

DATA WAREHOUSING & DATA MINING

Ruchi Bathla

Assistant Professor

Computer Science & Application, Geeta Degree College, Shera

City – Panipat (Haryana)

Abstract: In the real world, there are huge amount of data store in the database which increase day by day. A Query comes in our mind “How I extract the knowledge from the huge amount of data?”. For this purpose, we used the concept of data warehouse and data mining. Data warehouse store those data, which is beneficial for us, from the database. Data in database leads to noise, incomplete and inconsistent. Firstly we make it smooth by removing the dirt using the data cleaning methods, integrate the data from the heterogeneous database and store in Data warehouse. Now Select the data from the data warehouse and transform it according to the target. Then make the patterns and extract the information from the data warehouse using the data mining technique.

Keyword: Data warehouse, Data Mining, Knowledge Discovery from Data (KDD), Data Pre – processing

1. Data warehousing

1.1 Introduction

In the real world, Most of processing for strategic information will have to be analytical. This new environment that users need to obtain strategic information happens to be new paradigms of Data warehousing. Enterprise that built the system environment, which perform day to operation, is also built the new environment, which is used to obtain strategic decision.

Sometimes the information cannot obtain from the database. Then we need data warehouse in which strategic decision can be take place from data warehouse. Data Warehouse created from the database using the data pre – processing technique.

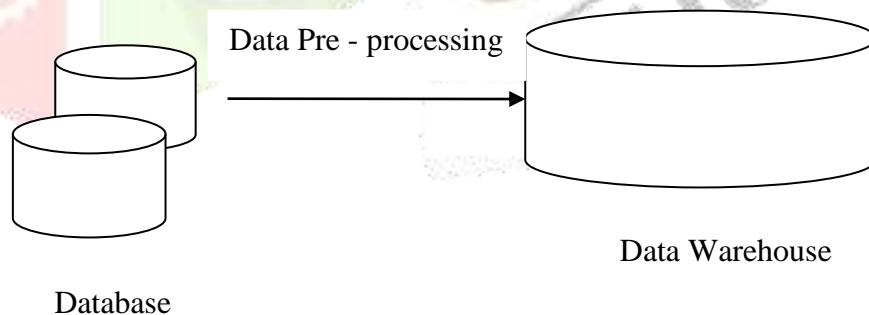


Figure 1

Data warehouse have been defined in many ways. It refers to a database that maintain separately from the organisational database.

Proper definition of data warehouse is “A data warehouse is a subject – oriented, integrated, and time – variant and non volatile collection of data that support to take the managerial decision”.

The main four feature of data warehouse is

- Subject – oriented

- Integrated
- Time – variant
- Non volatile

Subject – oriented: A data warehouse is organised only for major subjects, such as sale, customer, production rather than day to day operation. It focus only on the modelling a data in such a way to take the decision based on data warehouse.

Integrated: It is constructed by multiple heterogeneous sources. Data cleaning and data integration technique apply on the data to ensure consistency.

Time – Variant: Data are store to provide the information from the historical point of view.

Non Volatile: it also store separately from the operational database because it doesn't require transaction processing, recovery and concurrency control. It requires only two operations: loading the data & accessing the data.

Data warehouse is also known as **OLAP (On - line Analytical Processing)**.Data warehouse are based on a multidimensional Data Model.

1.2 Multidimensional data model

In this model, Data can be viewed in the form of data cube. A Data cube allows data that can be modelled and viewed in multiple dimensions. Dimension means different perspective with respect to which an organisation wants to keep records. **For example-** A 2D view of the production of different products in different branches in 2015 as follow

Table 1

Branch

| | American | London | Canada | Australia |
|---------------------------|----------|---------|---------|-----------|
| Products Sweets | £50,000 | £20,000 | £60,000 | £26,000 |
| Drinks | £5,000 | £20,000 | £80,000 | £28,000 |
| Crisps | £30,000 | £10,000 | £34,000 | £54,000 |
| Fruits | £25,000 | £40,000 | £70,000 | £5,000 |
| Vegetable | £32,000 | £80,000 | £50,000 | £35,000 |

In this figure, data was presented in the form of table having row and column. Row show Number of product produce in an organisation in 2015. Column show number of product produces in each location in 2015.

When we aggregate the data of 2015, 2016 and 2017 as follow.

| | | American | London | Canada | Australia |
|------|-----------|----------|---------|---------|-----------|
| Swe | | | | | |
| Drin | Sweet | | | | |
| Cris | Drinks | £50,000 | £20,000 | £60,000 | £26,000 |
| Fru | Crisps | £5,000 | £20,000 | £80,000 | £28,000 |
| Veg | Fruits | £30,000 | £10,000 | £34,000 | £54,000 |
| | Veget | £25,000 | £40,000 | £70,000 | £5,000 |
| | Vegetable | £32,000 | £80,000 | £50,000 | £35,000 |

After Aggregation

| | American | London | Canada | Australia |
|-----------|----------|----------|----------|-----------|
| Sweets | £250,000 | £420,000 | £260,000 | £426,000 |
| Drinks | £40,000 | £720,000 | £340,000 | £280,000 |
| Crisps | £690,000 | £510,000 | £304,000 | £504,000 |
| Fruits | £205,000 | £460,000 | £670,000 | £63,000 |
| Vegetable | £362,000 | £880,000 | £390,000 | £350,000 |

