www.ranapk.com

presents

February 26, 2017 at 2:03pm

all mcq,s from past papers zain nasir and mooaz files
subjective
2=special indexing technique names

Inverted index
- Bit map index
- Cluster index
- Join indexes

2= why water fall model is used .

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.

2=ssl stand for

secure sockets layer (SSL)

2=why analytic track is called fun part kuch ise thra tha statment yad nhe sahe

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

3=dense index sparse index

Dense index: every key in the data file is represented in the index file.

Sparse index: normally keeps only one key per data block. Some keys in the data file will not have ann entity in the index file.

3=complete data warehouse deliverables

3=2 statment the btna ka k y kis topic me hen
1= measure of correctness of your model answer =accuracy
2= ability to the technique work accurately even in conditions of noisy or dirty data answer=Robustness:
5= precedence constraint wala tha on success on failuer wala

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
5= table dia tha accurasy find krne the and gender se related kuch column ko set krna tha

suppose there is a record with
name Firdous, time spent 15 minutes and 1 item purchased. We predict the gender by using our classification model and as per our model the customer is assigned 'F' (15/1=15
which is greater than 6). Thus we can easily predict the missing data with some reasonable accuracy using classification.

5=name of three automatic data cleasing technique other than clustring.

1) Statistical: Identifying outlier fields and records using the values of mean, standard deviation, range, etc., based on Chebyshev's theorem, considering the confidence intervals for each field
2) Pattern Based: Combined techniques (partitioning, classification, and clustering) are used to identify patterns that apply to most records
3) Clustering: Identify outlier records using clustering based on Euclidian (or other) distance
4) Association Rules: Association rules with high confidence and support define a different
kind of pattern.

March 1, 2017 at 4:31pm

CS614 Final Term Paper Fall 2016
By Juste

70% MCQS were from past papers. 3,4 from mid-term papers. Rest were new & conceptual.

Short & Long questions, most are from past papers:

1. What is unsupervised learning in data mining?

Type and number of classes are not known in advance.

- One-way Clustering
- Two-way Clustering

2. What is Reverse Proxy?

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.

3. Write any two Parallel Software Architectures?

   Parallel Software Architectures
   - Shared Memory
   - Shard Disk
   - Shared Nothing

4. Why RAD methodology is successful as compare to other methodologies? Write at least two reasons.

1. Assemble a small, very bright team of database programmers, hardware technicians, designers, quality assurance technicians, documentation and decision support specialists, and a single manager.
2. Define and involve a small "focus group" consisting of users (both novice and experienced) and managers (both line and upper). These are the people who will provide the feedback necessary to drive the prototyping cycle. Listen to them carefully.

5. Time complexity of K-means algorithm is O(tkn). What does "t", "k" and "n" represent here?

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

6. If we write "Date" in text format and then write in d-MMM-yy format. Explaination is required.

   d-MMM-yy.
   Profiling Date needs to change our flow of work. For columns before this we just did profiling and no transformation has been done yet. We can not profile date without transforming it. So we need to transform date first and then profile. The problem is, when
   we loaded all records to SQL server from text files they are loaded as strings (character
   arrays). While profiling we may want to get the range of dates (minimum and maximum
   dates). We can not identify date ranges until or unless we transform it as date data type.

7. In context of Web data warehousing, consider the web page dimension, list at least three possible attributes of this dimension.

   The page dimension is small. If the nominal width of a single row is 100 bytes and we have a big Web site with 100,000 pages, then the unindexed data size is 100 x 100,000 =

10 MB. If indexing adds a factor of 3, then the total size of this dimension is about 40 MB. Page key, page source, page function

8. Clustering is the part of automatic data Cleansing. Write three methods of automatic data Cleansing except Clustering.

Repeated

9. Identify the given statement correct or incorrect and justify in either case. A "Dense index consist of a number of bit vector or in B-tree"?

significant advantage of bitmapped indexes is that multiple bitmapped indexes can be used to evaluate the conditions on a single table. Thus, bitmapped indexes are very appropriate for complex ad-hoc queries that contain lengthy WHERE-clauses. Performance, storage requirements, and maintainability should be considered when evaluating an indexing scheme.

