## Data preprocessing & ETL

Douta preprocessing: - It is adulta mining technique that transforms raw days into a more understandable suseful efficient format. Real world days generally incomplete, Moisy, Inconsist

Datacleaning: - It is cuso known as scrubbing. This task involves filling of missing values, smoothing or remain hoisy data

Reasons for dirty or unclean duta.

- Dummy values
- 2) Absence of data
- 3) violation of business rules
- 4) contradictory duta
- 5) Reused primary Heys.

How to Clean duta?

- 1) Handle missing values
- 2) Handle noisy duta
- 3) Remove unwanted deuta.

D Handle missing values: - There are differently courte problem domain & good of data mining process.

max. amount of data is missing.

It is avoided where less attribute will

are missing b) Fill missing values manually! 2+ is very time consuming method & infecisible for almost au scenarios. c) use global constant to fillin for missing values: - It is used when missing values are difficult to predict. D'Handle Moisy delta - It may be introduce due to faut in daya collection error during data entering. roise can be bandled using binning. 3) Remove up wanted duta - Itisa duplicate or irrelevant duta due to redundant records, the model may not provide accurate results as the dupliceve duta interferes with analysis mocess. Data Integration: - In this Step a coherent duta source is prepared. mis is don-e by collecting & integrating data from mutiple sources like dayabases, data cubes etc. Issues in data Integration a) schema Integration - schema from diff sources may not be compatible. This leads to entity identification problem. D) Data value conflicts - The values or metrics or representations of the same dara maybe diffin for the same

Date Page

recu coored entity in diffidate sources

c) Redundant data: Dupicate attribute

or tuples may occur as a result of
integrating data from various sources

Data Reduction - IP the data is sent large, data reduction is performed Data reduction techniques are Data at a cube aggregation - Data is reduced by applying of P operations like sire, dice or rollup.

D) Dimensionality reduction - Data atting or dimensions are reduced.

C) Data compression - large volum of data is compression - large volum of data is compression to a using lossy or lossless compression d) numerosity reduction - It can be done using bistograms, clustering of Sampling of data.

Discretization in data mining! - It refers
to a method, of converting a huge number
of data values into small er ones so
that the evaluation & management of
data become easy. There are two
forms of data discretization first is
supervised & second is unserpervised
in which the class day is used



in discretization - It is a method in which operation moreeds. 6.0 attribute - Age. before discretization: Age - 1,5,7,9,11,16,17,18,20,21,25,28 After Discretization-Attribute Age Age Age 1,5,7,9 17,16,17,18,20 21,25,28 After Discretization child young mature. Concept hierachy generation- It is an organizational structure mapping is done from low-level concepts to high level concepts. There are too types of hierachy is Top-down ii) bottom-Up. i) Top-docon - It generally starts with top with some general info & ends with bottom to specialized info. 11) Bottom- up: - It generally starts with bottom with some specialized info & ends with top to the generalized info. Delta transformation - This involves normalization & aggregation of data according to the needs of data set. = involves the following Smoothing - It can world to remove

maise from the duta.

Aggregation - This phase is generally used in making a data cube for the conalysis of the data at multiple of Generalization - where low level duto are restored by larger - level concertion as the condition - where the attributed as a cre scaled to fau within a specified range.

D)

ETL: - It is a process that extracts.

data from diff. source system then

transforms the data & finally loadst

data into the data coarehouse system

ETL means Extract, Transform & load

Extraction! - Data is extracted from

source system for further use ind

data warehouse. It is a time consumit

process.

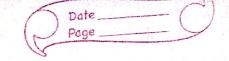
Extraction method!

a) Logical Eservaction 6) Physical Fall

a) Logical Eletraction! - It is divided in two kinds.

i) Full Exetraction ii) incremental

i) Full Exerraction. Data is extracted completely from source system.



there is no need to keep truck of Change to the data source since the last successful Exetraction:

Incremental Exetraction: - At a specific

incremental Exercaction: - At a specific point in time, only the data that has changed since a well-defined event back in history will be exercacted.

b) physical extraction

There are two methods.

i) online Extraction ii) offline Extraction i) online Extraction - Data is extracted from Source system itself. with online extractions, you need to consider whether the distributed transaction are using original source objects or mepared object.

in offine - The duta is not estractly directly from the source system.

Desta loading: - In this step extracted dave 8 transformed desta are loaded into torget database. It is the physical movement to takes three ways is initial load ii) Incremental load iii) full refresh methods for data loading are cloud-based, Batch processing open - source tools used are sprinkle.

Ab initio, IRI voracity.