Scheme of Evaluation

Internal Assessment Test 1 – August 2024

Sub:	**Data Mining & Business Intelligence**							**Sub Code:**	**22MCA252**
Date:	14-08-24	Duration:	90 mins	Max Marks:	50	**Sem:**	I	**Branch:**	**MCA**

<table>
<tr><th>Q.NO</th><th>Description</th><th>Marks Distribution</th><th>Max Marks</th></tr>
<tr><td>1</td><td><b>Describe about the major tasks of data preprocessing.</b><ul><li>List out the tasks of data preprocessing and also explain it in brief.</li></ul></td><td><b>10</b></td><td><b>10</b></td></tr>
<tr><td>2</td><td><b>Define Data mining. List out the steps in data mining.</b><ul><li>Definition of Data mining</li><li>Explanation of KDD Process</li></ul></td><td><b>2</b>
<b>8</b></td><td><b>10</b></td></tr>
<tr><td>3</td><td><b>Define an efficient procedure for cleaning the noisy data</b><ul><li>List out the procedure for cleaning the noisy data</li><li>Explain it in brief</li></ul></td><td><b>2</b>
<b>8</b></td><td><b>10</b></td></tr>
<tr><td>4</td><td><b>Describe the issues of data mining.</b><ul><li>Explanation of the issues of data mining</li></ul></td><td><b>10</b></td><td><b>10</b></td></tr>
<tr><td>5</td><td><b>Explain Normalization by Min-Max score, z-score, and Decimal scaling with examples.</b><ul><li>Explanation of the Normalization and its types</li></ul></td><td><b>10</b></td><td><b>10</b></td></tr>
<tr><td>6</td><td><b>For the below given 2X2 contingency table with two attributes “gender” and “Preferred reading” conduct correlation analysis between the given attribute using Chi-square test.</b>
Referred reading Gender<table><tr><th></th><th>Male</th><th>Female</th><th>Total</th></tr><tr><td>Fiction</td><td>250</td><td>200</td><td>450</td></tr><tr><td>Non Fiction</td><td>50</td><td>1000</td><td>1050</td></tr><tr><td>Total</td><td>300</td><td>1200</td><td>1500</td></tr></table><ul><li>Solutions given by stepwise</li></ul></td><td><b>10</b></td><td><b>10</b></td></tr>
<tr><td>7</td><td><b>Define an Attribute. Explain the different types of Attributes</b><ul><li>Definition of Attribute</li><li>Explanation of different types of attribute</li></ul></td><td><b>10</b></td><td><b>10</b></td></tr>
</table>

8	**Suppose that the data for analysis include the attributed age. The age values for the data tuples are 13,15,16,19,20,20,21,22,22,25,25,25,25,30,33,33,35,35, 35,35,36,40,45,46,52,70. Bin size: 3 Illustrate your steps using Binning method.** • Solutions given by stepwise	**10**	**10**
9	**What is Data Warehouse? What are the characteristics and features of Data Warehouse?** • Definition of Data warehouse • Explanation of the characteristics and features of Data Warehouse	**5** **5**	**10**
10	**Explain in detail the building blocks of Data Warehouse.** • Explanation of the Data warehouse components	**10**	**10**

Internal Assessment Test 1 – August 2024

Data Mining and Business Intelligence							Sub Code:	22MCA252
14/08/2024	Duration:	90 min's	Max Marks:	50	Sem:	I	Branch:	MCA

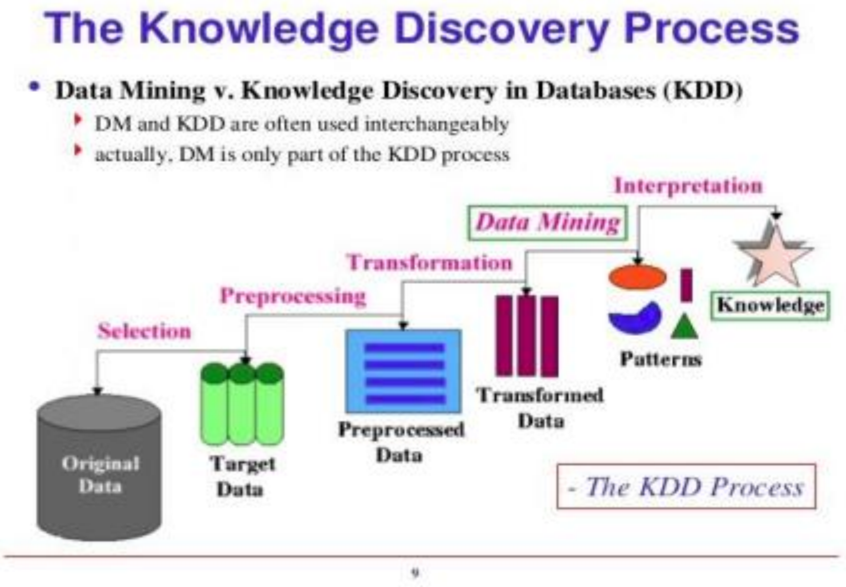
PART I

1. Describe about the major tasks of data preprocessing.

Data preprocessing is a crucial step in the data mining process, involving several tasks that prepare raw data for analysis. These tasks include:

1. **Data Cleaning**:
   - **Handling Missing Values**: This involves identifying and addressing missing data, either by removing incomplete records or imputing missing values using statistical methods.
   - **Smoothing Noisy Data**: Techniques such as binning, regression, or clustering are used to remove noise or outliers from the data.
   - **Correcting Inconsistencies**: Ensuring that data values are consistent across different datasets or within a single dataset.
2. **Data Integration**:
   - **Combining Data from Multiple Sources**: This involves merging data from different sources or databases to create a unified dataset. Techniques such as schema integration, entity resolution, and deduplication are used to ensure consistency and avoid redundancy.
3. **Data Transformation**:
   - **Normalization and Scaling**: Transforming data into a suitable format or range, such as scaling numerical attributes to a specific range or normalizing data to ensure consistency across features.
   - **Data Discretization**: Converting continuous attributes into discrete intervals or categories.
   - **Attribute Construction**: Creating new attributes or features by combining or transforming existing ones.
4. **Data Reduction**:
   - **Dimensionality Reduction**: Reducing the number of attributes in the dataset by using techniques like Principal Component Analysis (PCA) or feature selection methods.
   - **Data Compression**: Reducing the volume of data by encoding techniques or removing redundant information.
   - **Sampling**: Selecting a representative subset of the data to reduce the dataset size while maintaining its analytical value.
5. **Data Discretization and Concept Hierarchy Generation**:
   - **Discretization**: Transforming continuous data into categorical data by creating a set of ranges or intervals.
   - **Concept Hierarchy Generation**: Organizing data attributes into a hierarchy or levels of abstraction, which can simplify the analysis process.