10. Why analytic track is called as fun part?

Repeated

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

You are students, and whom you meet were also once students.
• You can do an assessment of the company for DWH potential at no cost.
• Since you are only interested in your project, so your analysis will be neutral.
• Your report can form a basis for a professional detailed assessment at a later stage.
• If a DWH already exists, you can do an independent audit

12. One question was regarding SQL.

Q1: Reverse Proxy? (Repeated)

Q2: Define Wed Data Ware house ?

The central repository Data Warehouse
has changed to a Web data warehouse and data mining tools are customized to cater for the semi structured and dynamic web data. 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.

Q3: we can the session in world wide web by using identify the persistent cookies"how ever there are some limitation of the techniques explain two limitation ?

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

Q4:In context of nested loop join mention two guidelines for selecting a table as inner table?

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. Despite these restrictions/limitations, we will begin our discussion with the traditional join technique i.e. nested loop join, so that you can appreciate the benefits of the join techniques typically used in a VLDB and DSS environment.

Q5: Identify the given statement correct or incorrect dense index consists of a number of bit vectors?

Q6:clustering is considered to be one of the most important data cleasing techniques name any three automatic data cleasing techniques other than clustering.

Q7:Give two reasons why rapis application development (RAD) is more suitable for data werehouse development as compared to other traditional development methodologies?

Q8:Write down any two draw back if date is stored in text form rather than using proper date format like"dd-MMM-yy etc.

Q9:since the user community must accept the warehouse for it to be deemed succefull ,education is critical,being part of the traning team ,specify at least three guidelines that you will consider as part of effective user education programe.

Education: Since the user community must accept the warehouse for it to be deemed successful, education is critical. The education program needs to focus on the complete warehouse deliverable: data, analytic applications, and the data access tool (as appropriate). Consider the following for an effective education program; (i) *Understand your target audience*, don't overwhelm, *(ii) Don't train* the business community early prior to the availability of data and analytic applications, *(iii) No release, no education:* Postpone the education (and deployment) if the data warehouse is not ready to be released, (iv) Gain the sponsor's commitment to a "no *education, no access*" policy.

List down any 2 parallel hardware architecture? (2)

Answer:Symmetric Multi Processing (SMP)

Distributed Memory or Massively Parallel Processing (MPP)

Non-uniform Memory Access (NUMA)

list down any four ways to identify the session in World Wide Web? (2)

Answer:

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

What is Web Data Warehouse? (2)

Answer:

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.

How time contiguous log entries and HTTP secure socket layer are used for user session identification? What are the limitations of these techniques? (3)

Answer:

A session can be combined by collecting time-contiguous log entries from the same host

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

Nested loop efficient way of accessing the inner table? (3)

Answer:
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

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))

Why students allowed by company to visit company data warehouse? (5)

Answer:

August 20, 2016 at 11:51am

My CS-614 Paper:
Subjective:
1) Issues of Click stream

Clickstream data has many issues.
1. Identifying the Visitor Origin
2. Identifying the Session
3. Identifying the Visitor
4. Proxy Servers
5. Browser Caches

2) Lexical error

Lexical errors: For example, assume the data to be stored in table form with each row representing a tuple and each column an attribute. If we expect the table to have five columns because each tuple has five attributes but some or all of the rows contain only
four columns then the actual structure of the data does not conform to the specified format

3) Limitation of HTTP

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

4) Name of Activities in DWH lifecycle

Design, prototype,deploy,operate,enhance

5) Issues of Data Acquisition cleansing n agriculture

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.

6) Table of Bitmap index :page 234

7) Gender guidelines

8) Choose tha Data mining metrics

Accuracy, speed, robustness, scalability , interpretability

9) Agree with the statistic algorithm

Statistics

Data Mining uses statistical algorithms to discover patterns and regularities (or "knowledge") in data. For example: classification and regression trees (CART, CHAID),

rule induction (AQ, CN2), nearest neighbors, clustering methods, association rules, feature extraction, data visualization, etc.

Data mining is, in some ways, an extension of statistics, with a few artificial intelligence

and machine learning twists thrown in. Like statistics, data mining is not a business solution, it is just a technology.

10) statement is corrct/incorrect abut Orr's Law: 181

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!"

August 21, 2016 at 11:20pm

all objective was from past papers...subjective was from last chapters

and some questions were from old subjectives papers e.g

Identify the given statement as correct or incorrect and justify your answer in either case.