Figure 2

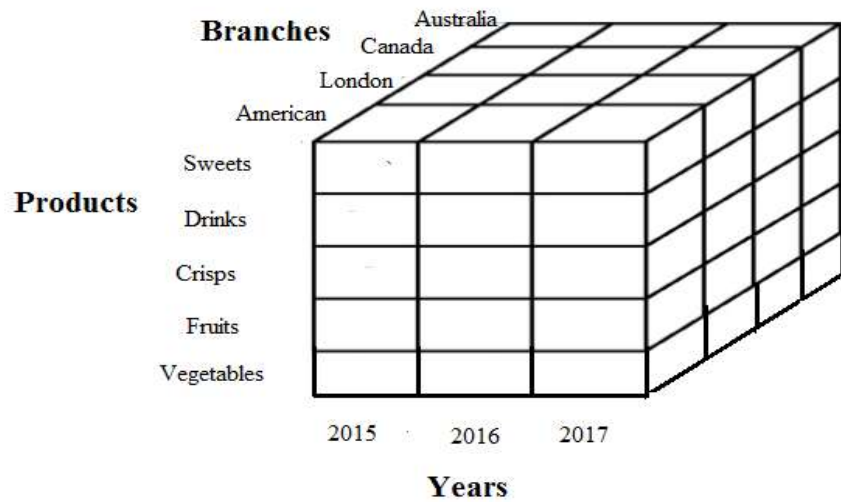


Figure 3

Figure 2 aggregate the production of 2015, 2016 and 2017 of an organisation. Figure 3 show a data cube of an organisation, which produce the different products in different branches, show three dimensions. The dimensions are products, Branches, Years

1.3 Measurement in Data warehousing

A data cube measure is numerical functions that can be calculating each point of the data cube space. A measured value is computed for a given point by collecting all the data.

Measures can be organised into the three methods

- **Distributive:** It can be computed in the distributive manners. Let the data are to be partition into n sets. We apply the function to each partition, get the n aggregate values. Now apply the function to the aggregate values. Example- Find the minimum value. Firstly apply the min () function to the n sub data cube. Then again apply min () function to all the measured values obtains in a sub data cubes.
- **Algebraic:** It can be computed by an algebraic function which is obtained by applying the distributive function. Example- Avg ()= Sum ()/ Count ().Here sum and Count are the distributive functions.
- **Holistic:** it is obtained by Holistic aggregate function. Example- Median, Mode, etc.

1.4 OLAP Operations

The following operations perform in Data Warehouse.

- Roll – Up (Dril – Up)

- Drill – Down
- Slice And Dice
- Pivot(Rotate)

Roll – Up: This operation performs on the data cube using climbing up concept hierarchy. The hierarchy of above data cube for Product Dimension is “Product < type of product” This operation show aggregate the data by ascending the Product hierarchy from the level product to the level of type of product.

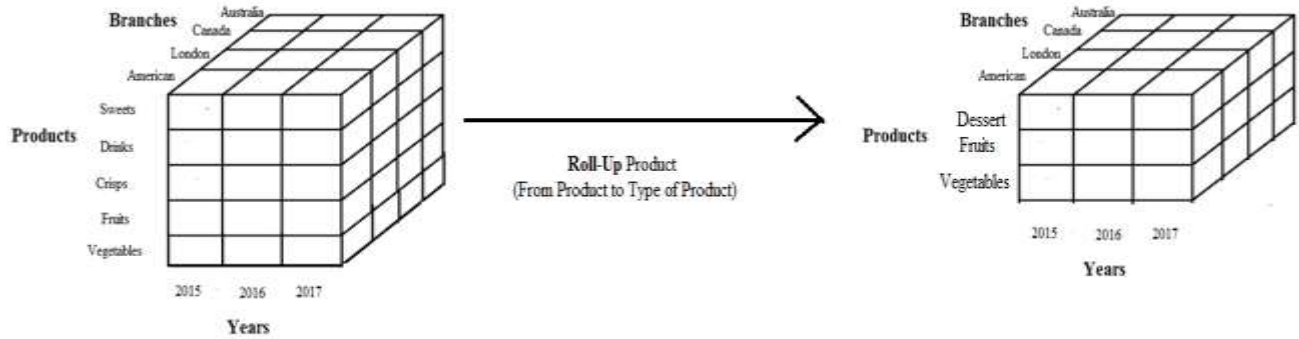


Figure 4

Drill – Down: It is the reverse of the Roll – Up. It converts less information to the more information by stepping down in the Concept Hierarchy. The concept Hierarchy for the Year is “month < quarter < half yearly < year”. This operation occurs by descending the year hierarchy from the level “year” to the level of “half yearly”.



Figure 5

Slice: This operation performs a selection on one dimension of the data cube to obtain a sub cube.
Example- Find the production of Sweets in an organisation.