2. **Define Data mining. List out the steps in data mining.**
Data Mining also known as Knowledge Discovery in Databases refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data stored in databases.



The Knowledge Discovery Process

- Data Mining v. Knowledge Discovery in Databases (KDD)
  - DM and KDD are often used interchangeably
  - actually, DM is only part of the KDD process

Figure 1.2 KDD Process

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 (Data Warehouse).
   ➢ 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 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.

PART II

3. Define an efficient procedure for cleaning the noisy data

Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data. Various methods for handling this problem:

Missing Values The various methods for handling the problem of missing values in data tuples include:

(a) Ignoring the tuple: This is usually done when the class label is missing (assuming the mining task involves classification or description). This method is not very effective unless the tuple contains several attributes with missing values. It is especially poor when the percentage of missing values per attribute varies considerably.

(b) Manually filling in the missing value: In general, this approach is time-consuming and may not be a reasonable task for large data sets with many missing values, especially when the value to be filled in is not easily determined.

(c) Using a global constant to fill in the missing value: Replace all missing attribute values by the same constant, such as a label like “Unknown,” or −∞. If missing values are replaced by, say, “Unknown,” then the mining program may mistakenly think that they form an interesting concept, since they all have a value in common — that of “Unknown.” Hence, although this method is simple, it is not recommended.

(d) Using the attribute mean for quantitative (numeric) values or attribute mode for categorical (nominal) values, for all samples belonging to the same class as the given tuple: For example, if classifying customers according to credit risk, replace the missing value with the average income value for customers in the same credit risk category as that of the given tuple.

(e) Using the most probable value to fill in the missing value: This may be determined with regression, inference-based tools using Bayesian formalism, or decision tree induction. For example, using the other customer attributes in your data set, you may construct a decision tree to predict the missing values for income.

Noisy data: Noise is a random error or variance in a measured variable. Data smoothing tech is used for removing such noisy data. Several Data smoothing techniques:

1 Binning methods: Binning methods smooth a sorted data value by consulting the neighborhood", or values around it. The sorted values are distributed into a number of 'buckets', or bins. Because binning methods consult the neighborhood of values, they perform local smoothing. In this technique,

1. The data for first sorted
2. Then the sorted list partitioned into equi-depth of bins.
3. Then one can smooth by bin means, smooth by bin median, smooth by bin
boundaries, etc.

a. Smoothing by bin means: Each value in the bin is replaced by the mean value of the bin.

b. Smoothing by bin medians: Each value in the bin is replaced by the bin median.

c. Smoothing by boundaries:

The min and max values of a bin are identified as the bin boundaries. Each bin value is replaced by the closest boundary value.

Example: Binning Methods for Data Smoothing • o Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34 o Partition into (equi-depth) bins(equi depth of 3 since each bin contains three values): -

Bin 1: 4, 8, 9, 15

Bin 2: 21, 21, 24, 25

Bin 3: 26, 28, 29, 34

Smoothing by bin means:

Bin 1: 9, 9, 9, 9

Bin 2: 23, 23, 23, 23

Bin 3: 29, 29, 29, 29

Smoothing by bin boundaries:

Bin 1: 4, 4, 4, 15

Bin 2: 21, 21, 25, 25

Bin 3: 26, 26, 26, 34

In smoothing by bin means, each value in a bin is replaced by the mean value of the bin. For example, the mean of the values 4, 8, and 15 in Bin 1 is 9. Therefore, each original value in this bin is replaced by the value 9. Similarly, smoothing by bin medians can be employed, in which each bin value is replaced by the bin median. In smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value.

Suppose that the data for analysis include the attribute age. The age values for the data tuples are (in increasing order): 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

(a) Use smoothing by bin means to smooth the above data, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given data. The following steps are required to smooth the above data using smoothing by bin means with a bin depth of 3.

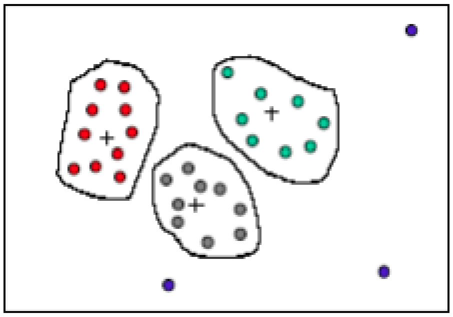
Step 1: Sort the data. (This step is not required here as the data are already sorted.)

Step 2: Partition the data into equi-depth bins of depth 3. Bin 1: 13, 15, 16 Bin 2: 16, 19, 20 Bin 3: 20, 21, 22 Bin 4: 22, 25, 25 Bin 5: 25, 25, 30 Bin 6: 33, 33, 35 Bin 7: 35, 35, 35 Bin 8: 36, 40, 45 Bin 9: 46, 52, 70

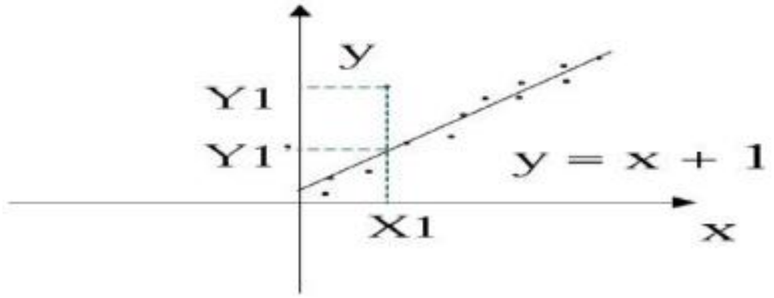
Step 3: Calculate the arithmetic mean of each bin.