"The problem of Referential Integrity always occurs in traditional OLTP system as well as in DWH". 3m

Synchronization is probably the biggest reason we have referential integrity problems in a

DWH i.e. lack of synchronization between different extracts of the source system.

August 23, 2016 at 8:01pm

My today CS614 paper

45% paper from past papers

2 question asked correct or not..and explain reason? ek question m ye 2 statements thi(Orr's law statement)and (Inteligent learning organization share thier information.)
The intelligent learning organization shares
information openly across the enterprise in a way that maximizes the throughput of the
entire organization.

mostly question and MCQs of K-means algorithm.

Software architecture name btany thy 2 marks

Repeated

In K-means algorithm O(tkn) what is t,k and n?

Repeated

context dimention for web data ware house 2 marks ka tha

Describes the page context for a Web page event
ƒ Definition of page must be flexible
ƒ Assume a well defined function that
ƒ Characterizes the page
ƒ Describe the page
ƒ The page dimension is small

Data Parallelism

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.

Data mining vs statistics btana tha

Requirements Definition ka short question tha

User requirement definition
Set an appointment to meet business users.
ƒ Collect answers to questions.
ƒ Get a copy of sample input.
ƒ Get a copy of sample output.
ƒ Compile report of interview.
ƒ Identify business processes.
ƒ Identify problems.
ƒ Identify measures of success.

August 24, 2016 at 1:17pm

My current cs614 paper 2016

Q.What may be possible implications if the developing organization never freezes the requirements throughout the DWH development i.e. it always behaves like an accommodating person. 5

Mistake 3: *Never freezing the requirements i.e. being an accommodating person.* 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)

Write down any two drawbacks if "Date" is stored in text format rather than using proper date format like "dd-MMM-yy" etc. 5

repeated

There are different data mining techniques e.g. "clustering", "description" etc. Each of the following statement corresponds to some data mining technique. For each statement name the technique the statement corresponds to. 5

a) Assigning customers to predefined customer segments (i.e. good vs. bad)(classification)

b) Assigning credit applicants to predefined classes (i.e. low, medium, or high risk) (classification)

c) Guessing how much customers will spend during next 6 months. (prediction)

d) Building a model and assigning a value from 0 to 1 to each member of the set. Then classifying the members into categories based on a threshold value. (estimation)

e) Guessing how much students will score more than 65% grades in midterm.(prediction)

There are two justifications for a task to be performed in parallel, either it manipulates significant amount of data (i.e. size) or it can be solved by divide and conquer (D&C) strategy. From the given list, provide the justification for each of the task to perform it in parallel. 5

a) Large table scans and joins

b) Creation of large indexes

c) Partitioned index scans

d) Bulk inserts, updates, and deletes

e) Aggregations and copying

Every operation can not be parallelized, there are some preconditions and one of them being that the operations to be parallelized can be implemented independent of each other. This means that there will be no interference between the operations while they

are
being parallelized. So what do we gain out of parallelization; many things which can be
divided into two such as with reference to size and with reference to divide and conquer.
Note that divide and conquer means that we should be able to divide the problem and then
solve it and then compile the results i.e. conquer. For example in case of scanning a large
table every row has to be checked, in such a case this can be done in parallel thus reducing the overall time. There can be and are many examples too.

What are the tasks performed through import / export data wizard to load data?Write any three 3

Import and Export Data Wizard provides the easiest method of loading data.
ƒ The wizard creates package which is a collection of tasks
ƒ Tasks can be as follows:
ƒ Establish connection through source / destination systems
ƒ Creates similar table in SQL Server
ƒ Extracts data from text files
ƒ Apply very limited basic transformations if required
ƒ Loads data into SQL Server table

In context of Four Cell Quadrant Technique, which business process (from the diagram below) will have highest priority? Justify with reason. [Marks 3]

Once the findings have been reviewed, it is
time to prioritize. The four-cell quadrant technique (shown in Figure 33.2) is an effective
tool for reaching consensus on a data warehouse development plan that focuses on the right initial opportunities. The quadrant's vertical axis refers to the potential impact or value to the business. The horizontal axis conveys feasibility of each of the findings.

