetition may not impair aggregate queries. Such an individualization task is hard to implement for multiple reasons.
- There is a skewness in the scouting data. Public extension personnel (scouts) are more likely to visit educated or progressive farmers, as it makes their job of data collection easy. Furthermore, large land owners and influential farmers are also more frequently visited by the scouts. Thus the data does not give a true statistical picture of the farmer demographics.
- Unlike traditional data warehouse where the end users are decision makers, here the decision-making goes all the way "down" to the extension level. This presents a challenge to the analytical operations‟ designer, as the findings must be fairly simple to understand and communicate.

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

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

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

Data profiling is a process which involves gathering of information about column through execution of certain queries with intention to identify erroneous records. In this process we identify the following:

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

We run different SQL queries to get the answers of above questions. During this process we can identify the erroneous records. Whenever we will come across an erroneous record, we will just copy it in error or exception table and set the dirty bit of record in the actual student table. Then we will correct the exception table. After this profiling process we will transform the records and load them into a new table Student_Info

Define additive and non additive facts

Additive facts are those facts which give the correct result by an addition operation. Examples of such facts could be number of items sold, sales amount etc. Non-additive facts can also be added, but the addition gives incorrect results. Some examples of non-additive facts are average, discount, ratios etc.

What are three fundamental reasons for warehousing web data?

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

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

What are the two basic data warehousing implementation strategies and their suitability conditions?

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

Bitmap Indexes: Concept

- Index on a particular column

- Index consists of a number of bit vectors or bitmaps
- Each value in the indexed column has a corresponding bit vector (bitmaps)
- The length of the bit vector is the number of records in the base table
- The ith bit is set to 1 if the ith row of the base table has the value for the indexed column

List and explain fundamental advantages of bit map indexing
Bitmap Index: Adv.

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

An obvious advantage of this technique is the potential for dramatic reductions in storage overhead. Consider a table with a million rows and four distinct values with column header of 4 bytes resulting in 4 MB. A bitmap indicating which of these rows are for these values requires about 500KB.

More importantly, the reduction in the size of index "entries" means that the index can sometimes be processed with no I/O and, more often, with substantially less I/O than would otherwise be required. In addition, many index-only queries (queries whose responses are derivable through index scans without searching the database) can benefit considerably.

Database retrievals using a bitmap index can be more flexible and powerful than a B-tree in that a bitmap can quickly obtain a count by inspecting only the index, without retrieving the actual data. Bitmap indexing can also use multiple columns in combination for a given retrieval.

Finally, you can use low-level Boolean logic operations at the bit level to perform predicate evaluation at increased machine speeds. Of course, the combination of these factors can result in better query performance.

List and explain fundamental disadvantages of bit map indexing
Bitmap Index: Dis. Adv.

- Locking of many rows
- Low cardinality
- Keyword parsing
- Difficult to maintain - need reorganization when relation sizes change (new bitmaps)

Row locking: A potential drawback of bitmaps involves locking. Because a page in a bitmap contains references to so many rows, changes to a single row inhibit concurrent access for all other referenced rows in the index on that page.

Low cardinality: Bitmap indexes create tables that contain a cell for each row times each possible value (the product of the number of rows times the number of unique values).
Therefore, a bitmap is practical only for low- cardinality columns that divide the data into a small number of categories, such as "M/F", "T/F", or "Y/N" values.

Keyword parsing: Bitmap indexes can parse multiple values in a column into separate keywords. For example, the title “Marry had a little lamb" could be retrieved by entering the word “Marry" or “lamb" or a combination. Although this keyword parsing and lookup capability is extremely useful, textual fields tend to contain high-cardinality data (a large number of values) and therefore are not a good choice for bitmap indexes.

What are major operations of data mining?

- Classification
- Estimation
- Prediction
- Clustering
- Description

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

DTS package is exactly like a computer program. Like a computer program DTS package is also prepared to achieve some goal. Computer program contains set of instructions whereas DTS package contains set of tasks. Tasks are logically related to each other. When a computer program is run, some instructions are executed in sequence and some in parallel. Likewise when a DTS package is run some tasks are performed in sequence and some in parallel. The intended goal of a computer program is achieved when all instructions are successfully executed. Similarly the intended goal of a package is achieved when all tasks are successfully accomplished

Package can also be programmed by using DTS object model instead of using graphical tools but DTS programming is rather complicated.