Step 4: Replace each of the values in each bin by the arithmetic mean calculated for the bin. Bin 1: 14, 14, 14 Bin 2: 18, 18, 18 Bin 3: 21, 21, 21 Bin 4: 24, 24, 24 Bin 5: 26, 26, 26 Bin 6: 33, 33, 33 Bin 7: 35, 35, 35 Bin 8: 40, 40, 40 Bin 9: 56, 56, 56 2

Clustering: Outliers in the data may be detected by clustering, where similar values are organized into groups, or 'clusters'. Values that fall outside of the set of clusters may be considered outliers.



Regression : smooth by fitting the data into regression functions.
Linear regression involves finding the best of line to fit two variables, so that• one variable can be used to predict the other.



Multiple linear regression is an extension of linear regression, where more than two variables are involved and the data are fit to a multidimensional surface. Using regression to find a mathematical equation to fit the data helps smooth out the noise.
Field overloading: is a kind of source of errors that typically occurs when developers compress new attribute definitions into unused portions of already defined attributes.
Unique rule is a rule says that each value of the given attribute must be different from all other values of that attribute .
Consecutive rule is a rule says that there can be no missing values between the lowest and highest values of the attribute and that all values must also be unique.
Null rule specifies the use of blanks, question marks, special characters or other strings that may indicate the null condition and how such values should be handled.

4. Describe the issues of data mining.

Major issues in data mining is regarding mining methodology, user interaction, performance, and diverse data types
1 .Mining methodology and user-interaction issues:
Mining different kinds of knowledge in databases: Since different users can be interested in different kinds of knowledge, data mining should cover a wide spectrum of data analysis and knowledge discovery tasks, including data characterization, discrimination, association, classification, clustering, trend and deviation analysis, and similarity analysis. These tasks may use the same database in different ways and require the development of numerous data mining techniques.
Interactive mining of knowledge at multiple levels of abstraction: Since it is difficult to know exactly what can be discovered within a database, the data mining process should be interactive.
Incorporation of background knowledge: Background knowledge, or information regarding the domain under study, may be used to guide the discovery patterns. Domain knowledge related to databases, such as integrity constraints and deduction rules, can help focus and speed up a data mining process, or judge the interestingness of discovered patterns.
Data mining query languages and ad-hoc data mining: Knowledge in Relational query languages (such as SQL) required since it allow users to pose ad-hoc queries for data **Presentation and visualization of data mining results:** Discovered knowledge should be

expressed in high-level languages, visual representations, so that the knowledge can be easily understood and directly usable by humans

Handling outlier or incomplete data: The data stored in a database may reflect outliers: noise, exceptional cases, or incomplete data objects. These objects may confuse the analysis process, causing over fitting of the data to the knowledge model constructed. As a result, the accuracy of the discovered patterns can be poor. Data cleaning methods and data analysis methods which can handle outliers are required.

Pattern evaluation: refers to interestingness of pattern: A data mining system can uncover thousands of patterns. Many of the patterns discovered may be uninteresting to the given user, representing common knowledge or lacking novelty. Several challenges remain regarding the development of techniques to assess the interestingness of discovered patterns,

2. Performance issues.

These include efficiency, scalability, and parallelization of data mining algorithms.

Efficiency and scalability of data mining algorithms: To effectively extract information from a huge amount of data in databases, data mining algorithms must be efficient and scalable. **Parallel, distributed, and incremental updating algorithms:** Such algorithms divide the data into partitions, which are processed in parallel. The results from the partitions are then merged.

3. Issues relating to the diversity of database types

Handling of relational and complex types of data: Since relational databases and data warehouses are widely used, the development of efficient and effective data mining systems for such data is important.

Mining information from heterogeneous databases and global information systems: Local and wide-area computer networks (such as the Internet) connect many sources of data, forming huge, distributed, and heterogeneous databases. The discovery of knowledge from different sources of structured, semi-structured, or unstructured data with diverse data semantics poses great challenges to data mining.

PART III

5. Explain Normalization by Min-Max score, z-score, and Decimal scaling with examples.

Data transformation can involve the following:

Smoothing: which works to remove noise from the data

Aggregation: where summary or aggregation operations are applied to the data. •

For example, the daily sales data may be aggregated so as to compute weekly and annual total scores.

Generalization of the data: where low-level or “primitive” (raw) data are replaced • by higher-level concepts through the use of concept hierarchies. For example, categorical attributes, like street, can be generalized to higher-level concepts, like city or country. Normalization: where the attribute data are scaled so as to fall within a small • specified range, such as −1.0 to 1.0, or 0.0 to 1.0. Attribute construction (feature construction): this is where new attributes are • constructed and added from the given set of attributes to help the mining process. Normalization In which data are scaled to fall within a small, specified range, useful for classification algorithms involving neural networks, distance measurements such as nearest neighbor classification and clustering. There are 3 methods for data normalization. They are: 1) min-max normalization 2) z-score normalization 3) normalization by decimal scaling Min-max normalization: performs linear transformation on the original data values. It can be defined as,

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

v is the value to be normalized minA,maxA are minimum and maximum values of an attribute A new_ maxA, new_ minA are the normalization range. Z-score normalization / zero-mean normalization: In which values of an attribute A are normalized based on the mean and standard deviation of A. It can be defined as,