Consider the following two statements. Specify that each statement correspond to which activity of data quality analysis project. 3

a) Identify functional user data quality requirements and establish data quality metrics. (define)

b) Measure conformance to current business rules and develop exception reports. (measure)

Identify the given statement as correct or incorrect and justify your answer in either case.

“Bayesian Modeling is an example of Unsupervised Learning”. 3

Bayesian Modeling is an example of supervised Learning

The problems associated with the extracted data can correspond to non-primary keys. List down any four problems associated with the non-primary key.5

Non primary key problems…
1. Different encoding in different sources.
2. Multiple ways to represent the same information.
3. Sources might contain invalid data.
4. Two fields with different data but same name.

What is Reverse Proxy?2

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

Why analytics track is called as the “fun part" while designing a data warehouse?2

Remember, this is the fun part!" We're finally using the investment
in technology and data to help users make better decisions. The applications provide a key mechanism for strengthening the relationship between the project team and the business community. They serve to present the data warehouse's face to its business users, and
they bring the business needs back into the team of application developers

List two main types of unsupervised learning.2

One way clustering

Two way clustering

Q1 mark 5 ka tha

There are different data mining techniques e.g. “clustering”, “description” etc. Each of the following statement corresponds to some data mining technique. For each statement name the technique the statement corresponds to. 5m

Q2: 5 nmber ka tha

Write down any two drawbacks if “Date” is stored in text format rather than using proper date format like “dd-MMM-yy” etc. 5m
In context of Web data warehousing, consider the “web page” dimension, list at least five possible attributes of this dimension. 5m

Q3 : page key, page source, page function, page template, page filename, item type, graphic type

We can identify the Session in Word Wide Web by using “Time-contiguous Log Entries” however there are some limitations of this technique. Briefly explain any two limitations. 3m

Q4:

- 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.

. Identify the given statement as correct or incorrect and justify your answer in either case. (3 marks)

"One-way clustering is used to get local view and Two-way clustering is used to get global view."

Q5: Biclustering
gives a local view of your data set while one-way clustering gives a global view

1. . Is there any strategy to standardize a column

Q6: In the this slide we may see three data management systems. Data is extracted from two systems, top and bottom, and is loaded into the standardized system shown in the middle. We may see two transformations over here. First one is name transformation and the other one is date transformation.

List down any 2 parallel hardware architecture (2)

Q 7: 3 nmber ka tha

A dataware house project is more like scientific research than anything in traditional informational system do you agree or not justify in either case(2)

Q8: The normal Information System (IS) approach emphasizes on knowing what the expected
results are before committing to action. In scientific research, the results are unknown up front, and emphasis is placed on developing a rigorous, step-by-step process to uncover the truth.

Q: Be a diplomat not a technologist"
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

February 28, 2016 at 5:18pm

MCQs mostly past mn sy thy
what is unsupervised learning in Data mining?
List down any two parallel software Architectures?
briefly explains any two types of precedence constraints that we can use in DTS
unconditional, on success and on failure
which common measurements can be used to measure the success of specific email, advertisement, marketing?
Number of visitors
• Number of sessions
• Most requested pages
• Robot activity
who record tha pest scouting data?
department of pest warning Punjab
difference between data mining and data warehouse
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*

and aik tha jis mn table dia hua tha
and index find krna tha
columns k
Write any three complete warehouse deliverable.

Data
- Analytic applications
- Data access tools
- Education tools

February 28, 2016 at 5:48pm

run length encoding

11110000111100000011111100000011111 INPUT
14#04#14#06#15#06#15 OUTPUT

requirement preplanning phase

special indexing techniques.....bitmap index, inverted and cluster

what is web data warehousing

nested loop join concept

data mining techniques

queries

and moaaz file for mcqs

remember in prayers

and plz share current papers

Write any three complete warehouse deliverable.

Data Mining Vs. Statistics

Formal statistical inference is assumption driven i.e. a hypothesis is formed and validated against the data.
- Data mining is discovery driven i.e. patterns and hypothesis are automatically extracted from data.

March 1, 2016 at 5:53pm

My Today CS614 Papar

Subjective

41.List down any four partitioning type of shared nothing RDBMS

Hash partitioning
- Key range partitioning.
- List partitioning.
- Round-Robin
- Combination

42.What is the purpose in justifing planning phase of kimball approuches for DWH

43.Name that activities tha is performed through DWH develpment life cycle.....