Write down the steps of handling skew in range partitioning?

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

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

Q12what type of anomalies exists if a table is in 2NF not in 3NF? [2]

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

What are three methods for creating a DTS package?

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

Write two extremes of Tech. Arch Design?

Attacking the problem from two extremes, neither is correct.

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

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

Q1: Explain analytic data application specification in Kimball 5 marks

- Starter set of 10-15 applications.
- Prioritize and narrow to critical capabilities.
- Single template use to get 15 applications
- Set standards: Menu, O/P, look feel.
- From standard: Template, layout, I/P variables, calculations.
- Common understanding between business & IT users

Analytic applications development

- Standards: naming, coding, libraries etc.
- Coding begins AFTER DB design complete, data access tools installed, subset of historical data loaded.
- Tools: Product specific high performance tricks, invest in tool-specific education.
- Benefits: Quality problems will be found with tool usage => staging.
- Actual performance and time gauged.

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

Data profiling is a process which involves gathering of information about column through execution of certain queries with intention to identify erroneous records. In this process we identify Validation of business rules

Q3: 2 real life examples of clustering 5 marks

Examples of Clustering Applications

Marketing: Discovering distinct groups in customer databases, such as customers who make lot of long-distance calls and don‟t have a job. Who are they? Students. Marketers use this knowledge to develop targeted marketing programs.

Insurance: Identifying groups of crop insurance policy holders with a high average claim rate. Farmers crash crops, when it is "profitable".

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

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

Q5: What issues may occur during data acquisition and cleansing in agriculture case study? 3marks

- The pest scouting sheets are larger than A4 size (8.5" x 11"), hence the right end was cropped when scanned on a flat-bed A4 size scanner.
- The right part of the scouting sheet is also the most troublesome, because of pesticide names for a single record typed on multiple lines i.e. for multiple farmers.
- As a first step, OCR (Optical Character Reader) based image to text transformation of the pest scouting sheets was attempted. But it did not work even for relatively clean sheets with very high scanning resolutions.

- Subsequently DEO‟s (Data Entry Operators) were employed to digitize the scouting sheets by typing.

Q6: Meant of classification process, how measure accuracy of classification? 3marks

First of the available data set is divided into two parts, one is called test set and the other is called the training set. We pick the training set and a model is constructed based on known facts, historical data and class properties as we already know the number of classes. After building the classification model, every record of the test set is posed to the classification model which decides the class of the input record. It should be noted that you know the class for each record in test set and this fact is used to measure the accuracy or confidence level of the classification model. You can find accuracy by

Accuracy or confidence level = matches/ total number of matches

In simple words, accuracy is obtained by dividing number of correct assignments by total number of assignments by the classification model

Q7: Data parallelism explain with example 3 marks

- Parallel execution of a single data manipulation task across multiple partitions of data.
- Partitions static or dynamic
- Tasks executed almost-independently across partitions.
- Query coordinator" must coordinate between the independently executing processes.

So data parallelism is I think the simplest form of parallelization. The idea is that we have parallel execution of single data operation across multiple partitions of data. So the idea here is that these partitions of data may be defined statically or dynamically fine, but we are requiring the same operator across these multiple partitions concurrently. And this idea actually of data parallelism has existed for a very long time. So the idea is that you are getting parallelization because we are getting semi-independent execution, data manipulation across the partitions. And as long as we keep the coordination required, we can get very good speedups. Well again this query coordinator, the thing that keeps the query distributed but still working and then collects its results. Now that query coordinator can potentially be a bottleneck, because if it does too much work, that is serial execution. So the query coordination has to be very small amount of work. Otherwise the overhead gets higher and the serialization of the workload gets higher.

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

Under the things which can be divided into two such as with reference to size and with reference to divide and conquer an operation can be execute in parallel.

Q9: What sorts of objectives metric are use by companies what are possible issues in formulation these metric? 2 marks

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

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

Q11: Cleansing can be break down in Who many steps, write their names? 2 marks
One can break down the cleansing into six steps: elementizing, standardizing, verifying, matching, house holding, and documenting.
Q12: What does u mean by "keep competition hot in context of production selection and transformation while designing a data warehouse ". 2 marks

Q13: Who merge column is selected in case of sort merge? 2 marks
The Sort-Merge join requires that both tables to be joined are sorted on those columns that are identified by the equality in the WHERE clause of the join predicate. Subsequently the tables are merged based on the join columns.

