16/08/19.

DATA WAREHOUSE

Data Sources ↳ To extract knowledge from the data based on the analysis.

ETL → Load
↓ └→ Transfer
Extract

→ Data sources are heterogenous (∵ different representations). (depends on developers)

Therefore, all application programs maynot extract same data.

→ Therefore, all data should be transformed into some standard form. (accepted by datawarehouse / datamining system) (some time)

Extract → extracting data from different sources.

Transfer → transformed into standard form

Load → loading into data warehouse system

Pre-processing: (Takes more data)

Includes attributes reduction, data reduction

Cleaning of data.

(If age is not entered, age can be calculated from DOB)

→ Identify outlayers in the raw data (noise data)

(Age = -5 etc)

Data Warehouse:

→ To design datawarehouses, data warehouse models are to be used.

→ It is ~~represe~~ implemented using Advanced SQL, DML.

Scanned by CamScanner

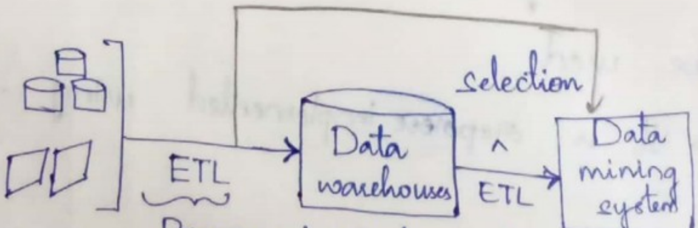
	DBMS	DW
Data:	Current Up-to-date	Historical data
Data Size:	MB-GB	Minimum: TB ≥TB
19/08/19 Operation Accessing	Read and write	Read only
No of Users:	Thousands of users End users, DataBase A(DBA) other Administrators	Hundreds of users Data Analyst Domain Experts Business Analysts. (BA)
Period of data:	Short-term Day-to-day operations	Long-term Data analysis/ Information processing
Models to Design:	ER / Relational Models	Star / Snowflake (Max. commercial datawareho
Performance on:	Transaction throughput	Response Time
	Thousands of records	Millions of records.
Representation:	2-Dimensional/ (Simple)	(High)/Multi-dimensional form (Time, Location, items etc gives co
	Detailed data	Summarized data

Knowledge Discovery Process from Data (KDD):

Data sources:

S1: Set of databases

S2: Data files



Preprocessing (includes cleaning)

Some times, transfer include integration also. Since, data set may include more than one database.

Loading ⇒ write into datawarehouse system.

23/08/19

→ Extracting data cannotbe done by everyone.

Data can be extracted. through own applications (API).

Data sources may be heterogenous.

→ Data should be transformed into acceptable form.

Integration is done (major task)

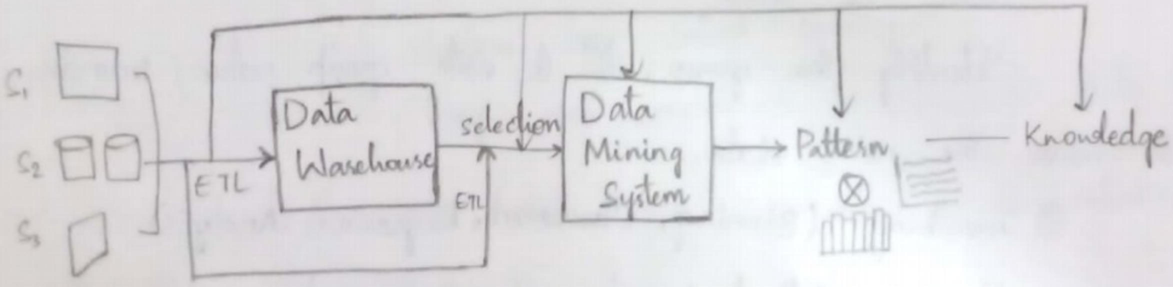
→ Data should be loaded into data warehouse. It has various models. Data should be converted considering the model types.

→ ETL is done for data warehouse → data mining.

Integration is not needed here.