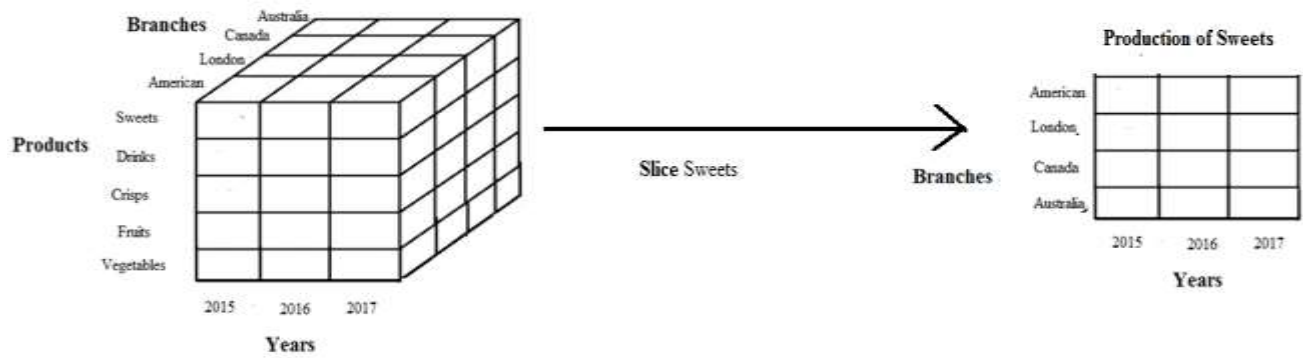


Figure 6

Dice: This operation perform selection of two or more dimension to obtain a sub cube. **Example-** Find the production of dessert in America, Australia location in 2016 and 2017 in an organisation.

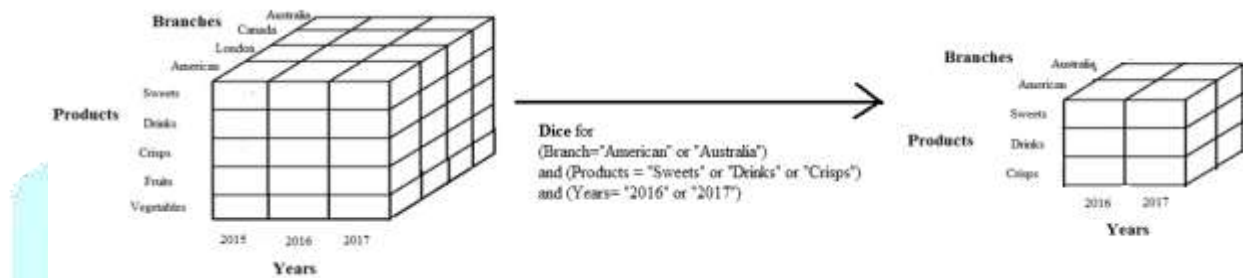
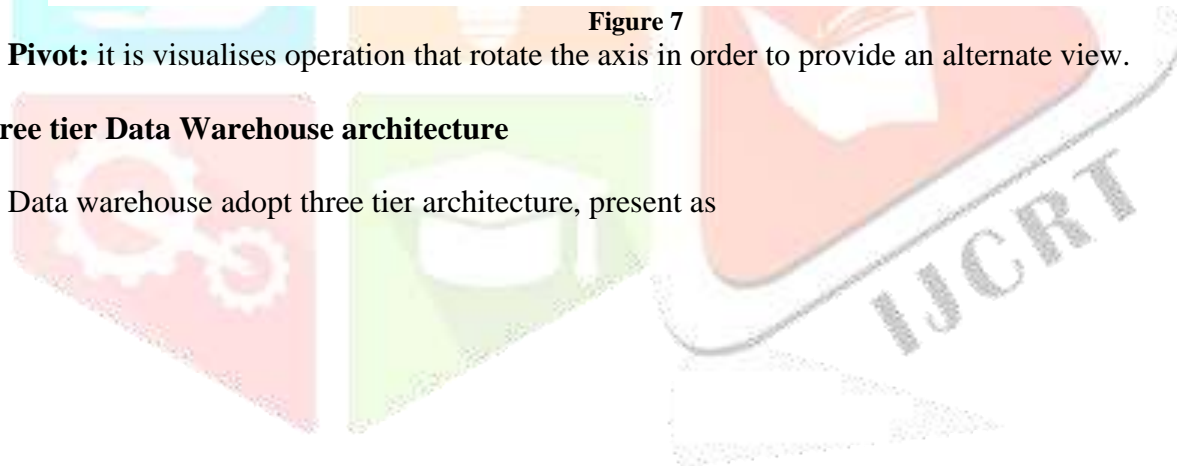


Figure 7

Pivot: it is visualises operation that rotate the axis in order to provide an alternate view.