44.which scripting languages are use to perform complex transformation inDTS...
Providing support of different scripting
languages(by default VB-Script and J-Script)

45.Correct Statement ..... In a shred memory each proces has its own memory

In a shared-memory machine, all processors have access to
a common main memory. In a distributed-memory machine, each processor has its own
main memory,

46.in contextof nested loop join mentiuon two guide line .....

47.specify atleast o implication if you dont provide dacumentation as a part of DWH

48.what are task though import export data wizard to load Data writes any three

49.table given ,,bitmap index .....tickless type or flo..... kuch aisa quistion tha

50.Diffrience B/w stastics and Data mining as peramters type of data .....

51.what were the issue of data achvvation & clear approaches DWH .....\

52. How gender guide is used for a larg number gender is missing ..

A mechanism can be formulated to correct gender
• Use a standard gender guide
• Create another table “Gender guide” with
columns Name and Gender
• Copy distinct first names to “Gender guide”
• Manually put the gender of all names in “Gender
Guide”
• Transform St_Name in Exception such that first
name gets separated and stored in another
column
• Make a join of Exception table and Gender guide
to fill missing gender

Best of luck .

Duaon main Yaad rakhna ..

End friends help other and shared your papar .....

March 7, 2016 at 2:27am

which common measurements can be used to measure the success of specific email, advertisement, marketing?
Success of email or other marketing campaign can be measured by integrating with other operational systems. Common

measurements are: • Number of visitors • Number of sessions • Most requested pages • Robot activity Etc.

March 7, 2016 at 2:32am

difference between data mining and data warehouse
Ans:
#254 page

Write any three complete warehouse deliverable. 3 marks
Data
Analytic applications
Data access tools
Education tools

February 29, 2016 at 11:27am

Today's exam 29/2/2016

52 total Questions, following is the subjective part;

2 Marks

- What should be done in a case where golden copy of missing dates is not available?

Type of corruption or error
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 persom

- There are many variants of the traditional nested-loop join. Name any two.
- What is meant by the statement "Be a diplomat NOT a technologist" in the context of a data warehouse development project?
- List down any two Parallel Software Architectures.

3 Marks

- Briefly explain the three Major Approaches for accessing the information stored on the Web

  Keyboard based search, querying deep web source, random surfing
- How it is possible that a sparse index takes even less space than dense index?

Because sparse index keeps only one key per data block and some keys in the data file won't have an entry in index file. On the other hand in inex file every key in the data file is represented in the index file.

- Time complexity of K-means algorithm is $O(tkn)$. What does "t", "k" and "n" represent here?
- Data profiling is a process of gathering information about columns, what are the purposes that it must fulfill? Describe Briefly.

5 Marks

Before combining all tables we need to
standardize them

• Number and types of columns, date formats and storing conventions all of them should be consistent in each table
• The process of standardization requires transformation of data elements
• To identify the degree of transformation required we will perform data profiling

- Consider a software drink factory where soft drink bottles are bottled. Suppose bottling is completed in three (3) stages. Each stage takes one (1) second to complete. Suppose there are 10 bottles that are bottled. How much speed up you will get if you apply the pipelining concept to this bottling process

  Note: mention complete calculations.

  10 CRATES LOADS OF BOTTLES
  Sequential execution = 10 × T
  Fill bottle, Seal bottle, Label Bottle pipeline = T + T × (9-1)/3 = 4 × T
  Speed-up = 2.50

  S=NT/T+(N-1)*T/M

- Give two reasons, why Rapid Application Development (RAD) is more suitable for data warehouse development as compared to other traditional development methodologies.
- Suppose there is a large enterprise which uses the same server for the development and production environments. What problems can arise if it uses single server for both purposes?

  Sometimes it is possible that the server needs to be rebooted for the development environment. Having a separate development environment will prevent the production environment from being effected by this.
  • There may be interference while having different database environments on a single server. For example, having multiple long queries running on the development server could affect the performance on the production server, as both are same.