Based on the mining task, the ETL processing should be made. Selection is made from the data warehouse data.

Sometimes, data warehouse is not needed. Data can be extracted from individual data sources and can directly be sent to data mining. Historical data is not stored. (or) individual sources may contain data.



→ The pattern obtained depend on the quality of data.

→ If the patterns obtained is not correct, the obtained data is noise data.

Static : If data is taken once and extracted.

Dynamic: Extracted and incremented.

Scanned by CamScanner

26/08/19

→ 60-80% of total time is spent in "pre-processing"

Data pre-processing

Methods:

① Data cleaning / cleansing

② Data Integration

③ Data Transformation

→ If data is incomplete (some data missing) and noisey data, it has to be cleaned. The data values which are important may be missing.

→ To solve data with missing values,

① Remove / Delete that entry (if Large database and few tuples have more missing values).

② Fill with column mean (if tuples have less missing values)

③ Fill with global constant (if highly influencing, change global constant

④ Fill with maximum and minimum value (boundary)

⑤ If categorical data, then group and analyse what must be filled.

Identify the group, fill it with group value / boundary.

→ To solve the noisy data,

① Smoothing (Binning, Clustered, Regression Analysis)

Binning Method: (works column by data)

Arrange attribute by attribute and ~~Bins~~. divide into bins

Eg:- A = {10 45 5 90 25 75 85 1 15 99 20 30 40}

Sorting / arranging:

A = {1 5 $\underset{B_1}{10}$ 20 | 25 $\underset{B_2}{30}$ 40 | $\underset{B_3}{75}$ 85 90 99 | ~~99~~ $_{B_4}$}

Divide into bins (size depends on the input size)

Size of bins must be same except one or two.

→ After dividing into bins, replace the values with the nearest boundary value. (Nearest - decided based on difference).

$$\left|\, \cancel{1} \quad \overset{1}{\cancel{5}} \quad \overset{1}{\cancel{10}} \quad 20 \,\right| \; 25 \quad \overset{25}{\cancel{30}} \quad \overset{45}{\cancel{40}} \quad 45 \,\Big|\; 75 \quad \overset{75}{\cancel{85}} \quad \overset{99}{\cancel{90}} \quad 99 \,\Big|$$

→ The values can also be replaced with mean of the bin ~~boundary~~ value

$$|\,9 \quad 9 \quad 9 \quad 9\,|\; 35 \quad 35 \quad 35 \quad 35\,|\; 88 \quad 88 \quad 88 \quad 88$$

→ This is smoothing the data set.

→ Data mining methods can be used to pre-processing the data.

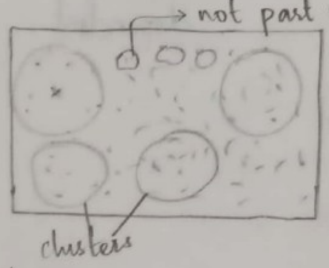
Eg:- Association based algorithm.

Grouping data items (clusters) based on some criteria.

30/06/19.

Pre-processing: (is done to improve quality of data)

- Data cleaning
  - Missing values
  - Noisy Data → Regression Analysis.
    - Binning method
    - Data clustering. (grouping of data based on some similarities).

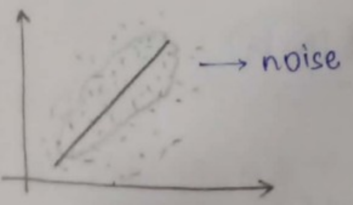


→ The items which are not part of any clusters are called "outlayers".

They are considered to be noise data and may be removed.

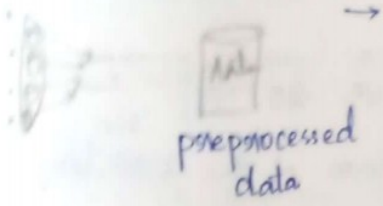
→ After removing noise data, good patterns can be obtained through data mining.

Regression Analysis: To identify noise data.



Scanned by CamScanner

Data Integration:



→ Data integration is difficult because diff data sources have different representation of data.