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

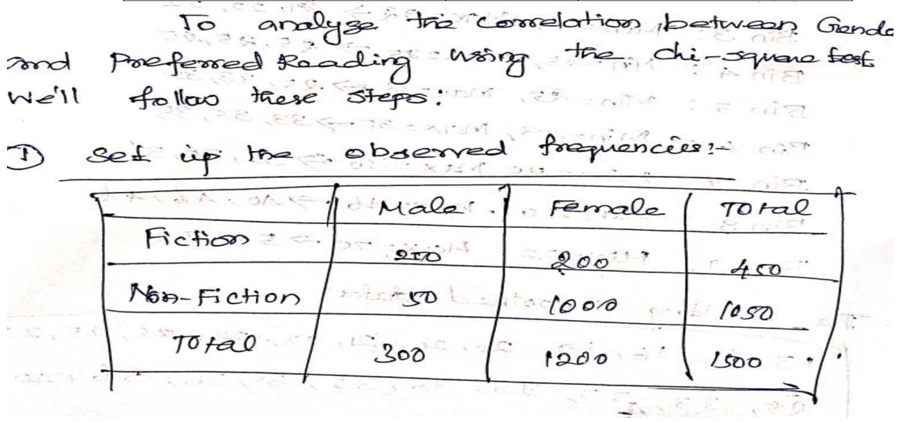
This method is useful when min and max value of attribute A are unknown or when outliers that are dominate min-max normalization. Normalization by decimal scaling: normalizes by moving the decimal point of values of attribute A. The number of decimal points moved depends on the maximum absolute value of A. A value v of A is normalized to v' by computing,

$$v' = \frac{v}{10^j} \quad \text{Where } j \text{ is the smallest integer such that } \mathrm{Max}(|v'|) < 1$$

6. **For the below given 2X2 contingency table with two attributes "gender" and "Preferred reading" conduct correlation analysis between the given attribute using Chi-square test.**

Referred reading Gender

	Male	Female	Total
Fiction	250	200	450
Non Fiction	50	1000	1050
Total	300	1200	1500



To analyze the correlation between Gende and Preferred Reading using the chi-square test we'll follow these steps:

1) Set up the observed frequencies:-

	Male	Female	Total
Fiction	250	200	450
Non-Fiction	50	1000	1050
Total	300	1200	1500

(2) Calculate the Expected frequenicies

The expected frequency for each cell is calculated using formula:

$$E = \frac{(\text{Row Total}) \times (\text{column Total})}{\text{Grand Total.}}$$

Fiction/Male:

$$E_{11} = \frac{450 \times 300}{1500} = 90$$

Fiction / Female:-

$$E_{12} = \frac{450 \times 1200}{1500} = 360$$

Non-Fiction/Male;

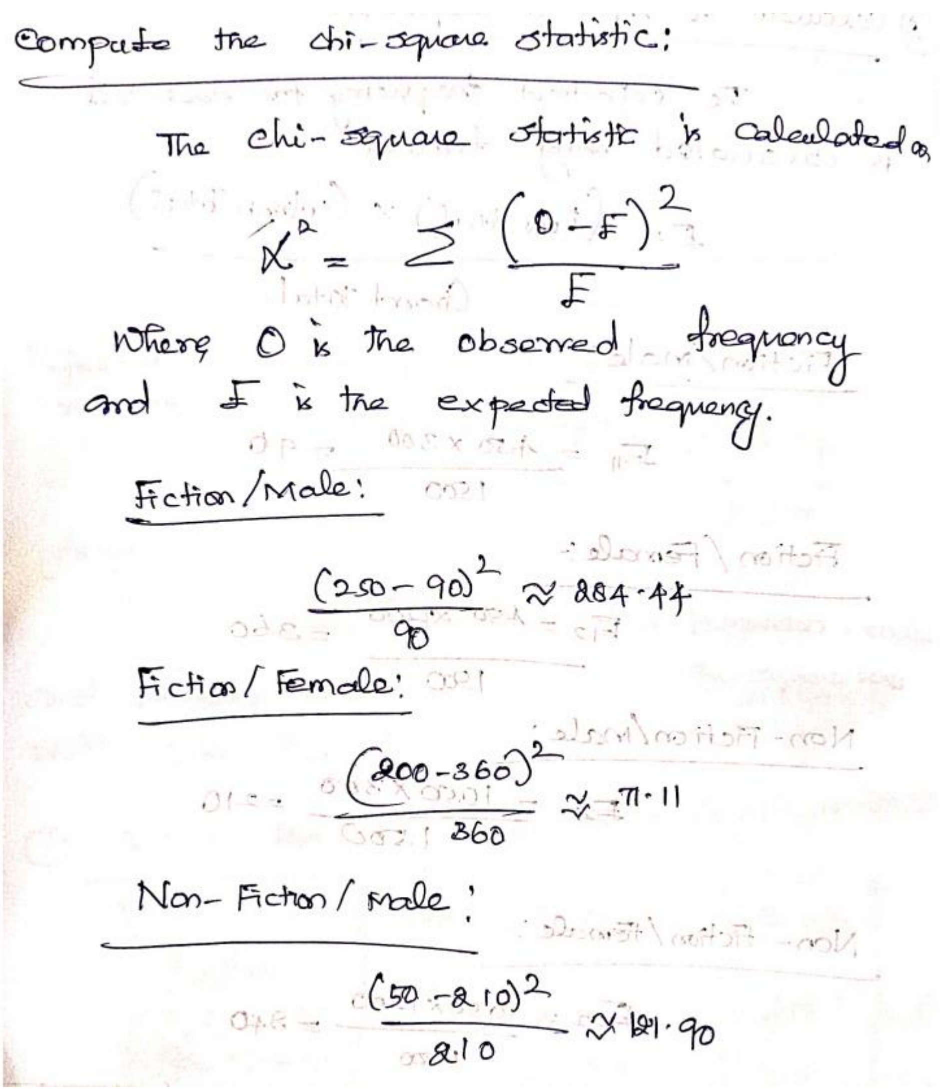
$$E_{21} = \frac{1050 \times 300}{1500} = 210$$

Non- Fiction / Female :-

$$E_{22} = \frac{1050 \times 1200}{1500} = 840$$

1500

	Male(Or)	Er	Female(Or	Er	Total
Fiction	250	90	200	360	450
Non-Fiction	50	210	1000	840	1050
Total	300		1200		1500



Compute the chi-square statistic:

The chi-square statistic is calculated as

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Where O is the observed frequency and E is the expected frequency.

