Module 2

Sources of Data

The sources of data can be classified into two types: statistical and non-statistical. Statistical sources refer to data that is gathered for some official purposes, incorporate censuses, and officially administered surveys. Non-statistical sources refer to the collection of data for other administrative purposes or for the private sector.

What are the different sources of data?

The following are the two sources of data:

Internal sources

When data is collected from reports and records of the organisation itself, they are known as the internal sources.

For example, a company publishes its annual report' on profit and loss, total sales, loans, wages, etc.

External sources

When data is collected from sources outside the organisation, they are known as the external sources. For example, if a tour and travel company obtains information on Karnataka tourism from Karnataka

Transport Corporation, it would be known as an external source of data.

Other sources of Data

A) Primary data

Primary data means first-hand information collected by an investigator.

It is collected for the first time.

It is original and more reliable.

For example, the population census conducted by the government of India after every ten years is primary data.

B) Secondary data

Secondary data refers to second-hand information.

It is not originally collected and rather obtained from already published or unpublished sources.

For example, the address of a person taken from the telephone directory or the phone number of a company taken from Just Dial are secondary data.

Data Readiness Level

The "readiness level" of data typically refers to how prepared or suitable data is for a specific purpose or analysis. The concept of data readiness level is not as standardized as technology readiness levels (TRLs) in engineering but can be adapted for various contexts.

Data readiness levels can be simplified into five main categories:

Raw Data (Data Readiness Level 1): At this level, data is in its most basic, unprocessed form, often collected or generated without any initial cleaning or transformation. It may be in the form of raw sensor readings, unstructured text, or images.

Cleaned and Structured Data (Data Readiness Level 2): Data at this level has undergone basic cleaning and structuring to remove errors, duplicates, and inconsistencies. It is typically organized into a more usable format, such as tables or databases.

Processed and Enhanced Data (Data Readiness Level 3): Data readiness level 3 includes data that has been processed to derive relevant features or metrics. This processing can involve calculations, aggregations, and transformations to make the data more suitable for analysis. Additionally, the data may be enriched with additional information, such as external data sources or metadata.

Annotated or Labeled Data (Data Readiness Level 4): In this category, data has been labeled or annotated, especially in supervised machine learning contexts. Labels are added to data instances to indicate the desired output or category, making it suitable for training machine learning models.

Curated and Validated Data (Data Readiness Level 5): This is the highest level of data readiness. Data at this level has been thoroughly curated, validated, and is

considered highly reliable and accurate for decision-making and analysis. It has undergone extensive quality control and validation processes.

Structured and Unstructured Data

Structured vs. Unstructured Data: 5 Key Differences

Here are the five main differences between structured vs. unstructured data:

Defined vs. Undefined Data

Structured data is clearly defined data in a structure. While unstructured data is usually stored in its native format, structured data lives in rows and columns and can be mapped into predefined fields.

Qualitative vs. Quantitative Data

Another difference between structured and unstructured data is that structured data is often quantitative data, meaning it usually consists of hard numbers or things that can be counted. (For example, product information in a customer relationship management system, or CRM.)

Unstructured data, on the other hand, is often categorized as qualitative data and cannot be processed and analyzed using conventional tools and methods.

Data Storage in Data Warehouses vs. Data Lakes

Businesses often store structured data in data warehouses and unstructured data in data lakes. A data

warehouse is an endpoint for the data's journey through an ETL pipeline. A data lake, on the other hand, is a sort of almost limitless repository where you store data in its original format or after undergoing a basic "cleaning" process.

Ease of Analysis

One of the most significant differences between structured and unstructured data is how well-structured data lends itself to analysis. Structured data is easy to search, both for data analytics experts and for algorithms. Unstructured data, on the other hand, is intrinsically more difficult to search and requires processing to become understandable.