3) Different b/w non key or key data access? 2
Non-keyed access uses no index. Each record of the database is accessed sequentially, beginning with the first record, then second, third and so on. This access is good when you wish to access a large portion of the database (greater than 85%). Keyed access provides direct addressing of records. A unique number or character(s) is used to locate and access records. In this case, when specified records are required (say, record 120, 130, 200 and 500), indexing is much more efficient than reading all the records in between.

4) "Be a diplomat not a technologist"? 2
The biggest problem you will face during a warehouse implementation will be people, not the technology or the development. You"re going to have senior management complaining about completion dates and unclear objectives. You"re going to have development people protesting that everything takes too long and why can"t they do it the old way? You"re going to have users with outrageously unrealistic expectations, who are used to systems that require mouse-clicking but not much intellectual investment on their part. And you"re going to grow exhausted, separating out Needs from Wants at all levels. Commit from the outset to work very hard at communicating the realities, encouraging investment, and cultivating the development of new skills in your team and your users (and even your bosses).

5) Dirty bit?2
- Add a new column to each student table
- This new column is named as "Dirty bit"
- It can be Boolean type column
- This column will help us in keeping record of rows with errors, during data profiling

6) What are the problem face industry when the growth in usage of master table file increase?3
The spreading of master files and massive redundancy of data presented some very serious problems, such as:
- Data coherency i.e. the need to synchronize data upon update.
- Program maintenance complexity.
- Program development complexity.

- Requirement of additional hardware to support many tapes.
-

7) Indexing using I/0 bottelneck?3

Need For Indexing: I/O Bottleneck

Throwing more hardware at the problem doesn't really help, either. Expensive and multiprocessing servers can certainly accelerate the CPU-intensive parts of the process, but the bottom line of database access is disk access, so the process is I/O bound and I/O doesn't scale as fast as CPU power. You can get around this by putting the entire database into main memory, but the cost of RAM for a multi-gigabyte database is likely to be higher than the server itself! Therefore we index. Although DBAs can overcome any given set of query problems by tuning, creating indexes, summary tables, and multiple data marts, or forbidding certain kinds of queries, they must know in advance what queries users want to make and would be useful, which requires domain-specific knowledge they often don't have. While 80% of database queries are repetitive and can be optimized, 80% of the ROI from database information comes from the 20% of queries that are not repetitive. The result is a loss of business or competitive advantage because of the inability to access the data in corporate databases in a timely fashion.

9) W8 is Click stream? Limitations?3

Click stream

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

Click stream data has many issues.

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

10) Import/export wizard tasks?3

- First of all load data

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

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

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

Problem with simple approach

- Number of required queries increases exponentially with the increase in number of dimensions.
- It‟s wasteful to compute all queries.
- In the example, the first query can do most of the work of the other two queries
- If we could save that result and aggregate over Month_Id and Product_Id, we could compute the other queries more efficiently

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

Data Mining Vs. Statistics

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

13) Persistent cookies limitations? 5

Using Persistent Cookies

Establish a persistent cookie in the visitor's PC. The Web site may establish a persistent cookie in the visitor's PC that is not deleted by the browser when the session ends.

Limitations

- No absolute guarantee that even a persistent cookie will survive.
- Certain groups of Web sites can agree to store a common ID tag

Misconception about data quality

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

Issues of data cleansing

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.

Classification and estimation

- Classification consists of examining the properties of a newly presented observation and assigning it to a predefined class.
- Assigning customers to predefined customer segments (good vs. bad)
- Assigning keywords to articles

- Classifying credit applicants as low, medium, or high risk
- Classifying instructor rating as excellent, very good, good, fair, or poor

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

Star schema
Star Schema: A star schema is generally considered to be the most efficient design for two reasons. First, a design with de-normalized tables encounters fewer join operations. Second, most optimizers are smart enough to recognize a star schema and generate access plans that use efficient "star join" operations. It has been established that a "standard template" data warehouse query directly maps to a star schema.

Why a pilot project strategy is highly recommended in DWH construction? 5
A pilot project strategy is highly recommended in data warehouse construction, as a full blown data warehouse construction requires significant capital investment, effort and resources. Therefore, the same must be attempted only after a thorough analysis, and a valid proof of concept.