Fiction/Male:

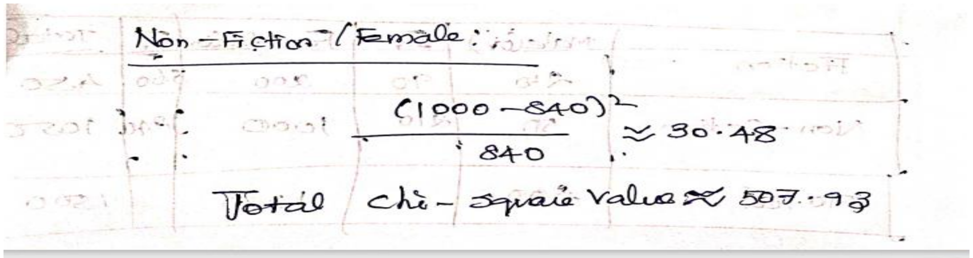
$$\frac{(250-90)^2}{90} \approx 284.44$$

Fiction/Female:

$$\frac{(200-360)^2}{360} \approx 71.11$$

Non-Fiction/Male:

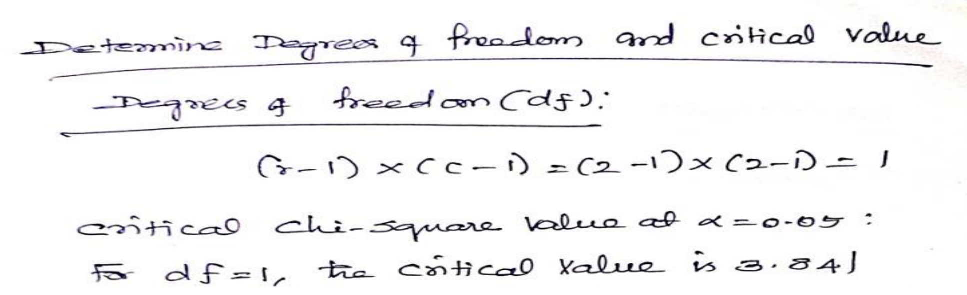
$$\frac{(50-210)^2}{210} \approx 121.90$$



Non-Fiction/Female:

$$\frac{(1000-840)^2}{840} \approx 30.48$$

Total chi-square value ≈ 507.93



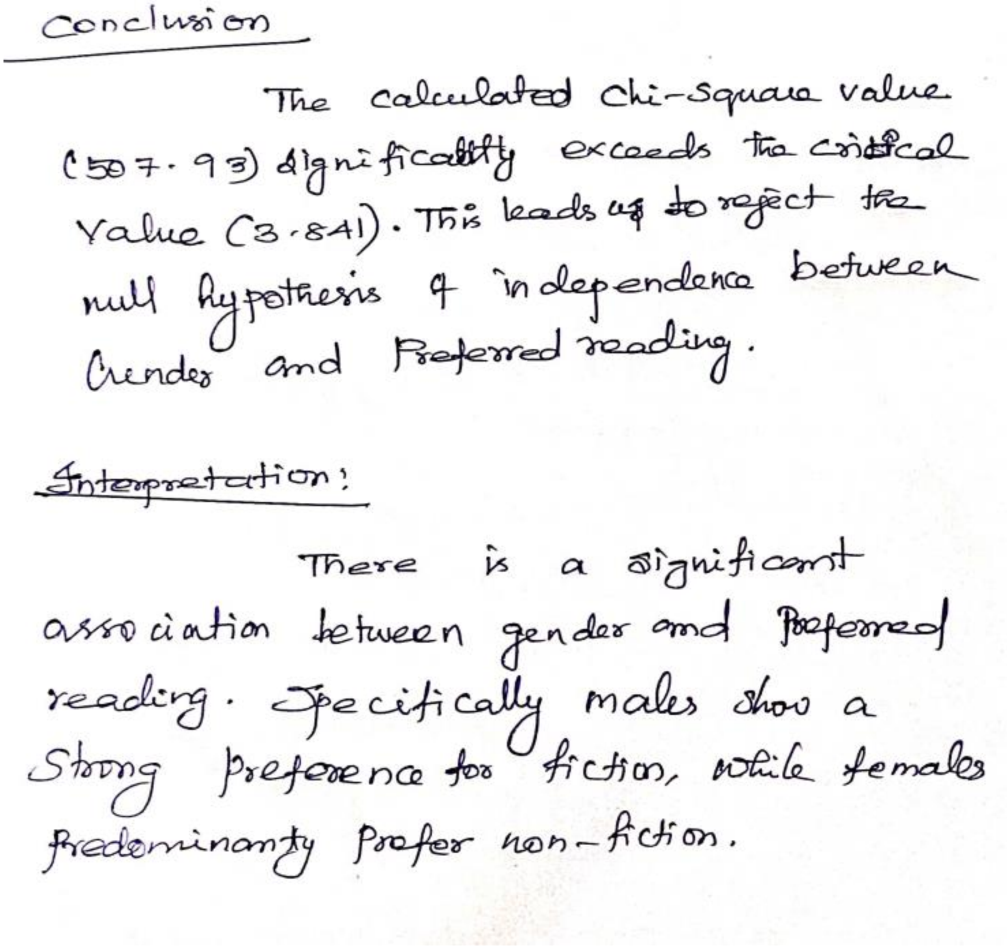
Determine Degrees of freedom and critical value

Degrees of freedom (df):

$$(r-1)\times(c-1) = (2-1)\times(2-1) = 1$$

critical chi-square value at $\alpha = 0.05$:

For df=1, the critical value is 3.841



Conclusion

The calculated Chi-square value (507.93) significantly exceeds the critical value (3.841). This leads us to reject the null hypothesis of independence between Gender and Preferred reading.

Interpretation:

There is a significant association between gender and preferred reading. Specifically males show a strong preference for fiction, while females predominanty prefer non-fiction.

PART IV

7. **Define an Attribute. Explain the different types of Attributes**

It can be seen as a data field that represents characteristics or features of a data object. For a customer object attributes can be customer Id, address etc. We can say that a set of attributes used to describe a given object are known as attribute vector or feature vector. Type of attributes: This is the First step of Data Data-preprocessing. We differentiate between different types of attributes and then pre process the data. So here is description of attribute types.

