# CS614 – Data Warehouse
# Final TERM Solved Subjective

**What are three types of precedence Constraints that we can use in DTS?**
In DTS, you can use three types of precedence constraints, which can be accessed either through DTS Designer or programmatically:
**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.
**Ref: Handout Page No. 395**

**what is the difference between data matrix and similarity/dissimilarity in terms of rows and columns, which one is symmetric?**

**Data matrix**

- We can measure the similarity of the row1 in data matrix with itself that will be 1.
- 1 is placed at index 1, 1 of the similarity matrix.
- We compare row 1 with row 2 and the measure or similarity value goes at index 1, 2 of the similarity matrix and son.
- In this way the similarity matrix is filled.
- your data matrix has n rows and m columns then your similarity matrix will have n rows and n columns.

**Similarity/dissimilarity**

- Similarity or dissimilarity matrix is the measure the similarity
- time complexity of computing similarity/dissimilarity matrix
- m accounts for the vector or header size of the data.
- measure or quantify the similarity or dissimilarity
- Pearson correlation and Euclidean distance

It should be noted that the similarity between row1 and row2 will be same as between row 2 and 1. Obviously, the similarity matrix will then be a square matrix, **symmetric** and all values along the diagonal will be same (here 1
**Ref: Handout Page No. 270**

**Write ten common mistakes which occur during the development of the data warehouse**
We will avoid common mistakes that may halt data warehousing process.

1. Project proceeded for two months and nobody has touched the data.

2. End users are not involved hands-on from day one throughout the program.
3. IT team members doing data design (modelers and DBAs) have never used the access tools.
4. Summary tables defined before raw atomic data is acquired and base tables have been built.
5. Data design finished before participants have experimented with tools and live data.

**Ref: Handout Page No. 311**

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

First you will need to determine the source of the data. Depending on your selection, you might have to provide additional authentication information. For example, when importing data from another SQL Server database, you might be able to use Windows domain accounts, instead of SQL Server logins, while selecting Access or Oracle will force you to deal with different authentication choices. The choice of data source will also affect the available Advanced Connection Properties (which are OLE DB provider specific) displayed after clicking the Advanced button on the Choose a Data Source page of the wizard. On the next page of the wizard, you will be prompted for equivalent configuration options for destination of data transfer (including provider type and advanced connection properties).After you specify the source and destination, We will be asked to select one of three types of data that will be imported/exported:

**Ref: Handout Page No. 381**

**Why building a data warehouse is a challenging activity? What are the three broad categories of data warehouse development methods? 3 Marks**

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.

1. **Waterfall Model:**
   The model is a linear sequence of activities like requirements definition, system design, detailed design, integration and testing, and finally operations and maintenance. The model is used when the system requirements and objectives are known and clearly specified.
2. **RAD:**
   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.
3. **Spiral Model:**
   The model is a sequence of waterfall models which corresponds to a risk oriented iterative enhancement, and recognizes that requirements are not always available and clear when the system is first implemented

**Ref: Handout Page No. 283**

**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. 3 marks**

The major problem is the inconsistent data sources at different campuses. The attributes summarizes the data sources at two genders. The problem is non-standardized attributes across Genders, Different conventions for representing Gender across the store that uses 0/1 while store uses 1/0 for representing male and female respectively. Similarly, there are different conventions for representing degree attribute across different store.
**Ref: Handout Page No. 405**

**Write formula of pipeline speed up?**
**Speedup = Pipeline depth Clock Cycle$_{unpipelined}$**
**1 + Pipeline stall CPI Clock Cycle$_{pipelined}$**
**Link :**
http://www.google.com.pk/#hl=en&q=write+formula+of+pipeline+speed+up%3F....&oq=write+formula+of+pipeline+speed+up%3F....&gs_l=serp.3...1099734.1099734.7.1099972.1.1.0.0.0.0.0..0.0...0.0...1c.nvYnltI75y8&bav=on.2,or.r_gc.r_pw.,cf.osb&fp=6ed243b376015671&biw=1152&bih=720

