DWDM SESSIONAL - I IMPORTANT QUESTIONS

UNIT –1

1. Define Data warehouse.

- **Data Warehouse:**

1. A data warehouse refers to a data repository that is maintained separately from an organization's operational databases.
2. Data warehouse systems allow for integration of a variety of application systems.
3. They support information processing by providing a solid platform of consolidated historic data for analysis

- According to William H. Inmon, a leading architect in the construction of data warehouse systems.

"A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision making process".

NOTE:The four keywords—subject-oriented, integrated, time-variant, and nonvolatile—distinguish data warehouses from other data repository systems, such as relational database systems, transaction processing systems, and file systems

2. OLAP vs OLTP, OLAP Operations.

OLAP operations:

There are five basic analytical operations that can be performed on an OLAP cube:

1. **Drill down:** In drill-down operation, the less detailed data is converted into highly detailed data. It can be done by:
   - Moving down in the concept hierarchy
   - Adding a new dimension

4. **Slice:** It selects a single dimension from the OLAP cube which results in a new sub-cube creation. In the cube given in the overview section, Slice is performed on the dimension Time = "Q1".

3. **Dice:** It selects a sub-cube from the OLAP cube by selecting two or more dimensions. In the cube given in the overview section, a sub-cube is selected by selecting following dimensions with criteria:
   - Location = "Delhi" or "Kolkata"
   - Time = "Q1" or "Q2"
   - Item = "Car" or "Bus"

2. **Roll up:** It is just opposite of the drill-down operation. It performs aggregation on the OLAP cube. It can be done by:
   - Climbing up in the concept hierarchy
   - Reducing the dimensions

5. **Pivot:** It is also known as *rotation* operation as it rotates the current view to get a new view of the representation. In the sub-cube obtained after the slice operation, performing pivot operation gives a new view of it.

3. Schemas available in Multidimensional Data model with neat sketch.

Stars, Snowflakes, and Fact Constellations: Schemas for Multidimensional Data Models.

The most popular data model for a data warehouse is a **multidimensional model,**which can exist in the form of a **star schema, a snowflake schema, or a fact constellation schema.**

Star Scheme:

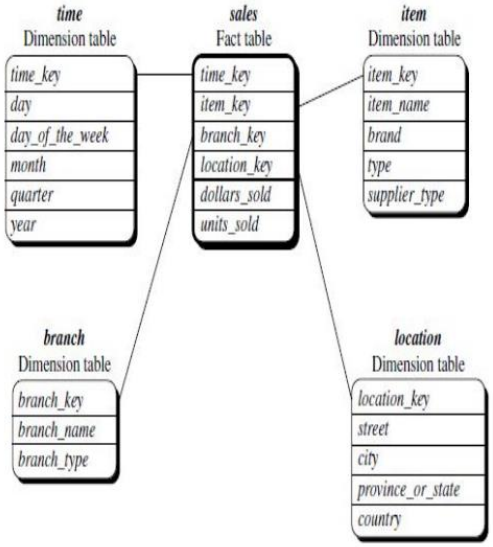
The most common modeling paradigm is the star schema, in which the data warehouse contains (1) a large central table (**fact table) containing the bulk of** the data, with no redundancy, and

(2) a set of smaller attendant tables (**dimension tables), one for each dimension.**

The schema graph resembles a starburst, with the dimension tables displayed in a radial pattern around the central fact table.

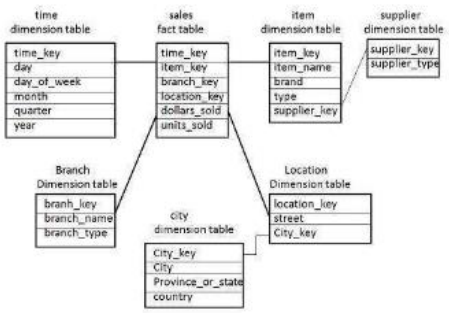
Stars, Snowflakes, and Fact Constellations: Schemas for Multidimensional Data Models.

- **Star Scheme:** Sales are considered along four dimensions: time, item, branch, and location. The schema contains a central fact table for sales that contains keys to each of the four dimensions, along with two measures: dollars sold and units sold. To minimize the size of the fact table, dimension identifiers (e.g., time key and item key) are system-generated identifiers.