1. Qualitative (Nominal (N), Ordinal (O), Binary (B)).
2. Quantitative (Discrete, Continuous)

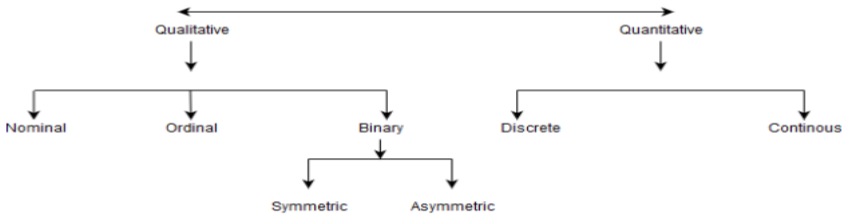
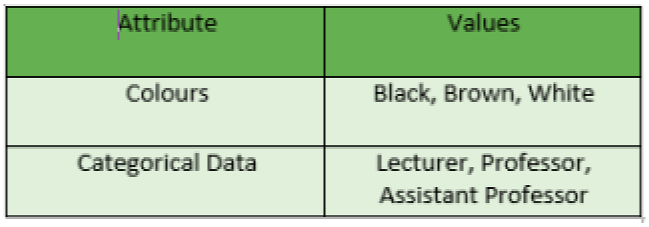


Figure 1.1 Type of attributes

Qualitative Attributes

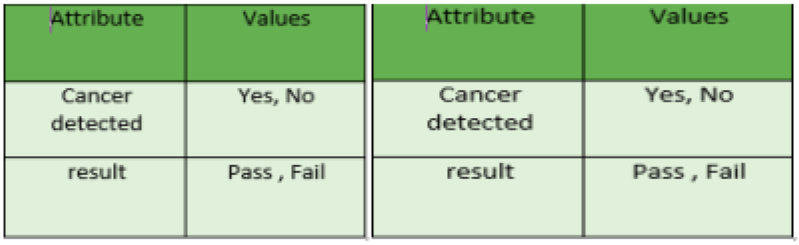
1. Nominal Attributes – related to names: The values of a Nominal attribute are name of things, some kind of symbols. Values of Nominal attributes represents some category or state and that’s why nominal attribute also referred as categorical attributes and there is no order among values of nominal attribute. Example



Attribute	Values
Colours	Black, Brown, White
Categorical Data	Lecturer, Professor, Assistant Professor

Table 1.1 Nominal Attributes

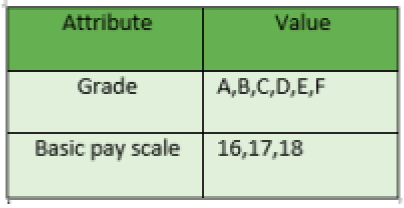
2. Binary Attributes: Binary data has only 2 values/states. For Example yes or no, affected or unaffected, true or false.
3. i) Symmetric: Both values are equally important (Gender).
ii) Asymmetric: Both values are not equally important (Result).



Attribute	Values	Attribute	Values
Cancer detected	Yes, No	Cancer detected	Yes, No
result	Pass , Fail	result	Pass , Fail

Table 1.2 binary Attributes

Ordinal Attributes : The Ordinal Attributes contains values that have a meaningful sequence or ranking(order) between them, but the magnitude between values is not actually known, the order of values that shows what is important but don’t indicate how important it is.



Attribute	Value
Grade	A,B,C,D,E,F
Basic pay scale	16,17,18

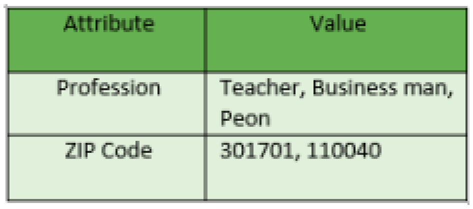
Table 1.3 Ordinal Attributes

Quantitative Attributes

1. **Numeric:** A numeric attribute is quantitative because, it is a measurable quantity, represented in integer or real values. Numerical attributes are of 2 types, interval and ratio.

   i) An **interval-scaled attribute** has values, whose differences are interpretable, but the numerical attributes do not have the correct reference point or we can call zero point. Data can be added and subtracted at interval scale but cannot be multiplied or divided. Consider an example of temperature in degrees Centigrade. If a day's temperature of one day is twice than the other day we cannot say that one day is twice as hot as another day.

   ii) A **ratio-scaled attribute** is a numeric attribute with an fix zero-point. If a measurement is ratioscaled, we can say of a value as being a multiple (or ratio) of another value. The values are ordered, and we can also compute the difference between values, and the mean, median, mode, Quantile-range and five number summaries can be given.
2. **Discrete:** Discrete data have finite values it can be numerical and can also be in categorical form. These attributes has finite or countable infinite set of values. Example

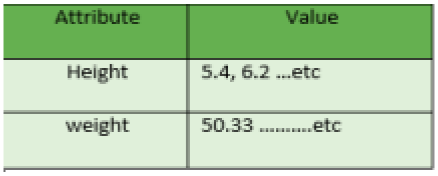


Attribute	Value
Profession	Teacher, Business man, Peon
ZIP Code	301701, 110040

Table 1.4 Discrete Attributes

3. Continuous: Continuous data have infinite no of states. Continuous data is of float type. There can be many values between 2 and 3.

   Example:



Attribute	Value
Height	5.4, 6.2 ...etc
weight	50.33 ....etc

Table 1.5 Continuous Attributes

8. **Suppose that the data for analysis include the attributed age. The age values for the data tuples are 13,15,16,19,20,20,21,22,22,25,25,25,25,30,33,33,35,35, 35,35,36,40,45,46,52,70. Bin size: 3 illustrate your steps using Binning method.**

To perform binning on the given data set using a bin size of 3, we can follow the following steps:

1. Sort the Data (if not already sorted):

The given data is already sorted:

13, 15, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70

2. Partition the Data into Bins:

With a bin size of 3, we'll divide the data into bins where each bin contains 3 consecutive values. Since there are 26 values, the last bin will contain only one value.