**how data can be load in DTS**
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.
**Ref: Handout Page No. 373**

**Define nested loop join ?What are types of nested loop join? Explain them**
We define a join of two tables to be a process in which we combine rows of one table with rows of another. There are three types of nested loop join
1. Naive nested-loop join
2. Index nested-loop join
3. Temporary index nested-loop join
**Ref: Handout Page No. 243**

**Write the steps to create a packages in DTS?**
Package can be created by one of the following three methods:
–Import/Export wizard
–DTS Designer
–Programming DTS applications
**Ref: Handout Page No. 380**

**Why "Justification" is required in project planning?**
Justification requires an estimation of the benefits and costs associated with a data warehouse. The anticipated benefits grossly outweigh the costs. IT usually is responsible for deriving the expenses. You need to determine approximate costs for the requisite hardware and software. Data warehouses tend to expand rapidly, so be sure the estimates allow some room for short-term growth.

**Ref: Handout Page No. 292**

**What is Inverted Index in simple words.**
An inverted index is an optimized structure that is built primarily for retrieval, with update being only a secondary consideration. The basic structure inverts the text so that instead of the view obtained from scanning documents where a document is found and then its terms are seen (think of a list of documents each pointing to a list of terms it contains), an index is built that maps terms to documents (pretty much like the index found in the back of a book that maps terms to page numbers).
**Ref: Handout Page No. 232**

List and explain fundamental advantages of bit map indexing

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

**Ref: Handout Page No. 235**

**DTS Operation? Explain**
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)
**Ref: Handout Page No. 375**

**Difference in between 1 way and 2 way clustering**
**1.One-way Clustering-**means that when you clustered a data matrix, you used all the attributes. In this technique a similarity matrix is constructed, and then clustering is performed on rows. A cluster also exists in the data matrix for each corresponding cluster in the similarity matrix.
**2. Two-way Clustering/**Biclustering-here rows and columns are simultaneously clustered. No any sort of similarity or dissimilarity matrix is constructed. Biclustering gives a local view of your data set while one-way clustering gives a global view. It is possible that you first take global view of your data by performing one-way clustering and if any cluster of interest is found then you perform two-way clustering to get more details. Thus both the methods complement each other.
**Ref: Handout Page No. 271**

**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.
**Ref: Handout Page No. 299**

What are design operations that are discussed in agri DWH 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.

**Ref: Handout Page No. 347**

**How time contiguous log entries and HTTP secure socket layer are used for user session identification? Limitation of this**
Web-centric data warehouse applications require every visitor session (visit) to have its own unique identity
 The basic protocol for the World Wide Web, HTTP, stateless so session identity must be established in some other way.
 There are several ways to do this
 Using Time-contiguous Log Entries
 Using Transient Cookies
 Using HTTP's secure sockets layer (SSL)
 Using session ID Ping-pong
 Using Persistent Cookies
**Ref: Handout Page No. 364**

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

```
SELECT COUNT(SID)
FROM   REGISTRATION,STUDENT
WHERE  REGISTRATION.SID = STUDENT.SID
```

AND DISCIPLINE = 'TC'
AND 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.

**Ref: Handout Page No. 347**

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

- 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

**Ref: Handout Page No. 225**

**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

**Ref: Handout Page No. 354**

**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.

**Ref: Handout Page No. 119**

**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.

**Ref: Handout Page No. 348**

**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.

**Ref: Handout Page No. 283**

**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

**Ref: Handout Page No. 233**

**List and explain fundamental advantages of bit map indexing**

Bitmap Index: Advantages

- 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.
**Ref: Handout Page No. 235**

**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.
**Ref: Handout Page No. 236**

**What are major operations of data mining?**
• Classification
• Estimation
• Prediction
• Clustering
• Description
**Ref: Handout Page No. 259**

**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.
**Ref: Handout Page No. 381**

**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.
**Ref: Handout Page No. 218**

**Q12what type of anomalies exists if a table is in 2NF not in 3NF?**
The table is in 2NF but NOT in 3NF Tables in 2NF but not in 3NF contain modification anomalies
**Ref: Handout Page No. 46**

**What are three methods for creating a DTS package?**
–Import/Export wizard
–DTS Designer
–Programming DTS applications
**Ref: Handout Page No. 381**

**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
**Ref: Handout Page No. 229**

**Explain analytic data application specification in Kimball**
• 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
**Ref: Handout Page No. 290**
**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.
**Ref: Handout Page No. 290**

**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
**Ref: Handout Page No. 42**

**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.
**Ref: Handout Page No. 264**

**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.
**Ref: Handout Page No. 340**

**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
**Ref: Handout Page No. 276**

**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.
**Ref: Handout Page No. 212**

**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.

**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.
**Ref: Handout Page No. 373**

**Cleansing can be break down in Who many steps, write their names?**
One can break down the cleansing into six steps:

- element zing,
- standardizing,
- verifying,
- matching,
- house holding,
- documenting.