Snowflake Schema

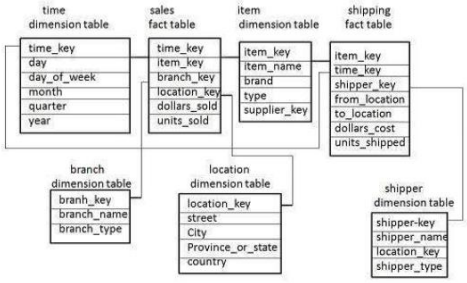
- Some dimension tables in the Snowflake schema are normalized.
- The normalization splits up the data into additional tables.
- Unlike Star schema, the dimensions table in a snowflake schema are normalized. For example, the item dimension table in star schema is normalized and split into two dimension tables, namely item and supplier table.



- Now the item dimension table contains the attributes item_key, item_name, type, brand, and supplier-key.
- The supplier key is linked to the supplier dimension table. The supplier dimension table contains the attributes supplier_key and supplier_type.

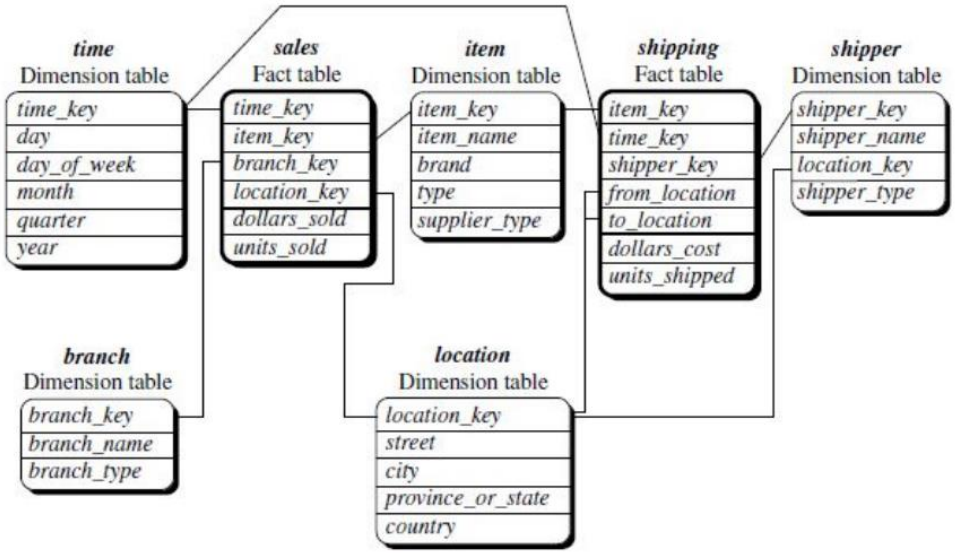
Fact Constellation Schema

- A fact constellation has multiple fact tables. It is also known as galaxy schema.
- The following diagram shows two fact tables, namely sales and shipping.



- The sales fact table is same as that in the star schema.
- The shipping fact table has the five dimensions, namely item_key, time_key, shipper_key, from_location, to_location.
- The shipping fact table also contains two measures, namely dollars sold and units sold.
- It is also possible to share dimension tables between fact tables. For example, time, item, and location dimension tables are shared between the sales and shipping fact table.

Stars, Snowflakes, and Fact Constellations: Schemas for Multidimensional Data Models.



4. 3 tier Data Ware house Architecture.

Three-Tier Data Warehouse Architecture

Data Warehouses usually have a three-level (tier) architecture that includes:

1. Bottom Tier (Data Warehouse Server)
2. Middle Tier (OLAP Server)
3. Top Tier (Front end Tools).

Bottom Tier(Data sources and data storage) :

1. The bottom Tier usually consists of Data Sources and Data Storage.
2. It is a warehouse database server. For Example RDBMS.
3. In Bottom Tier, using the application program interface(called gateways), data is extracted from operational and external sources.
4. Application Program Interface likes ODBC(Open Database Connection), OLE-DB(Open-Linking and Embedding for Database), JDBC(Java Database Connection) is supported.