→ During integration (if directly combined), data repitition happ

Redundancy should be removed.

→ We must be able to identify similar attributes (may not h same name)

→ Co-relation functions may be used to determine similarity.

Attributes: A, B

$$\boxed{r_{A,B} = \frac{\sum_{i=1}^{N}(a_i-\bar{A})(b_i-\bar{B})}{N\sigma_A\sigma_B}}$$

a_i/b_i – ith attribute in A/B

N – No. of tuples in the dataset

σ_A – Standard Deviation of A.

Either A or B needed. ← $r_{A,B}>0$ ⇒ +vely co-related ⇒ Redundant/dependent.

$r_{A,B}=0$ ⇒ independent

$r_{A,B}<0$ ⇒ -vely co-related

↳ both A, B are needed

$\bar{A}, \bar{B}$ – mean values.

$$\boxed{\chi^2 = \frac{\sum_{i=1}^{c}\sum_{j=1}^{r}(\underbrace{o_{ij}}_{\text{observed}} - \underbrace{e_{ij}}_{\text{expected}})}{e_{ij}}}$$

$$e_{ij} = \frac{count(A=a_i) + count(B=b_j)}{N}$$

Data Transformation:

→ Source data is transformed into destination format.

↳ Data smoothing

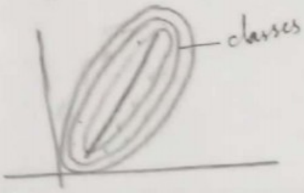
Dataset = {10 90 25 80. 75 60 5 15 40 55 45 35}
(N)

5 10 15 25 | 35 40 45 55 | 60 75 80 90

13 | 43.75

Binning method. ~~{13 13 13 13 43 43 43 43 75 75 75 75}~~
(replacing

5 5 5 25 35 35 55 55 60 60 90 90



Data Generalization:

→ If data represented in heirarchial form; highest level - generalization low level - specification.

Y M C C C Y S S S S
Eg: D = {20, 30, 10, 5, 15, 25, 75, 60, 45, 55, 85}
Age of customer

Child - (0-15)

Young (15-25)

Middle (25-45)

Senior (>45)

06/07/19

Pre-processing

(i) Data Cleaning
  ↳ Missing Value
  ↳ Noisy Data

(ii) Data Integration

(iii) Data transformation

Scanned by CamScanner

Data Transformation:

Eg: Students score are represented as

(i) (0-100) Marks – Source database

(ii) (0-10) CGPA – Dest. database

Here, data transformation is required.

→ Methods of transforming data:

(i) Normalization

(ii) Smoothing

(iii) Generalization

Normalization techniques:

(i) Min-max techniques

Let given 'v'

$$\boxed{v' = \frac{v - min_A}{max_A - min_A}(newmax_A - newmin_A) + newmin_A}$$

Eg: 1. Marks given = 70 ; Convert into CGPA

$\Rightarrow v = 70 \quad v' = \frac{70-0}{100-0}(10-0) + 0$

$$v' = \frac{7\cancel{0}}{10\cancel{0}}(1\cancel{0}) = 7$$

$$\boxed{v' = 7}$$

2. $D = \{4, 7, 8, \textcircled{5}, 2, 10\}$ – given min = 0
max = 10

Convert into range of 0-1

$$v' = \frac{5-0}{10-0}(1-0) + 0 = 0.5$$

(ii) Z - score normalisation:

Given v,

$$\boxed{v' = \frac{v - \bar{A}}{\sigma_A}}$$

$\bar{A}$ - Mean

σ_A - Standard Deviation

→ The distibution of data is also considered.

Scanned by CamScanner

(iii) Decimal point:

$\boxed{V' = \frac{V}{10^j}}$ j-depends on Data Base

j- power of 10 which is closest to max. value in the dataset.