Q2define nested loop join list and describe its variants? 5
Traditionally Nested-Loop join has been and is used in OLTP environments, but for many reasons, such a join mechanism is not suitable for VLDB and DSS environments. Nested loop joins are useful when small subsets of data are joined and if the join condition is an efficient way of accessing the inner table.
Nested-Loop Join: Variants
1. Naive nested-loop join
2. Index nested-loop join
3. Temporary index nested-loop join

Define Dense and Sparse index, adv and disadv (3)
For each record store the key and a pointer to the record in the sequential file. Why? It uses less space, hence less time to search. Time (I/Os) logarithmic in number of blocks used by the index can also be used as secondary index i.e. with another order of records.
Dense Index*:* Every key in the data file is represented in the index file
Pro: A dense index, if fits in the memory, costs only one disk I/O access to locate a record given a key
Con: A dense index, if too big and doesn‟t fit into the memory, will be expense when used to find a record given its key

Sparse Index: Concept

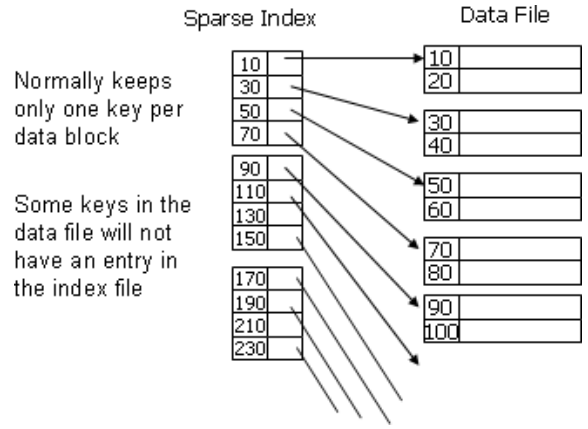


Figure-26.2: Sparse index concept

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

What happens when record 35 is inserted?

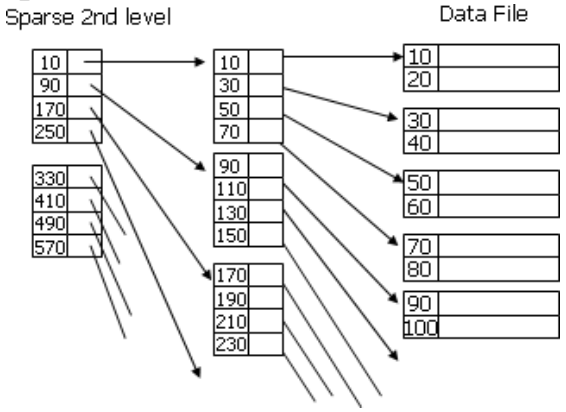
Sparse Index: Adv & Dis Adv

• Store first value in each block in the sequential file and a pointer to the block.

• Uses even less space than dense index, but the block has to be searched, even for unsuccessful searches.

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

Sparse Index: Multi level



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

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

Tasks performed through import/export data wizard

Tasks can be as follows:

- Establish connection through source / destination systems
- Creates similar table in SQL Server
- Extracts data from text files
- Apply very limited basic transformations if required
- Loads data into SQL Server table

Transient cookies

- Let the Web browser place a session-level cookie into the visitor's Web browser.
- Cookie value can serve as a temporary session ID

Limitations

You can't tell when the visitor returns to the site at a later time in a new session.

What is value validation process?

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

What is the difference between training data and test data?

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

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.

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

Why building a data warehouse is a challenging activity? What are the three broad categories of data warehouse development methods?

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

What are three fundamental reasons for warehousing Web data?

1. Web data is unstructured and dynamic, Keyword search is insufficient.
2. Web log contain wealth of information as it is a key touch point.
3. Shift from distribution platform to a general communication platform.

What types of operations are provided by MS DTS?

1. Providing connectivity to different databases
2. Building query graphically
3. Extraction data from disparate databases
4. Transforming data
5. Copying database objects
6. Providing support of different scripting languages (by default VB-script and Java –

What problems may be faced during Change Data Capture (CDC) while reading a log/journal tape?

Problems with reading a log/journal tape are many:

1. Contains lot of extraneous data
2. Format is often arcane
3. Often contains addresses instead of data values and keys
4. Sequencing of data in the log tape often has deep and complex
5. implications
6. Log tape varies widely from one DBMS to another.