Middle Tier :

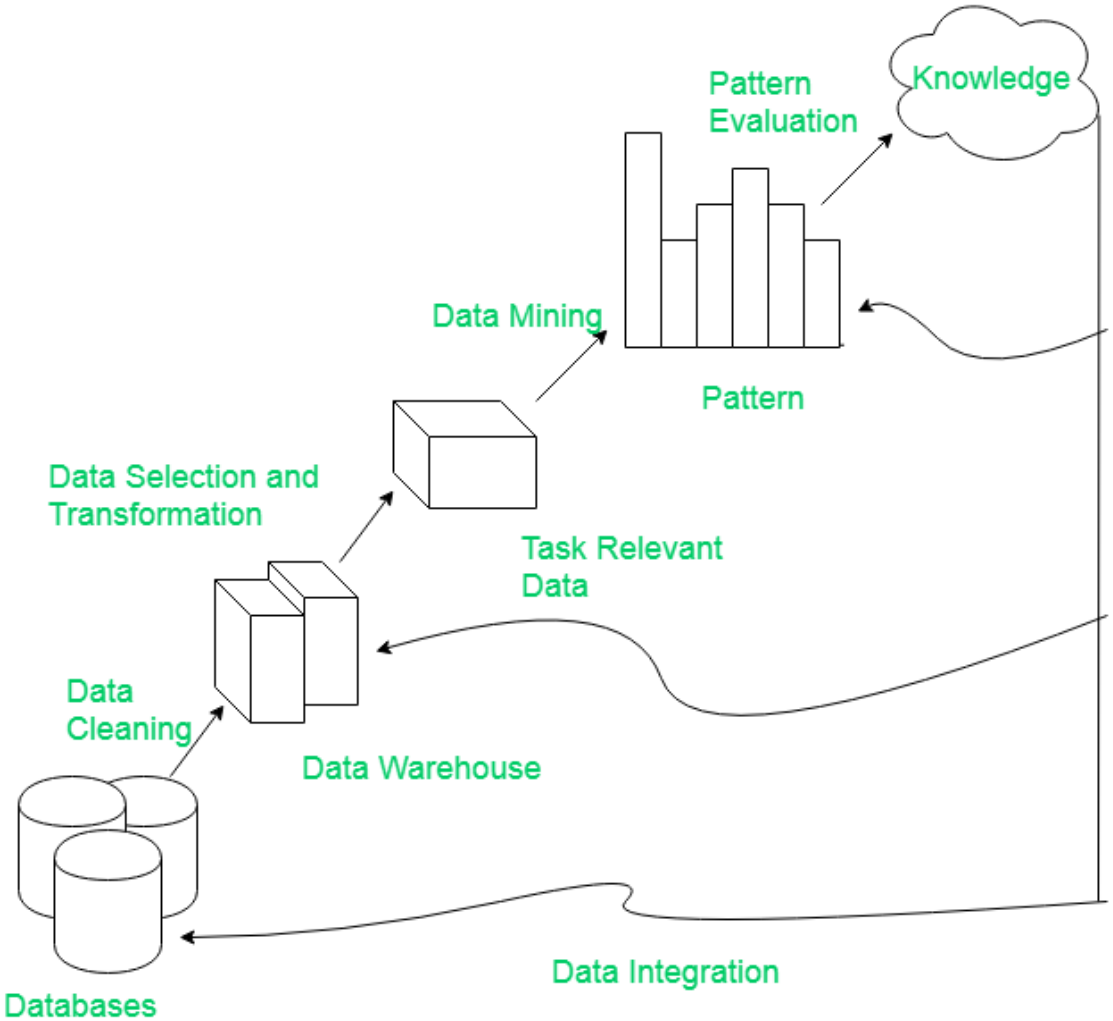
The middle tier is an OLAP server that is typically implemented using either :
A relational OLAP (ROLAP) model (i.e., an extended relational DBMS that maps operations from standard data to standard data); **or** A multidimensional OLAP (MOLAP) model (ie, a special purpose server that directly implements multidimensional data and operations).

Top Tier :

The top tier is a front-end client layer, which includes query and reporting tools, analysis tools, and/or data mining tools (eg, trend analysis, prediction, etc.).

5. Define Data Mining and Explain Data mining as a KDD Process with neat sketch.

Data Mining is defined as the procedure of extracting information from huge sets of data. In other words, we can say that data mining is mining knowledge from data. The tutorial starts off with a basic overview and the terminologies involved in data mining and then gradually moves on to cover topics such as knowledge discovery, query language, classification and prediction, decision tree induction, cluster analysis, and how to mine the Web.