The bins are as follows:

- **Bin 1**: 13, 15, 16
- **Bin 2**: 19, 20, 20
- **Bin 3**: 21, 22, 22
- **Bin 4**: 25, 25, 25
- **Bin 5**: 25, 30, 33
- **Bin 6**: 33, 35, 35
- **Bin 7**: 35, 35, 36
- **Bin 8**: 40, 45, 46
- **Bin 9**: 52, 70

3. Apply Binning Method:

There are different methods to handle the values within each bin. We'll demonstrate **Smoothing by bin means**, **Smoothing by bin medians**, and **Smoothing by bin boundaries**.

Smoothing by Bin Means:

Replace each value in the bin with the mean of the bin.

- **Bin 1**: (13 + 15 + 16) / 3 = 14.67 → 14.67, 14.67, 14.67
- **Bin 2**: (19 + 20 + 20) / 3 = 19.67 → 19.67, 19.67, 19.67
- **Bin 3**: (21 + 22 + 22) / 3 = 21.67 → 21.67, 21.67, 21.67
- **Bin 4**: (25 + 25 + 25) / 3 = 25 → 25, 25, 25
- **Bin 5**: (25 + 30 + 33) / 3 = 29.33 → 29.33, 29.33, 29.33
- **Bin 6**: (33 + 35 + 35) / 3 = 34.33 → 34.33, 34.33, 34.33
- **Bin 7**: (35 + 35 + 36) / 3 = 35.33 → 35.33, 35.33, 35.33
- **Bin 8**: (40 + 45 + 46) / 3 = 43.67 → 43.67, 43.67, 43.67
- **Bin 9**: (52 + 70) / 2 = 61 → 61, 61

The resulting smoothed data:

14.67, 14.67, 14.67, 19.67, 19.67, 19.67, 21.67, 21.67, 21.67, 25, 25, 25, 29.33, 29.33, 29.33, 34.33, 34.33, 34.33, 35.33, 35.33, 35.33, 43.67, 43.67, 43.67, 61, 61

Smoothing by Bin Medians:

Replace each value in the bin with the median of the bin.

- **Bin 1**: Median of 13, 15, 16 = 15 → 15, 15, 15
- **Bin 2**: Median of 19, 20, 20 = 20 → 20, 20, 20
- **Bin 3**: Median of 21, 22, 22 = 22 → 22, 22, 22
- **Bin 4**: Median of 25, 25, 25 = 25 → 25, 25, 25
- **Bin 5**: Median of 25, 30, 33 = 30 → 30, 30, 30
- **Bin 6**: Median of 33, 35, 35 = 35 → 35, 35, 35
- **Bin 7**: Median of 35, 35, 36 = 35 → 35, 35, 35
- **Bin 8**: Median of 40, 45, 46 = 45 → 45, 45, 45
- **Bin 9**: Median of 52, 70 = 61 → 61, 61

The resulting smoothed data:

15, 15, 15, 20, 20, 20, 22, 22, 22, 25, 25, 25, 30, 30, 30, 35, 35, 35, 35, 35, 35, 45, 45, 45, 61, 61

Smoothing by Bin Boundaries:

Replace each value in the bin with the closest boundary value (either minimum or maximum of the bin).

- **Bin 1**: Min = 13, Max = 16 → 13, 16, 16
- **Bin 2**: Min = 19, Max = 20 → 19, 20, 20
- **Bin 3**: Min = 21, Max = 22 → 21, 22, 22
- **Bin 4**: Min = 25, Max = 25 → 25, 25, 25
- **Bin 5**: Min = 25, Max = 33 → 25, 33, 33
- **Bin 6**: Min = 33, Max = 35 → 33, 35, 35
- **Bin 7**: Min = 35, Max = 36 → 35, 36, 36

- **Bin 8**: Min = 40, Max = 46 → 40, 46, 46
- **Bin 9**: Min = 52, Max = 70 → 52, 70

The resulting smoothed data:

13, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 33, 33, 33, 35, 35, 36, 36, 36, 40, 46, 46, 52, 70

PART V

9. **What is Data Warehouse? What are the characteristics and features of Data Warehouse?**

Data warehouse is an information system that contains historical and commutative data from single or multiple sources. It simplifies reporting and analysis process of the organization. It is also a single version of truth for any company for decision making and forecasting.

Characteristics of Data warehouse

- Subject-Oriented
- Integrated
- Time-variant
- Non-volatile

Subject-Oriented:

A data warehouse is subject oriented as it offers information regarding a theme instead of companies' on-going operations. These subjects can be sales, marketing, distributions, etc. A data warehouse never focuses on the on-going operations. Instead, it put emphasis on modelling and analysis of data for decision making. It also provides a simple and concise view around the specific subject by excluding data which not helpful to support the decision process.

Integrated:

In Data Warehouse, integration means the establishment of a common unit of measure for all similar data from the dissimilar database. The data also needs to be stored in the Data warehouse in common and universally acceptable manner. A data warehouse is developed by integrating data from varied sources like a mainframe, relational databases, flat files, etc. Moreover, it must keep consistent naming conventions, format, and coding. This integration helps in effective analysis of data. Consistency in naming conventions, attribute measures, encoding structure etc. has to be ensured.

Time-Variant

The time horizon for data warehouse is quite extensive compared with operational systems. The data collected in a data warehouse is recognized with a particular period and offers information from the historical point of view. It contains an element of time, explicitly or implicitly. One such place where Data warehouse data display time variance is in in the structure of the record key. Every primary key contained with the DW should have either implicitly or explicitly an element of time. Like the day, week month, etc. Another aspect of time variance is that once data is inserted in the warehouse, it can't be updated or changed.

Non-volatile

Data warehouse is also non-volatile means the previous data is not erased when new data is entered in it. Data is read-only and periodically refreshed. This also helps to analyze historical data and understand what & when happened. It does not require transaction process, recovery and concurrency control mechanisms. Activities like delete, update, and insert which are performed in an operational application environment are omitted in Data warehouse environment.

Only two types of data operations performed in the Data Warehousing are

1. Data loading
2. Data access

10. Explain in detail the building blocks of Data Warehouse.

The data warehouse is based on an RDBMS server which is a central information repository that is surrounded by some key components to make the entire environment functional, manageable and accessible.