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

A mechanism can be formulated to correct gender
• Use a standard gender guide
• Create another table “Gender guide” with columns Name and Gender
• Copy distinct first names to “Gender guide”
• Manually put the gender of all names in “Gender Guide”
• Transform St_Name in Exception such that first name gets separated and stored in another column
• Make a join of Exception table and Gender guide to fill missing gender

March 6, 2016 at 9:01pm

parallel Hardware Architectures:
==========
Symmetric Multi Processing (SMP)
Distributed Memory or Masively parallel processing (MPP)
Non-Uniform Memory Access (NUMA)

Parallel software Architectures:

Share Memory
Share Disk
Share Nothing

Scripting languages are used to perform complete transformation in DTS packages.
VB- Script and J-Script

A data warehouse project is more like scientific research than anything in traditional IS!
agree with this statement
The normal Information System (IS) approach emphasizes on knowing what the expected
results are before committing to action. In scientific research, the results are unknown up
front, and emphasis is placed on developing a rigorous, step-by-step process to uncover
the truth.

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

Drawbacks of traditional web searches:

Limited to keyword based matching.
Can not distinguish between the contexts in which a link is used.
Coupling of files has to be done manually

How would you determine out table in Nested-Loop join?
The smallest number of qualifying rows, and/or
The largest numbers of I/Os required to locate the rows.

Write names of first two steps of Kimball DWH lifecycle?
Project planning
Business Requirements Definition

Write at least three name of Shared nothing RDBMS architecture? (3 marks)
Hash partitioning

Key range partitioning.
List partitioning.

Why RAD methodology is successful, write at least two reasons? (5 marks)

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

Pest scouting:

Pest scouting is a systematic field sampling process that provide field specific information on pest pressure and crop injury.

what is reverse proxy?

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.

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

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.

The pest scouting data is being constantly recorded by:

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

Differentiate between DDS, Data mining and Data Warehouse DWH

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.

write any three complete warehouse deliverable:

? Data
? Analytic applications
? Data access tools
? Education tools

DW Lifecycle- Step 4: Deployment

• The three tracks converge at deployment.
• Not natural, require substantial pre-planning, courage, will-power and honesty.
• Something like waiting for arrival of Baarat, as only then meals can be served
• The Dulaaha is the key like data.
• Should serve uncooked data i.e. Dulaaha under a Sehara (not possible now) or wait and miss the deadline.

Special Index techniques

==========

? Inverted index
? Bit map index
? Cluster index
? Join indexes

data mininig technique:

==========

classifications
estimation
prediction
market masket analysis
clustering

There are many variants of the traditional nested-loop join. Name any two.

==========

There are many variants of the traditional nested-loop join. The simplest case is when an
entire table is scanned; this is called a naive nested-loop join. If there is an index, and that
index is exploited, then it is called an 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.

three major approaches when accessing information stored on the Web:

==========

keyword-based search
querying deep web sources
Random surfing

How it is possible that a sparse index takes even less space than dense index?

==========

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

Time complexity of K-means algorithm is O(tkn). What does “t”, “k” and “n” represent here?

==========

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 large enterprise which uses the same server for the development and

production environments. What problems can arise if it uses single server for both purposes?
==========
Sometimes it is possible that the server needs to be rebooted for the development environment. Having a separate development environment will prevent the production environment from being effected by this. • There may be interference while having different database environments on a single server. For example, having multiple long queries running on the development server could affect the performance on the production server, as both are same. pipelining concept: ==========
Pipelining: Speed-Up Calculation Time for sequential execution of 1 task = T Data Warehousing 215 Time for sequential execution of N tasks = N * T (Ideal) time for pipelined execution of one task using an M stage pipeline = T (Ideal) time for pipelined execution of N tasks using an M stage pipeline = T + ((N-1) × (T/M)) Speed-up (S) =NT/{T+(N-1) * T/M} Data profiling is a process of gathering information about columns, what are the purposes that it must fulfill?
========== 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 d-MMM-yy format: ========== Profiling Date needs to change our flow of work. For columns before this we just did profiling and no transformation has been done yet. We can not profile date without transforming it. So we need to transform date first and then profile. The problem is, when we loaded all records to SQL server from text files they are loaded as strings (character arrays). While profiling we may want to get the range of dates (minimum and maximum dates). We can not identify date ranges until or unless we transform it as date data type. In context of Web data warehousing, consider the “web page” dimension, list at least five possible attributes of this dimension.
========== Page key Page source Page function Page template Item type Graphic type Animation type Sound type Page file name There are different data mining techniques e.g. “clustering”, “description” etc. Each of the following statement corresponds to some data mining technique. For each statement name the technique the statement corresponds to.
========== a) Assigning customers to predefined customer segments (i.e. good vs. bad) classification b) Assigning credit applicants to predefined classes (i.e. low, medium, or high risk) classification c) Guessing how much customers will spend during next 6 months prediction d) Building a model and assigning a value from 0 to 1 to each member of the set. Then classifying the members into categories based on a threshold value. Estimation e) Guessing how much students will score more than 65% grades in midterm. Prediction There are two primary techniques for gathering requirements i.e. interviews or facilitated sessions. Which technique is preferred by Ralph Kimball?
========== Both have their advantages and disadvantages. Interviews encourage lots of individual participation. They are also easier to schedule. Facilitated sessions may reduce the elapsed time to gather requirements, although they require more time commitment from each participant. Kimball prefers using a hybrid approach with interviews to gather the gory details and then facilitation to bring the group to consensus. any two types of precedence constraints that we can use in DTS.
========== Unconditional on success There are four categories of data quality improvement. Write any two. (2