1.5 Three tier Data Warehouse architecture

Data warehouse adopt three tier architecture, present as



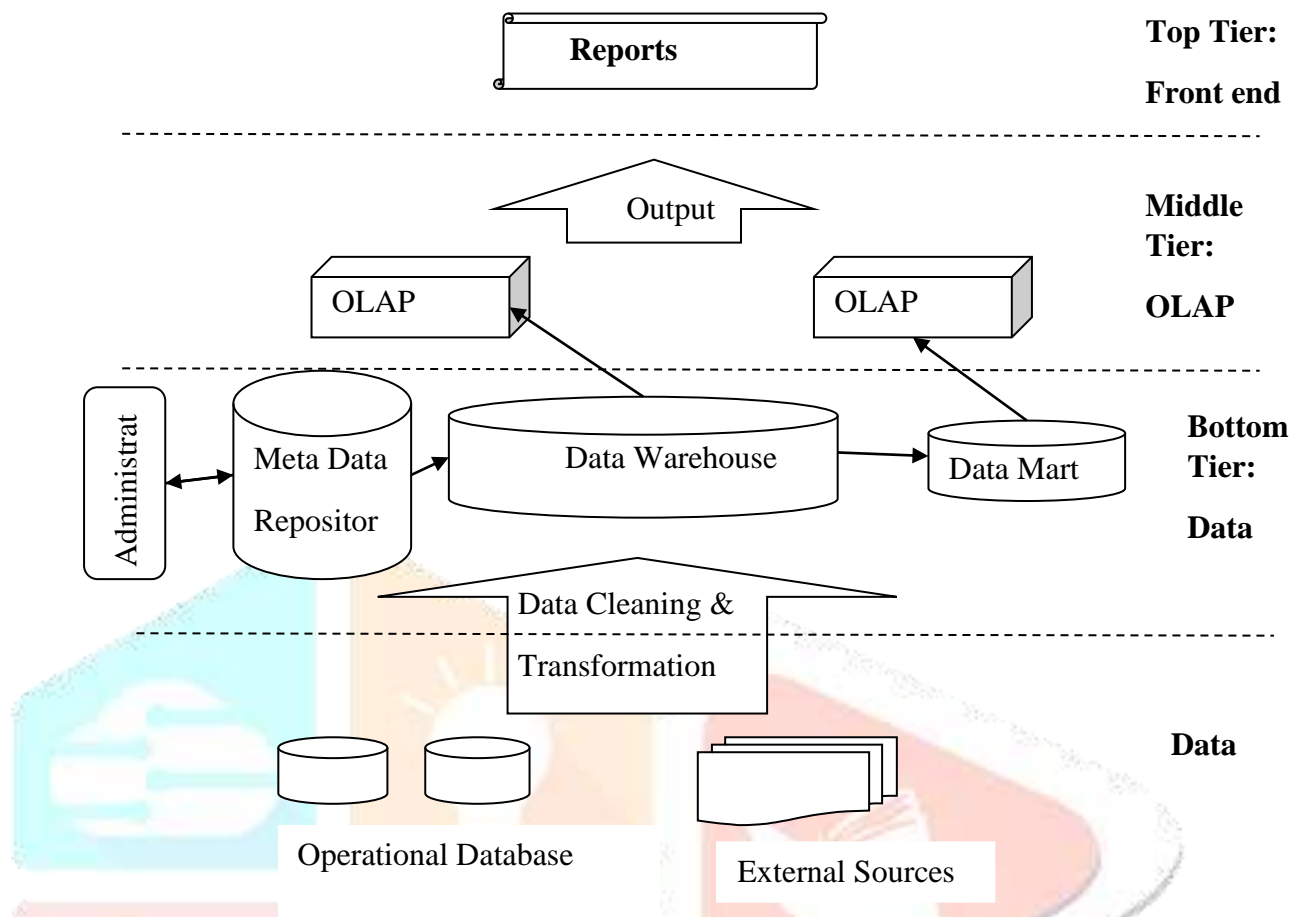


Figure 8

Bottom tier: Data warehouse Server

The data feed into the data warehouse using back – end tools that perform data cleaning, data integration techniques. This tier contains Meta data repository. **Meta data repository** store information about the data warehouse and its content also.

This tier also contains data marts. **Data mart** is a repository of data that is design to serve a specific group of users. Their design process start with to analysis of user needs.

Today, virtual data marts create using data virtualization software. This software pulls the data from different sources and combining it to meet the need of specific user.

Middle tier: OLAP server

It is implement using **ROLAP (Relational on - Line Analytical Processing)**, **MOLAP (Multidimensional OLAP)**. ROLAP is extended RDBMS that map the operations on multidimensional data to relational operations. MOLAP perform multidimensional operations.

Top tier: Front end layer

It contains report tools, analysis tools, Data Mining tools etc.

Data mining

2.1 Introduction

We have huge amount of data in the world and its amount increase day by day. There is billions of data increase per second. A Query comes in our mind – How we extract the knowledge from raw data? So we have to analyze the data and extract the knowledge which we need to get. This process is called **Data Mining**. Data mining is like mining the gold from Rocks and sands.

The meaning of Data Mining is “**Knowledge mining from data**”. There are lots of terms used for Data Mining like “Knowledge mining from data”, “Extracting Knowledge”, “Analyzing Data”, “Discovery of Knowledge”. Most of people treat Data Mining as Knowledge Discovery of Data (KDD)

Data mining is defined in many different ways. Some says- “**Data mining is the process of Discovery target pattern in huge amount of data and analysis the patterns to retrieve the knowledge**”. In other words, Data mining is a technique or a concept, in computer science, which deal with extracting useful and previously unknown information from raw data which is stored in big repository called database.

During Data Mining, Data miner used the functionality of Database Management System (DBMS) to extracting the knowledge from the pre-process raw data.