**Ref: Handout Page No. 168**

**Q12: What does u mean by "keep competition hot in context of production selection and transformation while designing a data warehouse ". 2 marks**

Keep the competition "hot" Even if a single winner is left, it is a good piece of advice that always keep at least two. What if you keep one The sole vendor may take benefit of the situation that he is the only player and create a situation favorable for him. He might get an upper hand in the bargaining process, and mold things according to his facility and benefit.

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

**Ref: Handout Page No. 305**

**Q13: Who merge column is selected in case of sort merge?**

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.

**Ref: Handout Page No. 243**

**Different b/w non key or key data 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. 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.

**Ref: Handout Page No. 231**

**"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).

**Ref: Handout Page No. 320**

**Dirty bit**

- 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
**Ref: Handout Page No. 438**

**What are the problem face industry when the growth in usage of master table file increase?**
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.
**Ref: Handout Page No. 12**

**Indexing using I/0 bottelneck?**
**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.
**Ref: Handout Page No. 221**

**What is hardware utilization different in DWH?**
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.
**Ref: Handout Page No. 24**

**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

**Ref: Handout Page No. 328**

**What is Click stream? Limitations?**
**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.
**Clickstream data has many issues with limitation**
1. Identifying the Visitor Origin
2. Identifying the Session
3. Identifying the Visitor
4. Proxy Servers
5. Browser Caches
**Ref: Handout Page No. 363**

**Import/export wizard tasks?**
• First of all load data
1. Connect to source Text files
2. Connect to Destination SQL Server
3. Create new database 'Lahore_Campus' for example
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
**Ref: Handout Page No. 412**

**In case of non-uniform distribution, what will be the impact on performance?**
Parallelization is based on the premise that there is a full utilization of the processors and all of them are bust most or all of the time. However, if there is a skew in the partitioning of data i.e. a non-uniform distribution, then some of the processors will be working while other will be idle.
**Ref: Handout Page No. 219**

**Problem using SQL to fill up tables of ROLAP cube?**
**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
**Ref: Handout Page No. 87**

**How data mining is different from statics? Which one is better?**
**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.
**Ref: Handout Page No. 256**

**Persistent cookies limitations?**
**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
**Ref: Handout Page No. 367**

**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
**Ref: Handout Page No. 198**

**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.
**Ref: Handout Page No. 341**

**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.
**Ref: Handout Page No. 259**

**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.
**Ref: Handout Page No. 334**

**Define nested loop join list and describe its variants?**
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
**Ref: Handout Page No. 239**

**Define Dense and Sparse index, adv and disadv**
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**
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
**Ref: Handout Page No. 223**

**What should be done 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
**Ref: Handout Page No. 456**

**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.
**Ref: Handout Page No. 159**

**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.
**Ref: Handout Page No. 261**

**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.**
The major problem is the inconsistent data sources at different campuses. The attributes summarizes the data sources at two genders. The problem is non-standardized attributes across Genders, Different conventions for representing Gender across the store that uses 0/1 while store uses 1/0 for representing male and female respectively. Similarly, there are different conventions for representing degree attribute across different store.
**Ref: Handout Page No. 405**

**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.

1. **Waterfall Model:**
   The model is a linear sequence of activities like requirements definition, system design, detailed design, integration and testing, and finally operations and maintenance. The model is used when the system requirements and objectives are known and clearly specified.
2. **RAD:**
   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.
3. **Spiral Model:**
   The model is a sequence of waterfall models which corresponds to a risk oriented iterative enhancement, and recognizes that requirements are not always available and clear when the system is first implemented