1. ***Data Cleaning***: Data cleaning is defined as removal of noisy and irrelevant data from collection.
   - Cleaning in case of ***Missing values***.
   - Cleaning ***noisy*** data, where noise is a random or variance error.
   - Cleaning with ***Data discrepancy detection*** and ***Data transformation tools***.

2. ***Data Integration***: Data integration is defined as heterogeneous data from multiple sources combined in a common source(DataWarehouse).
   - Data integration using ***Data Migration tools***.
   - Data integration using ***Data Synchronization tools***.
   - Data integration using ***ETL***(Extract-Load-Transformation) process.

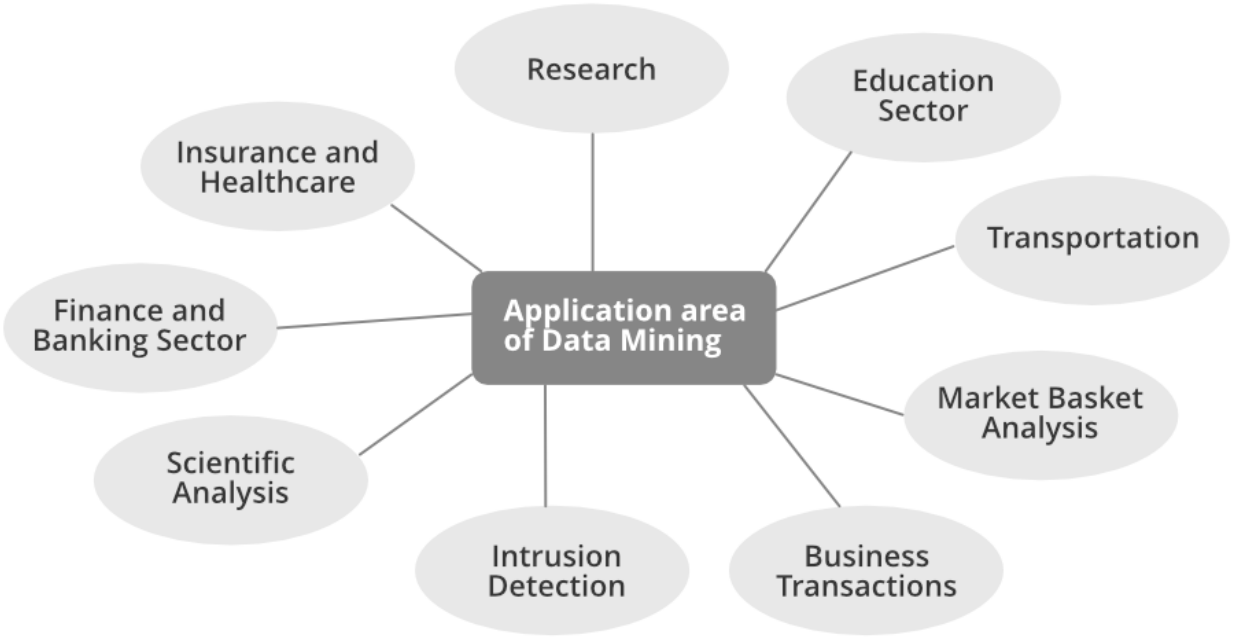
3. ***Data Selection***: Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection.
   - Data selection using ***Neural network***.
   - Data selection using ***Decision Trees***.
   - Data selection using ***Naive bayes***.
   - Data selection using ***Clustering***, ***Regression***, etc.
4. ***Data Transformation***: Data Transformation is defined as the process of transforming data into appropriate form required by mining procedure.
   Data Transformation is a two step process:
   - ***Data Mapping***: Assigning elements from source base to destination to capture transformations.
   - ***Code generation***: Creation of the actual transformation program.

5. ***Data Mining***: Data mining is defined as clever techniques that are applied to extract patterns potentially useful.
   - Transforms task relevant data into ***patterns***.
   - Decides purpose of model using ***classification*** or ***characterization***.

6. ***Pattern Evaluation***: Pattern Evaluation is defined as identifying strictly increasing patterns representing knowledge based on given measures.
   - Find ***interestingness score*** of each pattern.
   - Uses ***summarization*** and ***Visualization*** to make data understandable by user.