2.2 Why we need Data Mining?

We know that Data mining extract the knowledge from raw data which is store in the database. If we have a Query then we ask the query to the Database. But why we need Data Mining. Data Mining Reponses those query which can't predict by Database.

Consider the following Example as follow.

The number of accidents in Chandigarh as given below

Table 2

| Date | Location | No. of accidents | Weather |
|-----------------|-----------|------------------|---------|
| 1 July,2017 | Sector 22 | 2 | Sunny |
| 5September,2017 | Sector 18 | 5 | Windy |
| 9October,2017 | Sector 19 | 6 | Cloudy |
| 11December,2017 | Sector 17 | 3 | Sunny |
| 18December,2017 | Sector 15 | 2 | Rainy |
| 5February,2018 | Sector 26 | 1 | Sunny |
| 5March,2018 | Sector 28 | 3 | Windy |

Here we have given a database of number of accidents in Chandigarh at a particular date and also given what kind of weather is there at that time.

Suppose I want to know about how many accidents take place in Sector 17? We can easily predict using above database. But if we want to get the answer about the place that an accident takes place next week. We can't predict this type of query using the given database. We predict it using Data Mining techniques.

So we say that Data mining extract that knowledge which can't predict by Database and previously unknown information from raw data.

2.3 Knowledge Discovery from Data (KDD)

The term Knowledge discovery from Data (in short KDD) refers to the process of finding the knowledge from database. It involves the evaluation and interpretation of the patterns to get knowledge.

The main goal of KDD process is to extracting knowledge from data which store in a big repository. The step involve in Knowledge discovery from Data (KDD) as follow in diagram

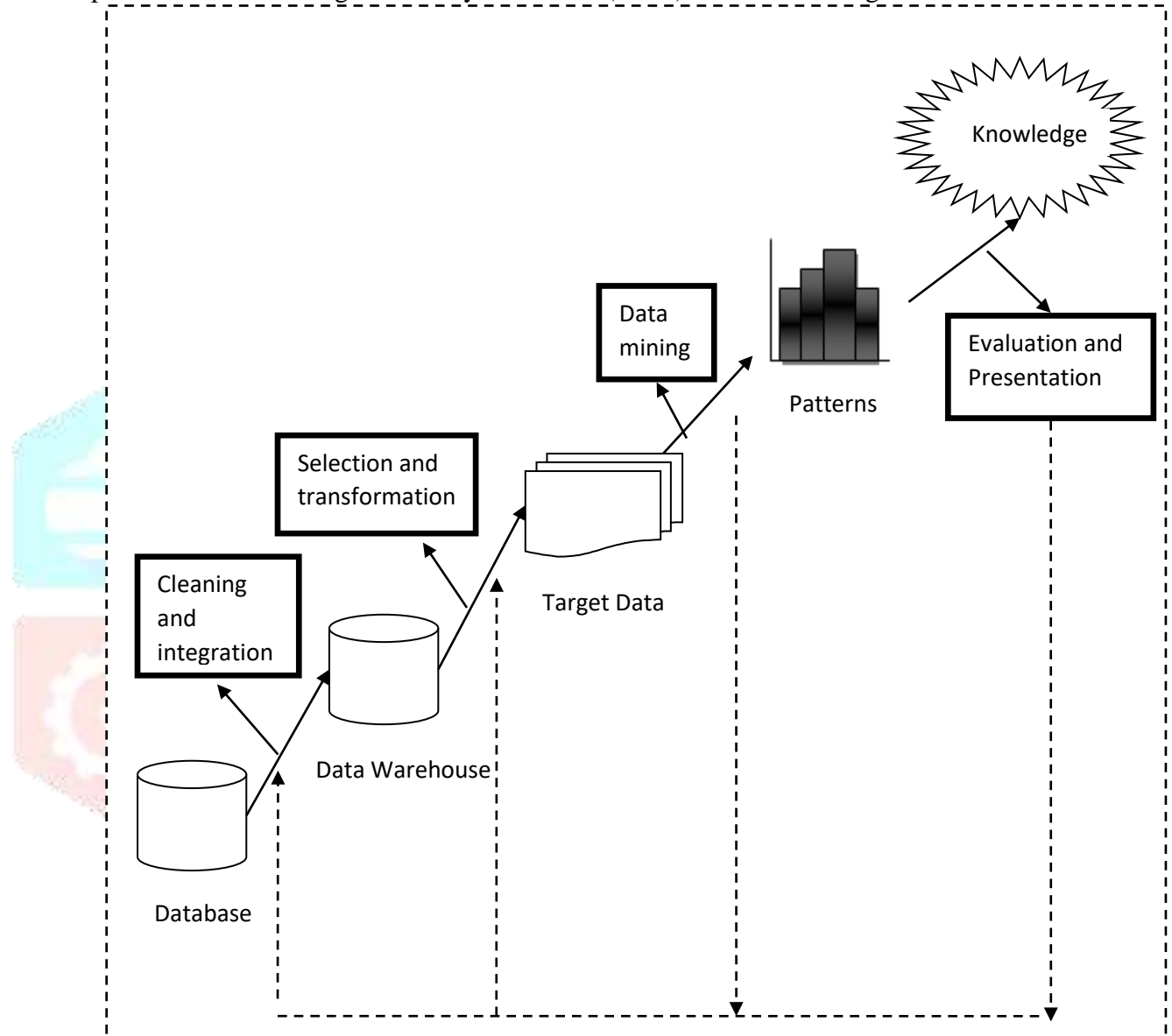


Figure 9

Steps of Knowledge Discovery from Data (KDD)

- Data pre-processing
 - Data Cleaning
 - Data integration
 - Data Transformation
 - Data selection
- Data Mining

➤ Data Evaluation & Presentation

2.4 Data Pre-processing

Data in real world lead to inconsistency, noisy, missing values.