**Ref: Handout Page No. 283**

**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.

**Ref: Handout Page No. 351**

**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 –

**Ref: Handout Page No. 375**

**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.

**Ref: Handout Page No. 151**

**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.
**Ref: Handout Page No. 299**

**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
**Ref: Handout Page No. 306**

**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.
**Ref: Handout Page No. 306**

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.
**Ref: Handout Page No. 307**

**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.
**Ref: Handout Page No. 307**

**Tech for de normalization (names)**
Areas for Applying De-Normalization Techniques

**What are the two extremes for technical architecture design? Which one is better?**
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. But in reality, this implies huge cost in disk space and the time for constructing the pre-aggregates. In the other case where performance is free i.e. infinitely fast machines and infinite number of them, then there is not need to build any summaries. Meaning zero cube space and zero pre-calculations, and in reality this would result in minimum performance boost, in the presence of infinite performance.
**Ref: Handout Page No. 95**

**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.
**Ref: Handout Page No. 351**

**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 –

**Ref: Handout Page No. 375**

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

1. Problems with reading a log/journal tape are many:
2. Contains lot of extraneous data
3. Format is often arcane
4. Often contains addresses instead of data values and keys
5. Sequencing of data in the log tape often has deep and complex
6. implications
7. 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.

**Ref: Handout Page No. 372**

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

Ans:

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.

**Ref: Handout Page No. 372**

**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:
- 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.

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 consisits 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.

When we do work into the development phase for the analytic applications, we again need to focus on standards. Standards for

- naming conventions,
- calculations,
- libraries, and
- coding

should be established to minimize future rework. The application development activity can begin once the database design is complete, the data access tools and metadata are installed, and a subset of historical data has been loaded. The application template specifications should be revisited to account for the inevitable changes to the data model since the specifications were completed.
We should take approperiate-specific education or supplemental resources for the development team.
While the applications are being developed, several ancillary benefits result. Application developers, should have a robust data access tool, quickly will find needling problems in the

data haystack despite the quality assurance performed by the staging application. we need to allow time in the schedule to
address any flaws identified by the analytic applications.
After realistically test query response times developers now reviewing performance-tuning strategies. The application development quality-assurance activities cannot be completed until the data is stabilized. We need to make sure that there is adequate time in the schedule beyond the final data staging cutoff to allow for an orderly wrap-up of the application development tasks.
**Ref: Handout Page No. 306**

**Elaborate the concept of data parallelism.**
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
**Ref: Handout Page No. 212**

**What is meant by the classification process? How we measure the accuracy of classifiers?**
Classification means that based on the properties of existing data, we have made or groups i.e. we have made classification.
**Ref: Handout Page No. 259**

**What are the issues regarding the record management tools at campuses where text files are used to store data?**
Main issues
Data duplication
Update the data
Data deletion
We can easily elaborate these issues

**What is fully de-normalized and highly de-normalized in DWH**
'Denormalization' does not mean that anything and everything goes. Denormalization does not mean chaos or disorder or indiscipline. The development of properly denormalized data structures follows software engineering principles, which insure that information will not be lost. De-normalization is the process of selectively transforming normalized relations into un-normalized physical record specifications, with the aim of reducing query processing time. Another fundamental purpose of denormalization is to reduce the number of physical tables that must be accessed to retrieve the desired data by reducing the number of joins required to answer a query.
**Ref: Handout Page No. 50**

**Differentiate between Range partitioning and Expression Partitioning**

Range and expression splitting:

- Can facilitate partition elimination with a smart optimizer.
- Generally lead to "hot spots" (uneven distribution of data).

Round-robin spreads data evenly across the partitions, but does not facilitate partition elimination (for the same reasons that hashing does not facilitate partition elimination). Round-robin is typically used only for temporary tables where partition elimination is not important and co-location of the table with other tables is not expected to yield performance benefits (hashing allows for co-location, but round-robin does not).

Roundrobin

is the “cheapest” partitioning algorithm that guarantees an even distribution of workload across the table partitions.

The most common use of range partitioning is on date. This is especially true in data warehouse deployments where large amounts of historical data are often retained. Hot spots typically surface when using date range partitioning because the most recent data tends to be accessed most frequently.

**Ref: Handout Page No. 66**