

Data preprocessing & ETL

Data preprocessing:- It is a data mining technique that transforms raw data into a more understandable, useful & efficient format. Real world data is generally incomplete, Noisy, Inconsistent

Data cleaning:- It is also known as scrubbing. This task involves filling of missing values, smoothing or removing noisy data.

Reasons for dirty or unclear data.

- 1) Dummy values
- 2) Absence of data
- 3) Violation of business rules
- 4) Contradictory data
- 5) Reused primary keys.

How to clean data?

- 1) Handle missing values
- 2) Handle Noisy data
- 3) Remove unwanted data.

1) Handle missing values:- There are different ways to handle missing data depends on the problem domain & goal of data mining process.

a) Ignore data row:- It is used when max. amount of data is missing. It is avoided where less attribute values

are missing.

b) Fill missing values manually:-

It is very time consuming method & infeasible for almost all scenarios.

c) use global constant to fill in for missing values:- It is used when missing values are difficult to predict.

2) Handle noisy data - It may be introduced due to fault in data collection, error during data entering. Noise can be handled using binning.

3) Remove unwanted data - It is a duplicate or irrelevant data. Due to redundant records, the model may not provide accurate results as the duplicate data interferes with analysis process.

Data integration :- In this step a coherent data source is prepared. This is done by collecting & integrating data from multiple sources like databases, data cubes etc.

Issues in data integration:

a) Schema integration - schema from different sources may not be compatible. This leads to entity identification problem.

b) Data value conflicts - The values or metrics or representations of the same data may be different for the same

Date _____
Page _____

real world entity in diff. data sources

c) Redundant data :- Duplicate attributes or tuples may occur as a result of integrating data from various sources

Data Reduction - If the data is very large, data reduction is performed. Data reduction techniques are

- a) Data cube aggregation - Data is reduced by applying OLAP operations like slice, dice or roll up.
- b) Dimensionality reduction - Data attributes or dimensions are reduced.
- c) Data compression - large volume of data is compressed. This can be done by using lossy or lossless compression.
- d) Numerosity reduction - It can be done using histograms, clustering or sampling of data.

Discretization in data mining :- It refers to a method of converting a huge number of data values into smaller ones so that the evaluation & management of data become easy. There are two forms of data discretization first is supervised & second is unsupervised.

i) Supervised - It refers to a method in which the class data is used.

ii) Discretization - It is a method in which operation proceeds.

e.g

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 | 11, 16, 17, 18, 20 | 21, 25, 28 |

After Discre-

tization child young mature.

Concept hierarchy generation - It is an organizational structure mapping is done from low-level concepts to high level concepts. There are two types of hierarchy

i) Top-down ii) bottom-up.

i) Top-down - It generally starts with top with some general info. & ends with bottom to specialized info.

ii) Bottom-up:- It generally starts with bottom with some specialized info. & ends with top to the generalized info.

Data transformation - This involves normalization & aggregation of data according to the needs of data set.

It involves the following

Smoothing - It can work to remove

noise from the data.

Aggregation - This phase is generally used in making a data cube for the analysis of the data at multiple granularities.

Generalization - where low level details are restored by larger-level concepts through use of hierarchies.

Normalization - where the attribute data are scaled to fall within a specified range.

ETL: - It is a process that extracts data from diff. source system then transforms the data & finally loads the data into the data warehouse system.

ETL means Extract, Transform & Load. Extraction: - Data is extracted from source system for further use in a data warehouse. It is a time consuming process.

Extraction method: -

a) Logical Extraction b) Physical Extraction

a) Logical Extraction: - It is divided into two kinds.

i) Full Extraction ii) Incremental Extraction.

i) Full Extraction - Data is extracted completely from source system.

there is no need to keep track of changes to the data source since the last successful extraction.

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

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 prepared object.

ii) offline - The data is not extracted directly from the source system.

Data loading: - In this step extracted data & transformed data are loaded into target database. It is the physical movement. It takes three ways

i) 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, JRI voracity.