- Noisy means it consist a lots of error.
- Inconsistent means it contain discrepancy in code
- Incomplete means data contain missing values

Major tasks involve in data pre processing are

- Data Cleaning (it fill the missing values, remove noisy data and resolve inconsistency)
- Data Integration (Integrate the data, which coming from multiple database, are store in data warehouse)
- Data Selection (Select those data which is relevant to analysis task from data Warehouse)
- Data transformation (Selected data transform in the appropriate form)

2.4.1 Data Cleaning

In real world, Data become incomplete, noisy and inconsistent. Data cleaning fill the missing values, smooth the data and resolve the inconsistency.

➤ **Missing Values**

In database, there are lots of missing values. Sometimes the tuple have no recorded values for several attributes. How do we fill the missing value? Let us study the following methods.

- **Ignore the tuple-** In this method; we ignore the tuple that doesn't have any value. This method is ineffective when there are several attribute which doesn't have any value.
- **Fill the missing values manually-** In this method, we fill the missing values manually.
- **Use global constant to fill the missing values-** we replace all the missing value by unknown or ∞ . If we fill the missing values by same constant and then data mining treat it as some interesting fact because it contain same value "unknown".
- **Use the central tendency (means mean, median) to fill the missing values-** we use mean for symmetrical distribution and median for skewed distribution.
- **Use the most probable values to fit the missing values-** to fill the missing values used the best tools like regression, decision tree etc.

➤ **Noisy data**

It means random error or variance in a data and also makes it smooth. The technique are-

- **Binning-** It smooth a sorted data by its neighbour which means surrounded values. The sorted data values are distribution into a number of buckets or bins.
Some binning techniques are
 - **Smoothing by mean-** values in bin are replaced by mean
Example- sorted data- 2, 4, 9, 12, 15, 18, 19, 21, 29

The sorted data are equal divided in a bin of size 3.

Bin 1- 2, 4, 9

Bin 2- 12, 15, 18

Bin 3- 19, 21, 29

Now replaced by mean

Bin 1- 5, 5, 5

Bin 2- 15, 15, 15

Bin 3- 23, 23, 23

- **Smoothering by median-** values in bin are replaced by median

Example- sorted data- 2, 4, 9, 12, 15, 18, 19, 21, 29

The sorted data are equal divided in a bin of size 3.

Bin 1- 2, 4, 9

Bin 2- 12, 15, 18

Bin 3- 19, 21, 29

Now replaced by median

Bin 1- 4, 4, 4

Bin 2- 15, 15, 15

Bin 3- 21, 21, 21

- **Smoothering by boundaries-** values in bin are replaced by neighbourhood minimum value.

Example- sorted data- 2, 4, 9, 12, 15, 18, 19, 21, 29

The sorted data are equal divided in a bin of size 3.

Bin 1- 2, 4, 9

Bin 2- 12, 15, 18

Bin 3- 19, 21, 29

Now replaced by mean

Bin 1- 2, 2, 9

Bin 2- 12, 12, 18

Bin 3- 19, 19, 29

- **Regression-** Here smothering of data can be done by regression. It includes linear regression, multiple regressions.
 - **Linear regression-** it finds the best line to fit two attribute so that one attribute predict other one. It can be express using the following linear function- $y = ax + b$ where y is response variable and x is predictor variable.
 - **Multiple regressions-** it finds the best among more than two attributes.
- **Clustering** – The process of grouping a set of objects into class of similar objects is known as clustering. Collection of similar objects is known as cluster and also collection of dissimilar object is also known as cluster. Clustering is also called as data segmentation because it partition large collection of data into small cluster which have similar objects. It can also used as outlier detector.

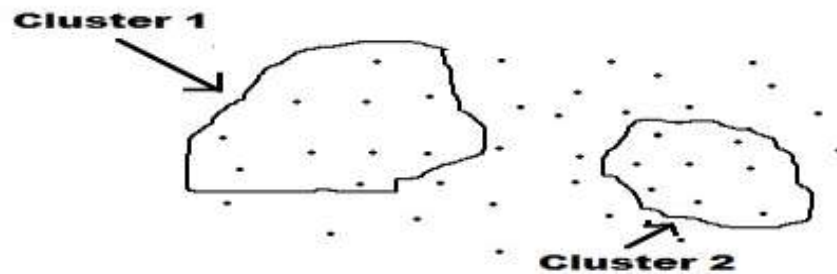


Figure 10

2.4.2 Data Integration

After cleaning the data, we integrate the heterogeneous data, which coming from multiple databases, in the data warehouse. In this way, new information merge with already exist information.