If max. absolute value = 999

$\Rightarrow j = 3$

Converting 285 is $V' = \frac{285}{1000} = 0.285 = V'$

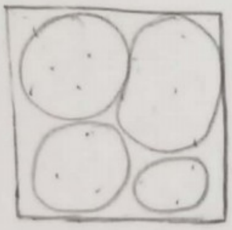
Smoothing Techniques:

(i) Binning Technique:

Eg:

	5	18	45
	3	12	30
	2	9	25
Replace with mean	3.3	13	33
	3.3	13	33
	3.3	13	33

(ii) Clustering:



No.of clusters is decided by user.

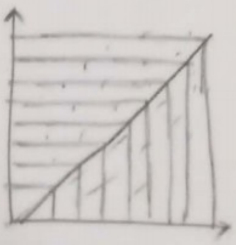
→ Every element is part of ~~atte~~ one cluster.

→ Elements of different clusters are dissimilar.

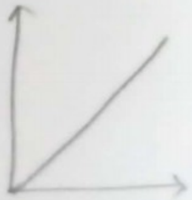
→ Elements of a cluster are replaced with the mean of cluster.

→ Clusters may be of different size.

(iii) Regression



All random values are replaced with the nearest regression value and the resultant graph is linear.



Scanned by CamScanner

Scanned by CamScanner

Generalization Technique:

Low level values are replaced with the high level values.

All values are divide into some groups based on some property and are replaced with the generalized group values.

Eg:- Source database contains ages

Destination database contains agegroups

Now, ages are grouped (generalized) and changed to required group value.

Scanned by CamScanner

Generalization Technique:

Low level values are replaced with the high level values.

-All values are divide into some groups based on some property and are replaced with the generalized group values.

Eg:- Source database contains ages

Destination database contains agegroups

Now, ages are grouped (generalized) and changed to required group value.

07/08/19.

Data Aggregation:

→ Aggregate level values are used for analysis. (better than individual values).

Eg:- Year-wise sales details.

Q1	210
Q2	320
Q3	130
Q4	250

2018

Q1	130
Q2	270
Q3	235
Q4	410

2017

Q1	260
Q2	350
Q3	450
Q4	230

2016

→

Year	Total
2018	910
2017	
2016	

Source Data / Raw Data

→ Quarterly-wise data is not needed, so data is aggregated and transformed as yearly-wise.

Attribute Construction:

$D = \{A_1, A_2, A_3 \ldots A_n\}$

$New = \{B_1, B_2 \ldots B_k\}$

The values of new-attributes are derived from old.

$D' = \{A_1, A_2 \ldots A_n, B_1, B_2 \ldots B_n\}$ → New

Newly constructed attributes are added to the original dataset

Eg:- Student report card.

Initial attributes - marks obtained in every subject.

Avg. grade or percentage - are newly constructed attributes are added to the dataset.

DATA REDUCTION:

→ Dataset size should be reduced.

Methods to reduce size:

(i) Attribute selection → Forward Selection
→ Backward Selection

(ii) Decision Tree Based Algorithm.

Eg:- While analysing student's performance, contact number (or) home address doesnot effect the patterns obtained.

Therefore, they can be removed (reduced).

Attribute Selection:

Forward Selection:

Initially the interesting set of attributes is empty.

$S = \{ \}$ $D = \{A_1, A_2 \ldots \ldots A_n\}$

All the attributes are scannedand high priority attributes are added to the interested set.

Domain Expert will assign weights to all the attributes

(Weight ↑ Priority ↓)

Based on priority, attributes are added to the required set.

Backward Selection:

Initially interesting set = $\{A_1, A_2 \ldots \ldots A_n\}$ - all attributes

The attributes with low priority are discarded from the dataset. (Reverse of forward Selection).

→ Combination of both methods - Use a Hash Table.

In same iteration find max priority and min. priority

Add max. priority to hash table and discard the min. priority

Decision Tree Based Algorithm:

↓

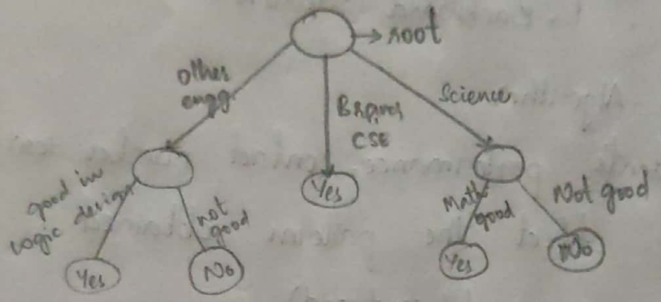
like flow chart

Tree - set of nodes.