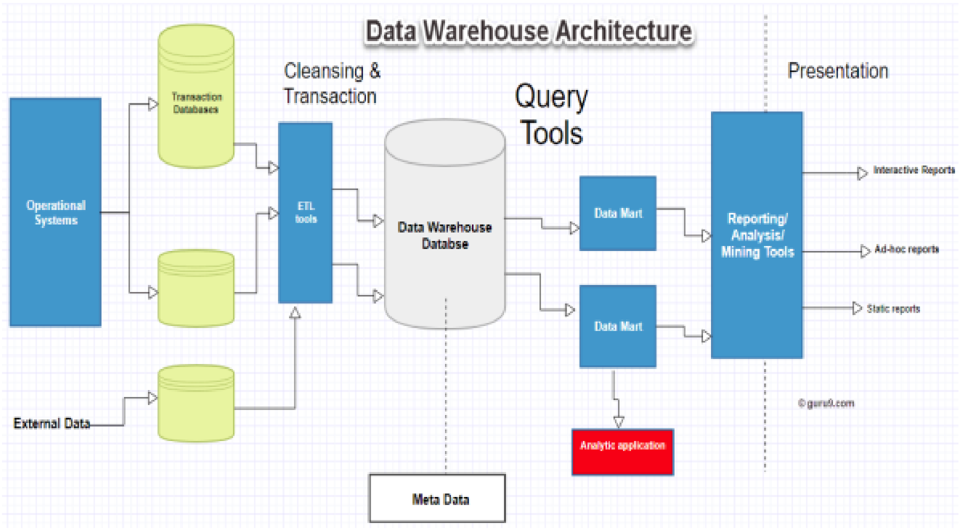


Figure 2.1 Data warehouse Components

There are mainly five components of Data Warehouse:

Data Warehouse Database: The central database is the foundation of the data warehousing environment. This database is implemented on the RDBMS technology. Although, this kind of implementation is constrained by the fact that traditional RDBMS system is optimized for transactional database processing and not for data warehousing. For instance, ad-hoc query, multi-table joins, aggregates are resource intensive and slow down performance.

Hence, alternative approaches to Database are used as listed below-

In a data warehouse, relational databases are deployed in parallel to allow for scalability.

- Parallel relational databases also allow shared memory or shared nothing model on various multiprocessor configurations or massively parallel processors.

- New index structures are used to bypass relational table scan and improve speed.
- Use of multidimensional database (MDDBs) to overcome any limitations which are placed because of the relational data model.

Example: Essbase from Oracle.

Sourcing, Acquisition, Clean-up and Transformation Tools (ETL)

The data sourcing, transformation, and migration tools are used for performing all the conversions, summarizations, and all the changes needed to transform data into a unified format in the data warehouse. They are also called Extract, Transform and Load (ETL) Tools. Their functionality includes:

Anonymize data as per regulatory stipulations.

- Eliminating unwanted data in operational databases from loading into Data warehouse.
- Search and replace common names and definitions for data arriving from different sources.
- Calculating summaries and derived data
- In case of missing data, populate them with defaults.
- De-duplicated repeated data arriving from multiple data sources.

These Extract, Transform, and Load tools may generate cron jobs, background jobs, Cobol programs, shell scripts, etc. that regularly update data in data warehouse. These tools are also helpful to maintain the Metadata. These ETL Tools have to deal with challenges of Database & Data heterogeneity.

Metadata

The name Meta Data suggests some high- level technological concept. However, it is quite simple. Metadata is data about data which defines the data warehouse. It is used for building, maintaining and managing the data warehouse. In the Data Warehouse Architecture, meta-data plays an important role as it specifies the source, usage, values, and features of data warehouse data. It also defines how data can be changed and processed. It is closely connected to the data warehouse.

Metadata helps to answer the following questions

- ➢ What tables, attributes, and keys does the Data Warehouse contain?
- ➢ Where did the data come from?
- ➢ How many times do data get reloaded?
- ➢ What transformations were applied with cleansing?

Metadata can be classified into following categories:

1. **Technical Meta Data:** This kind of Metadata contains information about warehouse which is used by Data warehouse designers and administrators.
2. **Business Meta Data:** This kind of Metadata contains detail that gives end-users a way easy to understand information stored in the data warehouse.

Query Tools:

One of the primary objects of data warehousing is to provide information to businesses to make strategic decisions. Query tools allow users to interact with the data warehouse system.

These tools fall into four different categories:

1. Query and reporting tools

2. Application Development tools

3. Data mining tools

4. OLAP tools

1. Query and reporting tools:

Query and reporting tools can be further divided into

- Reporting tools
- Managed query tool

Reporting tools: Reporting tools can be further divided into production reporting tools and desktop report writer.

1. Report writers: This kind of reporting tool is tools designed for end-users for their analysis.

2. Production reporting: This kind of tools allows organizations to generate regular operational reports. It also supports high volume batch jobs like printing and calculating. Some popular reporting tools are Brio, Business Objects, Oracle, Power Soft, SAS Institute.

2. Application development tools:

Sometimes built-in graphical and analytical tools do not satisfy the analytical needs of an organization. In such cases, custom reports are developed using Application development tools.

3. **Data mining tools:** Data mining is a process of discovering meaningful new correlation, pattens, and trends by mining large amount data. Data mining tools are used to make this process automatic.

OLAP tools: These tools are based on concepts of a multidimensional database. It allows users to analyse the data using elaborate and complex multidimensional views. Data warehouse Bus Architecture Data warehouse Bus determines the flow of data in your warehouse. The data flow in a data warehouse can be categorized as Inflow, Upflow, Downflow, Outflow and Meta flow. While designing a Data Bus, one needs to consider the shared dimensions, facts across data marts.

Data Marts A data mart is an access layer which is used to get data out to the users. It is presented as an option for large size data warehouse as it takes less time and money to build. However, there is no standard definition of a data mart is differing from person to person. In a simple word Data mart is a subsidiary of a data warehouse. The data mart is used for partition of data which is created for the specific group of users. Data marts could be created in the same database as the Data warehouse or a physically separate Database.