Predefined Format vs. Variety of Formats

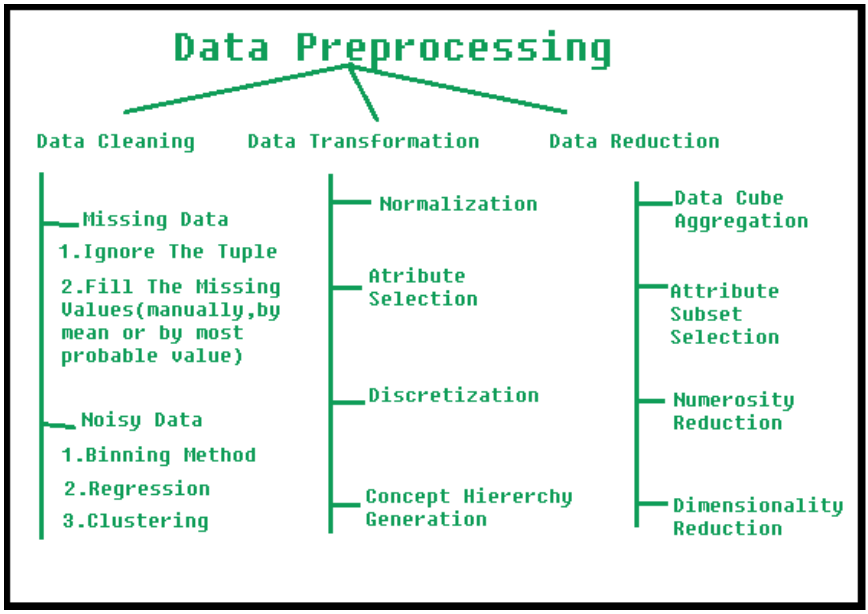
The most common format for structured data is text and numbers. Structured data has been defined beforehand in a data model.

Unstructured data, on the other hand, comes in a variety of shapes and sizes. It can consist of everything from audio, video, and imagery to email and sensor data.

Data Pre-processing Techniques

Introduction to Data Pre-processing Techniques

Data pre-processing is a data mining technique which is used to transform the raw data in a useful and efficient format.



Steps Involved in Data Pre-processing:

1. Data Cleaning (Error handling, Filtering)

2. Data Transformation

3. Data Reduction

4. Data Integration (Data Merging)

Each technique will be described below.

1. Data Cleaning Techniques (Error handling, Filtering)

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

Handling Missing Data:

This situation arises when some data is missing in the data. It can be handled in various ways.

Some of them are:

a) Ignore the tuples:

This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.

b) Fill the Missing values:

There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

Handling Noisy Data

Noisy data is a meaningless data that can't be interpreted by machines. It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :

a) Binning Method:

This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

b) Regression:

Here data can be made smooth by fitting it to a regression function. The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).

c) Clustering:

This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

2. Data Transformation Techniques

This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:

a) Normalization:

It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)

b) Attribute Selection:

In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

c) Discretization:

This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.

d) Concept Hierarchy Generation:

Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

3. Data Reduction Techniques

Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we use data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs.

The various steps to data reduction are:

a) Data Cube Aggregation:

Aggregation operation is applied to data for the construction of the data cube.

b) Attribute Subset Selection:

The highly relevant attributes should be used, rest all can be discarded. For performing attribute selection, one can use level of significance and p- value of the attribute. The attribute having p-value greater than significance level can be discarded.

c) Numerosity Reduction:

This enable to store the model of data instead of whole data, for example: Regression Models.

d) Dimensionality Reduction:

This reduce the size of data by encoding mechanisms. It can be lossy or lossless. If after reconstruction from compressed data, original data can be retrieved, such reduction are called lossless reduction else it is called lossy reduction. The two effective methods of dimensionality reduction are: Wavelet transforms and PCA (Principal Component Analysis).

4. Data Integration Techniques (Data Merging)

Data Integration is a data preprocessing technique that combines data from multiple heterogeneous data sources into a coherent data store and provides a unified view of the data. These sources may include multiple data cubes, databases, or flat files.

The data integration approaches are formally defined as triple <G, S, M> where,

G stand for the global schema,

S stands for the heterogeneous source of schema,

M stands for mapping between the queries of source and global schema.