Root → starting node

Leaf → Testing condition

Non-leaf → class labels



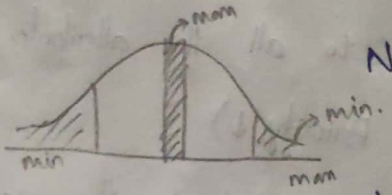
13/09/19 Random Sampling:

→ Some portion of data is taken and analysed and the patterns are obtained.

→ Selection is done based on sample (random).

Random sampling can be done in a variety of ways

Min. and max. of random numbers is selected.



Normal dist. may be selected

Avg. value numbers are more

Values with min. and max. values are less

Random sampling can be

1. With replacement
2. Without replacement

1. With replacement: If a number is selected, it has a chance to be selected again (since, the number is replaced into the original choices also).

→ Repetitions may be possible.

Sample: {4, 9, 1, 6, 8, 4} data is repeated

A conclusion can be made if large no.of times. Conclusion on behaviour can be made.

2) Without replacement:

No repetition

→ This approach can be applied to 2D and 3D data also. Many iterations can be performed. Each iteration some data is analysed.

If similar type of data analysed, percentage of accuracy is less. If random data selected, percentage of accuracy is more.

Data smoothing (i) Binning method:

→ Used for (i) cleaning of noisy data

(ii) data transformation

(iii) Data reduction

Suppose, bin size = 4

16 21 70 3 25 75 35 5 85 10 12 65

(following bin average technique)

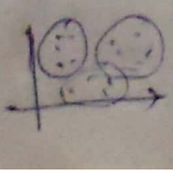
12	7.5	35	24	85	75
10	7.5	25	24	75	75
5	7.5	21	24	70	75
3	7.5	16	24	65	75
B1		B2		B3	

Large dataset is reduced into 3 different classes (7.5, 24, 75)

La Analysing this data is easy

(ii) Clustering technique:

Data is grouped into clusters and all the data in a cluster is replaced with cluster average (depends)



~~only~~ Lot of data is reduced.

No. of clusters is decided by user.

(iii) Regression Analysis:

- All data is replaced with the nearest regre value.
So, data reduction takes place.

20/09/19

Data discretization:

→ Divide into ranges

→ Classification - accept categorial attributes

→ Reduce size

→ Analysis

Interval labels replace actual data values

Supervised - discretization using class info.

unsupervised — no class info.

Split vs Merge
(top-down) (bottom-up)

Recursive discretization — collect and reduce.

Concept Heirarchy formation:

Eg:- Location data of org.

Same region branches clubbed together.

Methods used: (can be used recursively)

↳ Binning: Top-down, unsupervised

Bins are formed <u>without</u> any class labels and replace with bin boundaries.

Histogram Analysis: Grouped data

Clustering Analysis: Both top-down (split) & bottom-up (merge) can be use

also unsupervised.

Entropy based discretization - supervised, top-down (split

Interval merging - χ^2 analysis

Entropy: Interval splitting

→ Each value of A can be considered as a potential interval boundary or split point to partition the range of A.

→ Based on value before A and after A, the partition can be made.

→ Recursively applying them with threshold on intervals and partitions reached.

Interval Merging:

Based on χ^2 values, each pair of distinct intervals are tested and merged if χ^2 are low because low implies lightly related.

→ A data warehouse is a subject - oriented
integrated
time - variant
non - volatile

Collection of data in support of management's decision
making process.

· Data warehouse on subjects like sales, product, customer.

n-D base cube is called base cuboid

Apex cuboid : 0-D cuboid
(highly summarized [illegible] data)

Lattice of cuboid from data cube