When we collect data from multiple databases then redundant data occur.

A number of issues created during data integration-

- **Entity identification problem-** some attribute may have different names in different databases. **Example-** Some database consider Employee ID as Emp-ID, Some other consider as Employee_ID.
- **Redundancy-** Sometimes an attribute can be derived attribute from another database. Inconsistency in attribute name may cause redundancy.
- **Detection and resolution of data value conflicts-** attribute value from different sources may differ due to difference in scaling , representation techniques etc. example- some treat height as centimetres but other treat as feet.

Redundancy can be detecting using correlation analysis. Correlation can be measure how two attributes, say x and y, are strongly related to each other. We can calculate correlation between X and Y attribute by computing correlation coefficient (also called Pearson's correlation coefficient). This is denoted by r which lies between -1 to 1.

Formula of correlation coefficient is-

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}}$$

Where x and y be the attributes

And n is the total no of combination between x and y

Table 3

| R | Correlation |
|----------------|-------------------|
| -1 | Perfect Negative |
| -0.99 to -0.51 | Strongly Negative |
| -0.5 | Moderate Negative |
| -0.49 to -0.01 | Weakly Negative |
| 0 | No |

| | |
|--------------|-------------------|
| 0.01 to 0.49 | Weakly Positive |
| 0.5 | Moderate Positive |
| 0.51 to 0.99 | Strongly Positive |
| 1 | Perfect Positive |

2.4.3 Data Selection

Data selection is the process where we retrieve only those data from data warehouse which is relevant to the analysis task. Here the selected data from the data warehouse are very small in size. It is a technique that can create the reduced representation of the data.

It include-

- Data cube aggregation
- Attribute subset selection
- Dimensionality reduction
- Numerosity reduction
- Discretization and concept hierarchy generation

2.4.3.1 Data cube aggregation

A Data cube allows data that can be modelled and viewed in multiple dimensions in an organisation. Dimension means different viewpoint with respect to which an organisation wants to keep records. **For example-** A 2D view of the production of different products in different branches in 2015 as follow

Table 4

| | | Branch | | | |
|----------|-----------|----------|---------|---------|-----------|
| | | American | London | Canada | Australia |
| Products | Sweets | £50,000 | £20,000 | £60,000 | £26,000 |
| | Drinks | £5,000 | £20,000 | £80,000 | £28,000 |
| | Crisps | £30,000 | £10,000 | £34,000 | £54,000 |
| | Fruits | £25,000 | £40,000 | £70,000 | £5,000 |
| | Vegetable | £32,000 | £80,000 | £50,000 | £35,000 |

Here data was presented in the form of table having row and column. When we aggregate the data of 2015, 2016, 2017 as follow

| | | American | London | Canada | Australia | |
|-------|--------|-----------|---------|---------|-----------|---------|
| Swe | | | | | | |
| Drin | Sweet | American | London | Canada | Australia | |
| Cris | Drinks | Sweets | £50,000 | £20,000 | £60,000 | £26,000 |
| Fru | Crisps | Drinks | £5,000 | £20,000 | £80,000 | £28,000 |
| Veget | Fruits | Crisps | £30,000 | £10,000 | £34,000 | £54,000 |
| | | Fruits | £25,000 | £40,000 | £70,000 | £5,000 |
| | | Vegetable | £32,000 | £80,000 | £50,000 | £35,000 |

After Aggregation

| | American | London | Canada | Australia |
|-----------|----------|----------|----------|-----------|
| Sweets | £250,000 | £420,000 | £260,000 | £426,000 |
| Drinks | £340,000 | £720,000 | £340,000 | £280,000 |
| Crisps | £690,000 | £510,000 | £304,000 | £504,000 |
| Fruits | £205,000 | £460,000 | £670,000 | £63,000 |
| Vegetable | £362,000 | £880,000 | £390,000 | £350,000 |

Figure 11

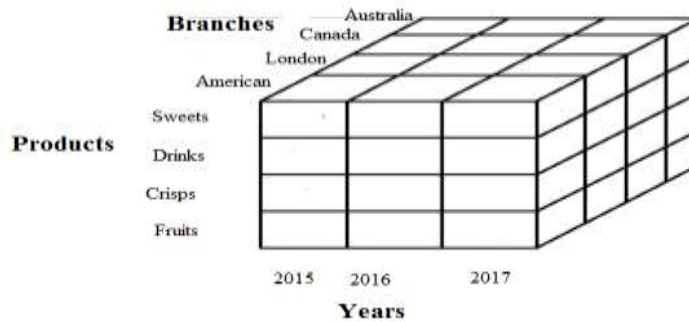


Figure 12

Data cubes store aggregate information. The figure shows the aggregate data values corresponding to each data point in multidimensional analysis of the productions.

2.4.3.2 Attribute Subset Selection

In data set, it contains lots of attribute. Some attributes is irrelevant to the mining tasks.

Using the attribute subset selection, we remove the irrelevant attributes using Forward selection, backward selection, Decision Tree

- **Forward Selection**

In this technique, the process starts with empty set. Now the relevant attribute add to the empty set. At each steps, the best of the remaining attribute added to the reduce set.