What are seven steps for extracting data using the SQL server DTS wizard?

SQL Server Data Transformation Services (DTS) is a set of graphical tools and programmable objects that allow you extract, transform, and consolidate data from disparate sources into single or multiple destinations. SQL Server Enterprise .Manager provides an easy access to the tools of DTS.

Explain Analytic Applications Development Phase of Analytic Applications Track of Kimball's Model?

The DWH development lifecycle (Kimball‟s Approach) has three parallel tracks emanating from requirements definition.

These are

1. technology track,
2. data track and
3. Analytic applications track.

Analytic Applications Track:

Analytic applications also serve to encapsulate the analytic expertise of the organization, providing a jump-start for the less analytically inclined.

It consists of two phases.

1. Analytic applications specification
2. Analytic applications development

Analytic applications specification:

The main features of Analytic applications specification are:

3. Starter set of 10-15 applications.
4. Prioritize and narrow to critical capabilities.
5. Single template use to get 15 applications.
6. Set standards: Menu, O/P, look feel.
7. From standard: Template, layout, I/P variables, calculations.
8. Common understanding between business & IT users.

VU Answer

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,

Before designing the initial applications, it's important 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.

Analytic applications development:

The main features of Analytic applications development consist of:

1. Standards: naming, coding, libraries etc.
2. Coding begins AFTER DB design complete, data access tools installed, subset of historical data loaded.
3. Tools: Product specific high performance tricks, invest in tool-specific education.
4. Benefits: Quality problems will be found with tool usage => staging.
5. Actual performance and time gauged.

5 tech for de normalization (names)

Areas for Applying De-Normalization Techniques

Dealing with the abundance of star schemas.

Fast access of time series data for analysis.

Fast aggregate (sum, average etc.) results and complicated calculations.

Multidimensional analysis (e.g. geography) in a complex hierarchy.

Dealing with few updates but many join queries.

Q2) Describe physical extractions what is the difference offline and online extraction?(5)

Physical Extraction

Online Extraction

Offline Extraction

Legacy vs. OLTP

Depending on the chosen logical extraction method and the capabilities and restrictions on the source side, the extracted data can be physically extracted by two mechanisms. The data can either be extracted online from the source system or from an offline structure. Such an offline structure might already exist or it might be generated by an extraction routine.

Online Extraction

The data is extracted directly from the source system itself. The extraction process can connect directly to the source system to access the source tables themselves or to an intermediate system that stores the data in a preconfigured manner (for example, snapshot logs or change tables). Note that the intermediate system is not necessarily physically different from the source system. With online extractions, you need to consider whether the distributed transactions are using original source objects or prepared source objects.

Offline Extraction

The data is not extracted directly from the source system but is staged explicitly outside the original source system. The data already has an existing structure (for example, redo logs, archive logs or transportable table-spaces) or was created by an extraction routine.

2. Output of run length encoding (2 marks)

Run length used in bitmap indexing

Out put 1 may be

15#02# 18# (mean 1 come 5 time and 0 come 2 times and 1 come 1 8 times (111110011111111))

Out put 2 may be

11#01#11#

Output may 3 be

112#012#

Qualifying blocks question using outer table order?

i/o cost different

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.

If the outer loop executes R times and the inner loop executes S times, then the time complexity will be O(RS).

The time complexity should be independent of the order of tables i.e. O(RS) is same as O(SR). However, in the context of I/Os the order of tables does matter.

Along with this the relationship between the number of qualifyin g rows/blocks between the two

tables matters.

Join cost = Blocks accessed for *Table_A* + Blocks accessed for *Table_A* Blocks accessed for *Table_B*

Example :

Table_A = 500 blocks and
Table_B = 700 blocks. (www.vustudents.net)
Qualifying blocks for *Table_A QB(A)* = 50
Qualifying blocks for *Table_B QB(B)* = 100

Join cost A&B = 500 + 50□ □ 700 = 35,500 I/OsJoin
cost B&A = 700 + 100□ 500 = 50,700 I/Os
i.e. an increase in I/O of about 43%.
For example, if qualifying blocks for *Table_A* QB(A) = 50 and qualifying blocks for *Table_B* QB(B) = 100 and size of Table_A is 500 blocks and size of Table_B is 700 blocks then Join cost
A&B = 500 + 50□ 700 = 35,500 I/Os and using the other order i.e.*Table_B* outer table and *Table_A* as inner table, the join cost B&A = 700 + 100□ 500 = 50,700 I/Os i.e. an increase inI/O of about 43%.