marks) ============================== The four categories of Data Quality Improvement • Process • ƒSystem • ƒPolicy & Procedure • ƒData Design "One-way clustering is used to get local view and Two-way clustering is used to get global view." ============================== Incorrect One-way clustering gives global view and bi-clustering gives local view Accuracy: ======= Accuracy or confidence level = matches/ total number of matches Is there any strategy to standardize a column: ============================== no List at least 2 barrier of linear speedup. ============================== Amdahl' Law Startup interference Skew Write 2 limitation of persistent cookies: ============================== • It's possible that the visitor will have his or her browser set to refuse cookies or may clean out
his or her cookie file manually, so there is no absolute guarantee that even a persistent cookie
will survive.
? • Although any given cookie can be read only by the Web site that caused it to be created,
certain groups of Web sites can agree to store a common ID tag that would let these sites
combine their separate notions of a visitor session into a super session

Write three activities requirement preplanning phase:
==========
Requirements preplanning: This phase consists of activities like choosing the forum, identifying and preparing the requirements team and finally selecting, scheduling
and preparing the business representatives.

March 7, 2016 at 4:07pm

1. "Median is an example of Distributive Aggregate".
Answer :
image: http://resources.infolinks.com/banners/fb_300x250-3.jpg

No. Median is an example of Holistic aggregate.

q2;supervised learning in data mining
Answer:

Supervised learning is when you are performing DM the supporting information that helps in the DM process is also available. Information means you may know your data that how many groups or classes your data set contains. What are the properties of these classes or clusters?

q3: down any four ways to identify the session in World Wide Web
Answer:
There are several ways to do this

- ƒ Using Time-contiguous Log Entries
- ƒ Using Transient Cookies

- *f* Using HTTP's secure sockets layer (SSL)
- *f* Using session ID Ping-pong
- *f* Using Persistent Cookies

q4:we have seen many issues during data acquisition & cleansing in agriculture data warehouse case study. In your point of view why those issues have arisen? Briefly explain….5m q;
Answer :

Data cleansing and standardization is probably the largest part in an ETL exercise. For Agri-DWH major issues of data cleansing had arisen due to data processing and handling

at four levels by different groups of people i.e. (i) Hand recordings by the scouts at the field level (ii) typing hand recordings into data sheets at the DPWQCP office (iii)

photocopying of the scouting sheets by DPWQCP personnel and finally (iv) data entry or digitization by hired data entry operators.

Briefly explain any two types of precedence constraints that we can use in DTS….
Answer :
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.

Identify kerna tha k ye statement correct he ya incorrect aur reason btana tha

Bayesian modeling is an example of unsupervised learning”
Answer :

Bayesian modeling is an example of supervised learning.

ik table diya hoa tha btna tha p index table

: Types of partitioning used in shared nothing environment

Answer:

There are five Types of partitioning used in shared nothing environment

1. *f* Hash partitioning
2. *f* Key range partitioning.
3. *f* List partitioning.
4. *f* Round-Robin

5. *f* Combinations (Range-Hash & Range-List)

Which script languages are used to perform complex transformation in DTS package?

Answer: VB Script or Java Script

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

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