Example- Let the set contains {A1, A2, A3, A4, A5, A6, and A7}

According to this procedure,

- First we take empty set { }
- Now add the best attribute { A3 }
- Now again add the best attribute among remaining attribute {A3, A5 }
- Repeat this steps and now the reduced set will be {A3, A5, A7 }
- Now there is no best attributes.

So Reduced set is {A3, A5, A7}

- **Backward Selection**

In this technique, the process starts with complete set and removes the irrelevant and worst data. At the end, we get best reduced set.

Example- let the set contains {A1, A2, A3, A4, A5, A6, and A7}

According to this procedure,

- First we take whole set { A1, A2, A3, A4, A5, A6, and A7 }
- Remove the irrelevant attribute { A1, A2 }and now the reduced set will be { A3, A4, A5, A6, andA7 }
- Again remove the worst attribute among reduced set {A4, A6} and now the reduced set will be{A3,A5, A7 }

So Reduced set is {A3, A5, A7 }

2.4.3.3 Decision Tree

Decision tree is a predictive modelling approach. In this tree structure,

- Leaves represent class labels
- Branches represent conjunctions that show the way to class labels.

A decision tree can be used to represent decision and decision making.

2.4.3.4 Dimensionality Reduction

In this reduction, Data transformation is apply in such a way to obtained a reduced representation of the original data. It include-

- **Loseless reduction-** when original data reconstructed using the compressed data without loss of any information
- **Lossy reduction-** when we reconstruct the aproximate of original data.

2.4.3.5 Numerosity Reduction

When we reduce the volume of data by choosing the smaller form of data representation, we use numerosity reduction technique.

These technique used two method

- **Parametic method-** In this method, we store parameter instead actual data. Example- regression
- **Non parametric method-** it store reduced representation of data.example- histograms, clustering, sampling.

2.4.3.6 Data discretization and concept hirrarchy generation

This technique can be used to reduced the attribute in the intervals. The actual data is replaced by the intervals.this leads to a concise, easy to use, easy to understand. This technique based on top down and bottom-up approach.

It include-

- Top – down discretization
- Bottom – up discretization

In **top down splitting**, firstly find the split point to split the entire rangeand then repeat this recursively.

In **bottom – up merging**, this process start by all the value. Then merge the neighbourhood values and this process repeat recursively.

2.4.4 Data transformation

In data transformation, the data are transform in the appropriate form for mining.

It involve-

- Smoothening
- Aggregation
- Generalization
- Attribute construction

In **smoothening**, we remove noisy data using Binning, regression, clustering

In **aggregation**, we summarise data using the data cube aggregation

In **Generalization**, the low level data are replaced by higher level concept using concept hierarchy.

In **attribute construction**, new attributes are constructed from the given attributes to help the mining process.

After the Data Pre – processing, We find the frequent patterns using a data mining tool. Frequent patterns are those patterns that occur frequently. There are many kinds of frequent patterns such as frequent itemset, sequential patterns, structure patterns etc. A frequent itemset refers to a set of items that frequently appear together in a data set.

A sequential pattern refers to the pattern that occurs in sequence. A structure pattern refers to the patterns in which the structure occurs frequently. After creating the patterns, we extract the knowledge based on patterns.

2.5 Conclusion

A huge amount of data is generated through the internet. Online transaction in financial market is also generating a huge amount of data. Unlike traditional data flow in and out of a computer system continuously varying day by day. It may be impossible to store on entire data. To discover knowledge or patterns from data stream, it is necessary to develop mining methods.

For this purpose, we construct a data warehouse that stores only interested data. Due to the enormous size of data streams, it is not possible to analyze each item. So we use data mining techniques like clustering, binning, data cube aggregation etc. we create the patterns using data mining techniques. To reduce the patterns, we use the clustering method. Along with frequent patterns is important as they play the major role in intrusion detection systems.

2.5 References

- 1) Jiawei Han and Micheline Kamber.” *Data Mining Concept and Technology*”, 2nd edition, Published by Morgan Kaufmann in 500 Sansome Street, Suite 400, San Francisco, CA94111, 2006; Pages 3-150
- 2) Alex Berson, Stephen J. Smith “*Data Warehousing, Data Mining & OLAP*”, ed. 2008; Published by Tata McGraw – Hill Publishing Company Limited, & West Patel Nagar, New Delhi 110053; Pages 115-126
- 3) Pieter Adriaans and Dolf Zantinge, ”*Data Mining*”; Published by Dorling Kindersley (India) Pvt. Ltd. Licensees of Pearson Education in south Asia; 2007; pages 37-90
- 4) Paolo Giudici and Silvia Figini, “*Applied Data Mining for business and industry*” 2nd edition; Published by John Wiley & Sons Ltd., The Atrium, Southern Gate, Chichester, West Sussex, PO19 85Q, United Kingdom; 2009; Pages 47-70
- 5) Gleen J. Myatt and Wayne P. Johnson, “*Making Sense of Data*”, 2nd Edition; Published by John Wiley & Sons Ltd., The Atrium, Southern Gate, Chichester, West Sussex, PO19 85Q, United Kingdom; 2007; Pages 2.1-5.99
- 6) S. Nagabhushana, ”*Data Warehousing OLAP and Data Mining*”, Volume 1; New age International; 2006; pages 24-35