Nested-Loop Join: Variants
1. Naive nested-loop join
2. Index nested-loop join
3. Temporary index nested-loop join

1. Naive nested-loop join
The simplest case is when an entire table is scanned; this is called a naive nested-loop join.
2. Index nested-loop join
If there is an index, and that index is exploited, then it is called an index nested-loop join
3. Temporary index nested-loop join
If the index is built as part of the query plan and subsequently dropped, it is called as a temporary index nested-loop join.

DTS Data Transformation Services (DTS)
DTS is 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)
• **DTS includes**
–Data Import/Export Wizard
–DTS Designer
–DTS Query Designer
–Package Execution Utilities
• DTS Tools can be accessed through “SQL Server Enterprise Manager”

What is Bit Mapped Index?
Answer: Bitmap indexes make use of bit arrays (bitmaps) to answer queries by performing bitwise logical operations. They work well with data that has a lower cardinality which means the data that take fewer distinct values. Bitmap indexes are useful in the data warehousing applications. Bitmap indexes have a significant space and performance advantage over other structures for such data. Tables that have less number of insert or update operations can be good candidates.

What are the advantages of Bit Mapped Index?
Answer: The advantages of Bitmap indexes are: They have a highly compressed structure, making them fast to read. Their structure makes it possible for the system to combine multiple indexes together so that they can access the underlying table faster.

What is the disadvantage of Bit Mapped Index?
Answer: The overhead on maintaining them is enormous.

What is Bi-directional Extract?
Answer: In hierarchical, networked or relational databases, the data can be extracted, cleansed and transferred in two directions. The ability of a system to do this is referred to as bidirectional extracts.

What is Data Collection Frequency?
Answer: Data collection frequency is the rate at which data is collected. However, the data is not just collected and stored. It goes through various stages of processing like extracting from various sources, cleansing, transforming and then storing in useful patterns. It is important to have a record of the rate at which data is collected because of various reasons: Companies can use these records to keep a track of the transactions that have occurred. Based on these records the company can know if any invalid transactions ever occurred. In scenarios where the market changes rapidly, companies need very frequently updated data to enable them make decisions based on the state of the market and then invest appropriately. A few companies keep launching new products and keep updating their records so that their customers can see them which would in turn increase their business. When data warehouses face technical problems, the logs as well as the data collection frequency can be used to determine the time and cause of the problem. Due to real time data collection, database managers and data warehouse specialists can make more room for recording data collection frequency.

Clustering vs. Cluster Detection (5 marks)

Solution:

In one-way clustering, reordering of rows (or columns) assembles clusters.

If the clusters are NOT assembled, they are very difficult to detect.

Once clusters are assembled, they can be detected automatically, using classical techniques such as K-means.

The *K-Means* Clustering

- Given *k*, the *k-means* algorithm is implemented in 4 steps:

1. Partition objects into *k* nonempty subsets
2. Compute seed points as the centroids of the clusters of the current partition. The centroid is the center (mean point) of the cluster.
3. Assign each object to the cluster with the nearest seed point.
4. Go back to Step 2, stop when no more new assignment.

PROS AND CONS OF KMEANS

1. K-means is a fairly fast technique and normally when terminates, then clusters formed are fairly good.
2. It can only work for data sets where there is the concept of mean (the answer to the question posed in a few slides back). If data is non numeric such as likes dislikes, gender, eyes color etc. then how to calculate means. So this is the first problem with the technique.
3. Another problem or limitation of the technique is that you have to specify the number of cluster in advance.
4. The third limitation is that it is not a robust technique as it not works well in presence of noise.
5. The last problem is that the clusters found by k-means have to be convex i.e. if you draw a polygon and join parameters of any two points in that polygon, that line goes out of the cluster boundaries.

12) Define the project planning task?

Lifecycle Key Steps

Lifecycle begins with project planning during which we assess the organization's readiness for a data warehouse initiative, establish the preliminary scope and justification, obtain resources, and launch the project.

Write down the limitations of pipelining parallelism?

Pipeline parallelism is a good fit for data warehousing (where we are working with lots of data), but it makes no sense for OLTP because OLTP tasks are not big enough to justify breaking them down into subtasks.