7. ***Knowledge representation***: Knowledge representation is defined as technique which utilizes visualization tools to represent data mining results.
   - Generate ***reports***.
   - Generate ***tables***.
   - Generate ***discriminant rules***, ***classification rules***, ***characterization rules***, etc.

5. Applications of Data Mining in detail.



Healthcare and Insurance: A Pharmaceutical sector can examine its new deals force activity and their outcomes to improve the focusing of high-value physicians and figure out which promoting activities will have the best effect in the following upcoming months, Whereas the Insurance sector, data mining can help to predict which customers will buy new policies, identify behavior patterns of risky customers and identify fraudulent behavior of customers.

Financial Data Analysis

The financial data in banking and financial industry is generally reliable and of high quality which facilitates systematic data analysis and data mining. Some of the typical cases are as follows −

- Design and construction of data warehouses for multidimensional data analysis and data mining.
- Loan payment prediction and customer credit policy analysis.
- Classification and clustering of customers for targeted marketing.
- Detection of money laundering and other financial crimes.

Transportation: A diversified transportation company with a large direct sales force can apply data mining to identify the best prospects for its services. A large consumer merchandise organization can apply information mining to improve its business cycle to retailers.

- Determine the distribution schedules among outlets.
- Analyze loading patterns.

UNIT– 2

1. Typesof attributes.

- **Attribute (**or **dimensions, features, variables**): a data field, representing a characteristic or feature of a data object.
  - *E.g., customer _ID, name, address*
- Types:
  - Nominal
  - Binary
  - Numeric: quantitative
    - Interval-scaled
    - Ratio-scaled

Attribute Types

- **Nominal:** categories, states, or "names of things"
  - *Hair_color = {auburn, black, blond, brown, grey, red, white}*
  - marital status, occupation, ID numbers, zip codes
- **Binary**
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - e.g., gender
  - Asymmetric binary: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to most important outcome (e.g., HIV positive)
- **Ordinal**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - *Size = {small, medium, large}*, grades, army rankings

Numeric Attribute Types

- Quantity (integer or real-valued)
- **Interval**
  - Measured on a scale of **equal-sized units**
  - Values have order
    - E.g., *temperature in C˚or F˚, calendar dates*
  - No true zero-point
- **Ratio**
  - Inherent **zero-point**
  - We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚).
    - e.g., *temperature in Kelvin, length, counts, monetary quantities*

2. Characteristics and Types of Data sets.

1. **Dimensionality** — # of attributes (very high leads to **Curse of Dimensionality**: it means many types of Data Analysis become difficult as the dimensionality of the data set increases. Specifically, the data becomes increasingly sparse in the space that it occupies.
2. **Sparsity** — # of non-zero entries is significantly less compared to the entire data
3. **Resolution** — identification of patterns in the data. Too fine: pattern not visible, Too coarse: pattern may disappear

Types of Datasets

In Statistics, we have different types of data sets available for different types of information. They are:

- Numerical data sets
- Bivariate data sets
- Multivariate data sets
- Categorical data sets
- Correlation data sets

Numerical Datasets

The numerical data set is a data set, where the data are expressed in numbers rather than natural language. The numerical data is sometimes called quantitative data. The set of all the quantitative data/numerical data is called the numerical data set. The numerical data is always in the numbers form, such that we can perform arithmetic operations on it.

Bivariate Datasets

A data set that has two variables is called a Bivariate data set. It deals with the relationship between the two variables. Bivariate dataset usually contains two types of related data.

Multivariate Datasets

A data set with multiple variables. When the dataset contains three or more than three data types (variables), then the data set is called a multivariate dataset. In other words, the multivariate dataset consists of individual measurements that are acquired as a function of three or more than three variables.

Correlation Datasets

The set of values that demonstrate some relationship with each other indicates correlation data sets. Here the values are found to be dependent on each other.

Categorical Datasets

Categorical data sets represent features or characteristics of a person or an object. The categorical dataset consists of a categorical variable also called the qualitative variable, that can take exactly two values. Hence, it is termed as a dichotomous variable. Categorical data/variables with more than two possible values are called polytomous variables. The qualitative/categorical variables are often assumed to be polytomous variable unless otherwise specified.

3. Measures of proximity b/w objects that involve multiple attributes (Euclidean Distance, Manhattan Distance, Minkowski Distance)
4. Statistical Descriptions of Data.

In statistics, there are two main categories:

- **Descriptive Statistics:** The purpose of descriptive statistics is to organize data and identify the main characteristics of that data. Graphs or numbers summarize the data. Average, Mode, SD(Standard Deviation), and Correlation are some of the commonly used descriptive statistical methods.
- **Inferential Statistics:** The process of drawing conclusions based on probability theory and generalizing the data. By analyzing sample statistics, you can infer parameters about populations and make models of relationships within data.

- **Correlation Analysis:** In statistical terms, correlation analysis captures the relationship between variables in a pair. The value of such variables is usually stored in a column or rows of a database table and represents a property of an object.
- **Regression Analysis:** Based on a set of numeric data, regression is a data mining method that predicts a range of numerical values (also known as continuous values). You could, for instance, use regression to predict the cost of goods and services based on other variables. A regression model is used across numerous industries for forecasting financial data, modeling environmental conditions, and analyzing trends.

- **Discriminant Analysis:** A Discriminant Analysis is a statistical method of analyzing data based on the measurements of categories or clusters and categorizing new observations into one or more populations that were identified a priori. The discriminant analysis models each response class independently then uses Bayes's theorem to flip these projections around to estimate the likelihood of each response category given the value of X. These models can be either linear or quadratic.
  - **Linear Discriminant Analysis:** According to **Linear Discriminant Analysis,** each observation is assigned a discriminant score to classify it into a response variable class. By combining the independent variables in a linear fashion, these scores can be obtained. Based on this model, observations are drawn from a Gaussian distribution, and the predictor variables are correlated across all k levels of the response variable, Y. and for further details linear discriminant analysis

5. Problem of Data quality with examples.

Duplicate data

Multiple copies of the same records take a toll on computing and storing, but may also produce skewed or incorrect insights when undetected. One of the critical problems could be human error — someone simply entering data multiple times by accident — or an algorithm that went wrong.

Unstructured data

Many times, if data has not been entered correctly in the system, or some files may have been corrupted, the remaining data has many missing variables. For example, if the address does not contain a zipcode at all, the remaining details might be of little interest, because it will be challenging to determine the geographical dimension.

1. Incomplete Data

This is by far the most common issue when dealing with DQ. Key columns are missing information, failing ETL jobs or causing downstream analytics impact. The best way to fix this is to put in place a reconciliation framework control. The control would check the number of records passing through your analytical layers and alert when records have gone missing.

4. Hidden data

Most organizations use only a part of their data, while the remaining may be lost in data silos or dumped in data graveyards. For example, customer data available with sales may not get shared with the customer service team, losing an opportunity to create more accurate and complete customer profiles. Hidden data means missing out on discovering opportunities to improve services, design innovative products, and optimize processes.

5. Inconsistent data

When you're working with multiple data sources, it's likely to have mismatches in the same information across sources. The discrepancies may be in formats, or units, or sometimes spellings. Inconsistent data can also get introduced during migration or company mergers. If not reconciled constantly, inconsistencies in data tend to build up and destroy the value of data. Data-driven organizations keep a close watch on data consistency because they want only trusted data powering their analytics.

6. Too much data

While we focus on data-driven analytics and its benefits, too much data does not seem to be a data quality issue. But it is. When you are looking for data relevant to your analytical projects, it's possible to get lost in too much data. Business users, data analysts, and data scientists spend 80% of their time locating the right data and preparing it. Other data quality issues become more severe with the increasing volume of data, especially with streaming data and large files or databases.

6. Data Preprocessing Techniques:

a)Data Cleaning

1. Data Cleaning:

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

- **(a). Missing Data:**

  This situation arises when some data is missing in the data. It can be handled in various ways. Some of them are:

  1. **Ignore the tuples:**

     This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.

  2. **Fill the Missing values:**

     There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

- **(b). Noisy Data:**

  Noisy data is a meaningless data that can't be interpreted by machines.It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :

  1. **Binning Method:**

     This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

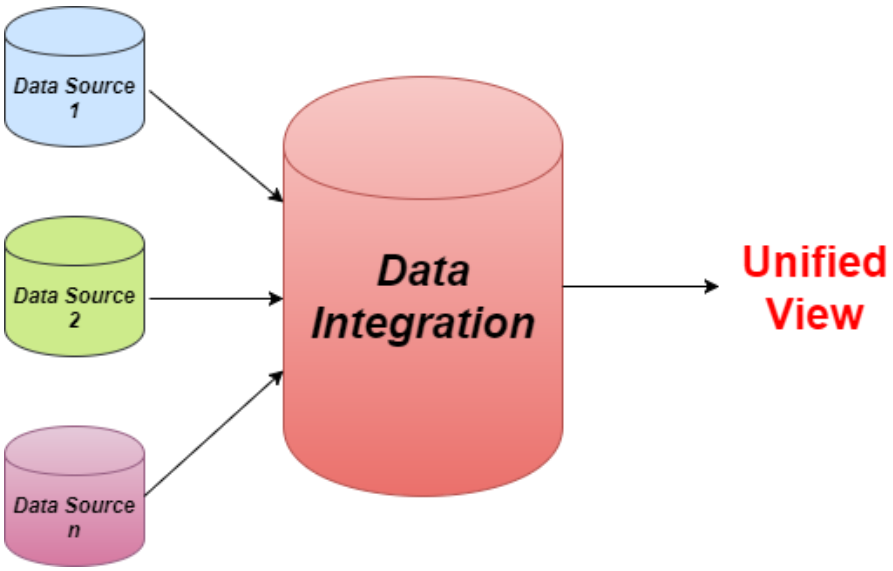
  2. **Regression:**

     Here data can be made smooth by fitting it to a regression function.The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).

  3. **Clustering:**

     This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

b)Data Integration



It is involved in a data analysis task that combines data from multiple sources into a coherent data store. These sources may include multiple databases. Do you think how data can be matched up ?? For a data analyst in one database, he finds Customer_ID and in another he finds cust_id, How can he sure about them and say these two belong to the same entity. Databases and Data warehouses have Metadata (It is the data about data) it helps in avoiding errors.

UNIT – 3

1. Define Classification.

There are two forms of data analysis that can be used to extract models describing important classes or predict future data trends. These two forms are as follows:

1. Classification
2. Prediction

Classification is to identify the category or the class label of a new observation. First, a set of data is used as training data. The set of input data and the corresponding outputs are given to the algorithm. So, the training data set includes the input data and their associated class labels. Using the training dataset, the algorithm derives a model or the classifier. The derived model can be a decision tree, mathematical formula, or a neural network. In classification, when unlabeled data is given to the model, it should find the class to which it belongs. The new data provided to the model is the test data set.

Classification is the process of classifying a record. One simple example of classification is to check whether it is raining or not. The answer can either be yes or no. So, there is a particular number of choices. Sometimes there can be more than two classes to classify. That is called ***multiclass classification***.

2. General approach to solve a classification problem (Steps in Data Classification).

A general approach to classification:

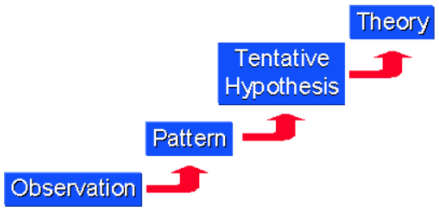
Classification is a two-step process involving,

Learning Step: It is a step where the Classification model is to be constructed. In this phase, training data are analyzed by a classification Algorithm.

Classification Step: it's a step where the model is employed to predict class labels for given data. In this phase, test data are wont to estimate the accuracy of classification rules.

3. Induction and deduction tasks in classification.

Inductive reasoning works the other way, moving from specific observations to broader generalizations and theories. Informally, we sometimes call this a "bottom up" approach (please note that it's "bottom up" and *not* "bottoms up" which is the kind of thing the bartender says to customers when he's trying to close for the night!). In inductive reasoning, we begin with specific observations and measures, begin to detect patterns and regularities, formulate some tentative hypotheses that we can explore, and finally end up developing some general conclusions or theories.



Deduction & Induction

In logic, we often refer to the two broad methods of reasoning as the *deductive* and *inductive* approaches.

Deductive reasoning works from the more general to the more specific. Sometimes this is informally called a "top-down" approach. We might begin with thinking up a *theory* about our topic of interest. We then narrow that down into more specific *hypotheses* that we can test. We narrow down even further when we collect *observations* to address the hypotheses. This ultimately leads us to be able to test the hypotheses with specific data – a *confirmation* (or not) of our original theories.